FOLLOW-UP QUALIFICATION
LIST OF ISSUES
RESPONSE
EMA/SA/0000139783

PathAI
1325 Boylston Ave.
10th Floor
Boston, MA 02215

Scientific Advice
Human Medicines Division
European Medicines Agency
Domenico Scarlattilaan 6
1083 HS Amsterdam
The Netherlands

Please find below PathAI's response to the Follow-up Qualification List of Issues based on the coordinators' reports by the Scientific Advice Working Party (SAWP) during its 25-28 September 2023 meeting regarding the AI-Based Histologic Measurement of NASH (AIM-NASH). This response has been uploaded to the IRIS platform. Additionally, the history of the FDA and EMA activities for the qualification of this NASH drug development tool are provided below for reference:

FDA Activities

- Held Pre-Letter of Intent (LOI) Meeting with FDA – January 28, 2020
- Submitted LOI to FDA – April 30, 2020
- Submitted Revised LOI to FDA – May 7, 2020
- Received LOI Determination Letter – September 18, 2020
- Held LOI Feedback Review Meeting with FDA – November 9, 2020
- Submitted Draft QP to FDA – December 21, 2020
- Received Reviewability Memorandum from FDA – January 22, 2021
- Response and Revised QP resubmitted to FDA by PathAI – February 3, 2021
- Received 2nd Reviewability Memorandum from FDA – June 17, 2021
- Informal FDA QP Feedback Meeting – October 15, 2021
- Submitted Revised QP to FDA – November 23, 2021
- Received 3rd Reviewability Memorandum from FDA – March 25, 2022
- Response and Revised QP resubmitted to FDA by PathAI – March 29, 2022
- Received Information Request from FDA – February 28, 2023
- Response and Revised QP resubmitted to FDA by PathAI – March 14, 2023

EMA Activities

- Submitted Draft Briefing Document to EMA – October 30, 2020
- Preparatory Meeting with EMA – May 17, 2021

- Submitted Final Briefing Doc to EMA – May 27, 2021
- Final Briefing Document Accepted and Review Began – June 6, 2021
- EMA Deliver's List of Issues (LoI) to be addressed – July 5, 2021
- Informal feedback SAWP meeting to discuss LOS responses – September 1, 2021
- Final Briefing Document acceptance and Qualification Advice delivered by EMA – November 19, 2021
- Submitted Draft Briefing Document to EMA – May 8, 2023
- Preparatory Meeting with EMA – June 13, 2023
- Submitted Final Briefing Doc to EMA – June 27, 2023
- Final Briefing Document Accepted and Review Began – June 29, 2023
- EMA Delivers List of Issues (LoI) to be addressed – October 13, 2023
- Submitted written response for LoI – November 16, 2023

Sincerely,

Dr. Katy Wack

Primary Contact:
Name: Katy Wack
Phone: +1 412-728-1217
Email: katy.wack@pathai.com

Alternative Contact:
Name: Nick Anderson
Phone: +1 571-242-6589
Email: nick.anderson@pathai.com

List of issues to be addressed in writing and during the discussion meeting

Based on the coordinators' reports the Scientific Advice Working Party (SAWP) determined that the Applicant should discuss the following points, before advice can be provided:

General Questions including context of use and life-cycle management

1. The applicant should justify the appropriateness of the context of use statement and propose a potential revision.

   PathAI Response:

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

   The AIM-NASH outputs will mirror the current EMA guidelines for NASH evaluation for enrollment in clinical trials, measurement at follow-up time points and histologic endpoint evaluation. AIM-NASH is accessible to users on a web-based platform, integrating into NASH clinical trials without significant impact to workflow or introducing new risk to patients. The pathologist will review the output of AIM-NASH and take an active role in its interpretation by accepting or rejecting each of the NAS components and fibrosis stage, after confirming sample evaluability and determining the presence of any additional findings. AIM-NASH should be used with slides scanned with the Aperio AT2 scanner.

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

   Updated Proposed 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 NASH clinical trials. The tool is an aid to the pathologist that is to be used during the enrollment of patients and to capture change in score over time in response to a therapeutic agent for evaluation of histologic-based endpoints.

   The AIM-NASH outputs will mirror the current EMA guidelines for NASH biopsy evaluation for enrollment in clinical trials, measurement at follow-up time points and for histologic endpoint evaluation. The biomarker is applicable to screening and follow-up time points for phase 2 and phase 3 NASH trials. This includes patients with fibrosis stages ranging from 0-4 and NAS <4 and ≥ 4.

   AIM-NASH is accessible to users on a web-based platform, integrating into NASH

clinical trials without significant impact to workflow or introducing new risk to patients. The pathologist will review the output of AIM-NASH and take an active role in its interpretation by accepting or rejecting each of the NAS components and fibrosis stage, after confirming sample evaluability and determining the presence of any additional findings. AIM-NASH should be used with slides scanned with the Aperio AT2 scanner.

Justification of Proposed COU:
This biomarker measurement tool aims to provide a more reliable histologic score to aid the pathologist and replace fully manual scoring during any relevant assessments of NASH trial biopsies. Depending on the enrollment or follow-up timepoint context of use, the histologic score may be seen as a diagnostic and monitoring biomarker based on the BEST categories. The studies presented here, including overlay validation, AV, and CV, along with the published exploratory efficacy studies in partnership with NASH drug developers, thoroughly test the accuracy, reproducibility, and clinical trial impact across a large and robust dataset to demonstrate this context of use. The datasets for these studies adequately represent the current NASH clinical trial screened and enrolled populations and include multiple drug candidates with baseline and follow-up timepoint biopsies collected from many global investigational sites, where multiple labs were used to stain and scan the samples to generate WSIs used for these prospective analyses.

The studies met primary, secondary, and exploratory endpoints for accuracy using a robust and large dataset and demonstrated superior accuracy of pathologists aided by AIM-NASH in determining key inclusion and endpoint criteria, as well as greater reproducibility compared to that of pathologists performing manual reads (both intra- and inter-pathologist). The evidence here demonstrates significant progress in providing a reliable, validated tool for pathologists to use to accurately and consistently enroll NASH patients into clinical trials and to more reliably detect drug response to bring effective treatments to patients.

2. Please describe plans for lifecycle management in full detail since it remains unclear what that applicant considers a “major change”. These plans should cover re-testing with changes to the model, future clinical trial data, generalizability of the tool to settings outside the US, potential extension to other scanner models and scanning procedures, changes in the training procedure(s) and expanded training populations. The applicant should also elaborate on the possibility that small changes over a long time would add up to a large change that would require revalidation. Please also describe whether plans for further development/improvement are already envisaged.

PathAI Response:
As part of lifecycle management, PathAI has an SOP describing our 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 does cover the retesting requirements for the situations you have cited, 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.

We recognize that it is possible for small changes over time to accumulate and require some revalidation. 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.

Currently, PathAI has no specific plans for further development/improvement of the AIM-NASH product.

3. The training and validation 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, it is questioned if the model was sufficiently trained to recognise the “extremes” of NASH histology features (e.g., only 3.6%, 4.5% and 7.7% of the F0/1/2 stages fibrosis, respectively, were included into the training dataset). Please elaborate on the potential implications of this low “abundance”
   a. During model development
   b. Within the validation studies especially regarding the analytical and clinical validation

   The applicant is also requested to provide per class (score) performance measures (aside weighted kappa's (WKs)) to provide a better understanding of how well the models perform on underrepresented classes. Please discuss the overall generalisability of the tool based on the available datasets, including the case studies, considering the overall still limited number of trials which have (re-)evaluated their data with the use of the AIM-NASH tool. Please also present additional "case study" results in case such have become available meanwhile.

PathAI Response:

a. From a Model Development Perspective:
We are aware that in the clinical trial data sets used for model development and validation, disease stages are not equally represented (e.g., low-stage fibrosis is underrepresented as highlighted by the reviewers) which is due to the patient populations encountered during the screening and follow-up time points for these trials. Our selection of model development and validation data reflects the target population in which the tool is intended to be used.

The main risk of such an "unbalanced" dataset in model development is that the model is exposed to fewer of the rare examples and as a result might not show as strong generalizability for these classes. We attempted to minimize this risk in two ways. During tissue model development, we applied augmentations to each sample as described in updated briefing document submitted August 24, 2023 (Briefing Document) Section 3.7.3, effectively increasing the number of samples even for rare classes. During development of the scoring model, we up-weighted samples of rare classes. Specifically, samples were not picked with uniform probability, but instead inversely proportional to the logarithm of their relative frequency. Additionally, please see supportive analyses in the LoI Response (Q6) for CHMP Qualification Advice EMADOC-1700519818-731375, which evaluates low and high fibrosis stage performance in our standalone verification test set. Importantly, when used in NASH trials, a pathologist will review all trial cases to ensure the sample is adequate for algorithm-assisted scoring.

b. From an Analytical and Clinical Validation Perspective:
The implications of low representation of some scores in the development data set for analytical and clinical validation are lower performance in these score categories. To mitigate this risk, we have included a diverse trial dataset in the validation studies, included biopsies from 4 separate NASH clinical trials with different drug candidates and different phases (one Phase 3 non-cirrhotic trial, two phase 2 non-cirrhotic trials and one phase 2 cirrhotic trial). This study population is representative of current screened and enrolled NASH clinical trial populations. If the clinical trial population evolves (e.g., patients with earlier stage disease are screened), we will determine through performance monitoring whether additional training is needed according to our change control processes.

In addition to the sub-analyses performed in this response including individual histologic score analyses as well as impact on inclusion criteria and endpoint composite scores (Appendix A Table 24, Table 25, and Table 26), we included trial relevant categories in our original data analyses for both analytical and clinical validation analysis. These include F0&F1 vs other fibrosis stages (Briefing Document Table 49 for AV and Table 87 for CV), F2&F3 vs other fibrosis stages (Briefing Document Table 50 for AV and Table 92 for CV), F4 vs other fibrosis stages (Briefing Document Table 49 for AV and Table 87 for CV), NAS ≥4 and ≥1 for each component vs. Other (Briefing Document Table 50 for AV and Table 92 for CV) and NASH resolution (Briefing Document Table 50 for AV and Table 92 for CV); for each of these categories (including the "extremes"), AIM-NASH and

pathologist assisted by AIM-NASH perform similar or better than an IMR, indicating the generalizability of this tool across different levels of disease activity in multiple clinical trials.

Per Histologic Score AV/CV Percent Agreement Tables (Appendix A Table 4, Table 7, Table 10, Table 13, Table 16, Table 18, Table 20, Table 22, Table 5, Table 8, Table 11, Table 14, Table 17, Table 19, Table 21, Table 23, Table 6, Table 9, Table 12, Table 15)

Although, for some features, AIM-NASH % over-call and under-call relative to the reference GT panel are not completely balanced (e.g., LI 0-1 border), it is important to note that the study was not designed to be powered for every score level, and this sample size and distribution would be extremely difficult and burdensome to attain. Additionally, many IMR's scores relative to GT are also not "balanced" in this respect, with different pathologists displaying various levels of over- vs. under-calling for different features and score levels. Additionally, to our knowledge, there are no published studies in the literature, describing intra or inter-pathologist agreement at the individual score levels for reference. Finally, it is hypothesized that this agreement imbalance would likely also occur between GT panels ((Sanyal et al. 2021)) consisting of different expert pathologists, depending on which panel is used as the reference. This is largely due to the lack of standardization (e.g., identifying different sub-types as ballooned cells) for NASH histologic feature identification and scoring (e.g, a panel of two conservative pathologists may agree on lower balloon scores than a panel of a two less conservative pathologists who may identify and score more ballooned cells; see (Brunt et al. 2022)).

Results to support potential clinical trial impact if individual score agreements:
Commenting on clinical trial impact, the results describing composite score agreement (inclusion criteria such as NAS>=4, NASH resolution, etc.) for AIM-NASH compared to GT relative to IMRs compared to GT should, in part, illustrate the potential impact since validation trial datasets used here were representative of current NASH trial screened and enrolled populations. Kappa results demonstrate non-inferiority for all features and superiority for NAS >=4, as well as NASH resolution, and % agreement tables included here (Appendix A Table 24, Table 25, Table 26) demonstrate superiority for AIM-NASH in identifying F2 and 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 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
2. the superior repeatability/reproducibility of AIM-NASH compared to manual pathology (intra- and inter-) should result in more accurate, standardized and consistent enrollment and detection of steatosis, balloon, and/or inflammation grade change or fibrosis stage change for a patient in a trial.

Efficacy Analyses supported by Published Case Studies:
In addition to a large number of published case studies across several trials and drug candidates (semaglutide, pegbelfermin, resmetirom; see Appendix C) in various phases included in Briefing Document Section 4.8, we also recently demonstrated that AIM-NASH recapitulates primary efficacy results from a Phase 3 study of resmetirom (Appendix C, case study #5) with 966 patients ((Iyer et al. 2023)), similar to what was demonstrated in the corresponding phase 2b (Appendix C, case study #2, also in the Briefing Document Section 4.8, case study #2) study for resmetirom. For both NASH resolution and fibrosis improvement endpoint, the percentage of patients that responded were comparable when assessed by AIM-NASH or manual pathology assessment. This supportive evidence, along with the AV/CV evidence supported here strongly demonstrates the robust nature of the AIM-NASH tool across a wide range of disease activity and in the phase of drug treatments.

4. All evaluations have in principle been conducted on an "observed cases" basis with exclusion of slides due to various reasons. This kind of evaluation does not comply with the ITT principle, which should also include cases with part of the data not being available. The applicant is requested to present an evaluation of the full set of cases for which a GT evaluation is available and re-analyse the data (with sensible imputation methods for missing data), as far as possible.

   PathAI Response:
   In order to perform sensible imputation methods, we evaluated the following scenarios for all cases in which a GT evaluation was present, but either an AIM-NASH, AIM-NASH-Assisted, or IMR value was missing per case:

   DEFINITIONS:
   1. Worst Case for AIM-NASH or AIM-NASH-Assisted for a missing value equals the score furthest from that supplied by GT for that component (e.g., GT is Steatosis=3, AIM-NASH would be assigned a score of Steatosis=0).
   2. Best Case for AIM-NASH or AIM-NASH-Assisted for a missing value equals the GT score for that component (e.g., GT is Steatosis=3, AIM-NASH would be assigned a score of Steatosis=3).
   3. Worst case for IMR for a missing value equals the score furthest from that supplied by GT for that component (e.g., GT is Steatosis=3, IMR would be assigned a score of Steatosis=0).
   4. Best Case for IMR for a missing value equals the GT score for that component (e.g., GT is Steatosis=3, IMR would be assigned a score of Steatosis=3).

   SCENARIOS (from highest potential negative impact on accuracy (A) to lowest potential negative impact on accuracy (D):

A. Worst Case for AIM-NASH/AIM-NASH-assisted plus the Best Case for IMR (Definitions 1+4), across all features and cases (Appendix A Table 36, Table 37, Table 38 for AV, CV AIM-NASH-assisted, and CV AIM-NASH respectively).
B. Worst Case for AIM-NASH/AIM-NASH-assisted plus the Worst Case for IMR (Definitions 1+3) across all features and cases (Appendix A Table 33, Table 34, Table 35 for AV, CV AIM-NASH-assisted, and CV AIM-NASH respectively).
C. Best Case for AIM-NASH/AIM-NASH-assisted plus the Best Case for IMR (Definitions 2+4), across all features and cases (Table 42, Table 43, **Table** 44 for AV, CV AIM-NASH-assisted, and CV AIM-NASH respectively).
D. Best Case for AIM-NASH/AIM-NASH-assisted plus the Worst Case for IMR (Definitions 2+3), across all features and cases (Appendix A Table 39, Table 40, Table 41 for AV, CV AIM-NASH-assisted, and CV AIM-NASH respectively).

Non-inferiority was met for all missing data scenarios and datasets, except for in the AV dataset for steatosis for best or worst case scenario for AIM-NASH and best case scenario for IMR (A, Appendix A Table 36: where for worst case, all AIM-NASH values across cases would be furthest away from GT, for best cases, all IMR and AIM-NASH missing values would need to have perfect agreement with GT), where non-inferiority p value was 0.052. For AV, the best and worse case scenario for AIM-NASH were exactly the same because there were no missing AIM-NASH reads, as AIM-NASH always returns a score if tissue is present. In a clinical trial setting, the pathologist has the ultimate control and is responsible for deeming the sample (tissue, stain, scan) as adequate/inadequate for evaluation purposes. In the larger and more robust CV dataset, where pathologists, both IMR and AI-assist, did deem some samples as inadequate or non-evaluable, non-inferiority was met for all best and worse case combinations and all features.

For all other imputations for every dataset, primary endpoint was met. In addition, for AV and CV, superiority was achieved for all histologic features for multiple scenarios (Appendix A Table 33, Table 39, Table 34, Table 40, Table 35, and Table 41) and superiority was met for inflammation and ballooning for (Appendix A Table 33, Table 39, Table 34, Table 40, Table 43, Table 35, Table 38, Table 41, **Table** 44). For ballooning, AIM-NASH algorithm and AIM-NASH-assisted achieved superiority for all scenarios and datasets.

Importantly, no values were missing for AIM-NASH or IMR for the AV repeatability/reproducibility dataset, so no imputations were needed.

We performed the above analyses for AV, CV AIM-NASH-Assisted, and CV AIM-NASH algorithm only datasets. Worst case scenarios where the test (IMR or AIM-NASH/AIM-NASH-Assisted) are furthest away in score from GT are highly unlikely to occur (agreement tables show that disagreements are only off by a score of 1 for a majority of the cases) but represent the extreme or border scenario for illustration purposes.

Validation of the AISight Trials Platform or Translational platform

5. During validation of both platforms, the primary analysis was based on the comparison of the agreement of NASH and non-NASH diagnosis ground truth (GT) vs Study pathologist Glass [reads] and agreement of NASH diagnosis GT vs study pathologist WSI. The choice is not completely understood as the PathAI tool will be used in clinical trials not to establish diagnosis of NASH, but rather to provide scores based on which the diagnosis will take place. In that respect, the comparison of GT to Glass/WSI study pathologists on the level of individual NASH components scores is of interest. The Applicant is asked to comment if such comparisons can be provided.

   PathAI Response:
   The diagnosis of NASH in the platform validation studies was defined in a trial inclusion context of use as NAFLD Activity Score (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, similar to the definition used during NASH clinical trial enrollment. Therefore, we feel that the endpoint is appropriate for platform validation, as it takes into account scoring of different NASH components and any additional findings that the pathologists might have identified during a clinical trial.

   In an additional analysis, as requested, each histologic component score was assessed for agreement to GT, for both AIM-NASH scores and manual pathologist reads. The agreements for all component scores and overall NAS, by weighted kappa, 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.

Whole Slide Image (WSI) and Glass reads agreement with Consensus Ground Truth

Feature	Modality	N	Weighted Kappa (95% CI)
Steatosis	WSI vs GT	159	0.580 (0.505, 0.640)
	Glass vs GT	159	0.593 (0.519, 0.655)
Lobular Inflammation	WSI vs GT	159	0.367 (0.300, 0.432)
	Glass vs GT	159	0.380 (0.315, 0.445)
Hepatocellular Ballooning	WSI vs GT	159	0.537 (0.457, 0.608)
	Glass vs GT	159	0.522 (0.435, 0.595)
Fibrosis	WSI vs GT	159	0.640 (0.574, 0.695)
	Glass vs GT	159	0.604 (0.536, 0.662)
NAS	WSI vs GT	159	0.527 (0.476, 0.573)
	Glass vs GT	159	0.525 (0.473, 0.571)

Additionally, in order to design a validation study that is powered for each NASH component score, the sample size would be unreasonably large for this type of a study. In the current platform validation studies, the comparisons for individual NASH component scores might not be representative of true scoring variability due to inadequate powering.

6. It would have been of interest to measure intra-reader variability when determined on glass reads (intra-modality) within the study, instead of comparing it to the reported one in the literature, as was done by the Applicant. The Applicant is asked to comment if intra-reader variability is different when reads are performed intra-modality on glass or WSI based on the current knowledge, and if intra-reader and inter-modality WKs are different from the intra-reader intra-modality WKs.

   PathAI Response:
   Given that the intra-pathologist inter-modality (glass to digital) agreements demonstrated in this study are higher than intra- pathologist, intra-modality (glass to glass) values by expert pathologists in the literature and given that non-inferiority in agreement with glass GT was achieved for primary endpoint, we feel this is not necessary for validation, but we agree this comparison could be of interest.

Analytical validation

7. 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 ground truth would help assessing whether the disagreement is 'balanced' or whether there is a systematic deviation (this also applies to the clinical validation).

   PathAI Response:
   In considering types of disagreements for AIM-NASH and IMRs compared to GT, please see response for question 3, referencing individual component score tables. Although, for some features, AIM-NASH % over-call and under-call relative to the GT panel are not completely balanced (e.g., LI 0-1 border) and neither are IMR reads compared to GT, it is important to note that the study was not designed to be powered for every score level, and this sample size and distribution would be extremely difficult to attain. Many average IMR scores relative to GT are also not "balanced" in this respect, with different pathologists displaying various levels of over- vs. under-calling for different features and score levels, as can be observed in the individual IMR agreement tables. Therefore, it is helpful to consider impacts for clinical trial relevant score evaluations (e.g., composite inclusion criteria score evaluations, endpoint composite score evaluations such as NASH Resolution (defined by ballooning=0, inflammation= 0 or 1, steatosis=any value).

   In addition to providing overall percent agreement (OPA) for the composite clinical trial endpoints (Appendix A Table 24, Table 25, Table 26), we have provided data tables for positive percent agreement (PPA; Appendix A Table 27 for AV, Table 28 for CV AIM-

NASH-assisted and Table 29 for CV AIM-NASH algorithm only) and negative percent agreement (NPA; Appendix A Table 30 for AV, Table 31 for CV AIM-NASH-assisted and **Table** 32 for CV AIM-NASH algorithm only) for relevant clinical trial composite endpoints (NAS ≥ 4 with at least 1 in each NAS category, Fibrosis score 2 or 3, Fibrosis score 4 and NASH resolution). For AV, PPA for AIM-NASH vs GT was over 75% for all composite score categories and was also higher for all the categories compared to IMR vs GT, with NAS ≥ 4 and NASH resolution being superior to IMR vs GT. Strikingly, NASH resolution PPA was 16.1% higher for AIM-NASH than for IMRs compared to GT. Similar results were achieved in CV for AIM-NASH only and AIM-NASH-assisted workflow, except Fibrosis score 2 and 3 also achieved superiority. For CV, both AIM-NASH-Assisted (Table 28) and AIM, AIM-NASH algorithm only detected NASH resolution with 18% greater sensitivity (PPA) than IMRs.
For NPA, AIM-NASH vs GT was over 70% for all of the composite score categories and similar to the PPA, AIM-NASH vs GT was also superior to IMR vs GT for NAS ≥ 4 and NASH resolution. For CV NPA, AIM-NASH alone and AIM-NASH-assisted both were superior to IMR vs GT for NASH resolution.

In calculating change over time-based endpoints (e.g., change in NAS score and/or fibrosis score), both accuracy and reproducibility will be important. Since AIM-NASH algorithm is more reproducible than manual readers (both intra- and inter-) and, equivalent or higher for component accuracy, one can deduce that subject level change over time will be able to be more accurately detected by AIM-NASH, with fewer false negatives or false positives due to reader variability and enrollment scoring bias. Our cases studies across several trials provide supportive evidence for AIM-NASH in accurately capturing all of the above composite scores which can affect detection of drug response across several trials (Appendix C)

8. In the accuracy testing, WKs for AIM-NASH vs. GT were on a lower side, also compared to the literature. Especially WKs for inflammation were extremely low, in the range of 0.173 – 0.372 for different inflammation stages. WKs for fibrosis were also quite low, especially for lower fibrosis stages, with WKs for F0-F2 fibrosis being in the range of 0.263- 0.455. The Applicant is asked to comment on these low WKs values, especially for fibrosis stages (AIM-NASH vs GT) and discuss implications for the use of the tool in clinical studies, taking into account that regression of fibrosis is often included as one of the study endpoints, and patients with less severe NASH are being included in the NASH trials (of note, reproducibility was also lower for lower fibrosis stages). Also, the extremely low WKs for inflammation both for AIM-NASH and IMR should be justified. The same concerns hold for the CV validation.

PathAI Response:
The analytical and clinical validation studies were powered for overall score per histologic feature, and not per individual score level within each feature and therefore, kappa analysis for some individual scores are underpowered. Additionally, concerning

individual score representation, the validation datasets are representative of the populations encountered during screening and follow-up of NASH clinical trials and do not have an equal spread of scores per histologic feature (e.g., trials rarely have fibrosis score 0). For all histologic components overall, the kappas achieved (Briefing Document, Table 48, 86) were within the range achieved in analogous clinical trial datasets in the literature for individual pair-wise pathologist comparisons (Briefing Doc, Table 3), but it's also important to note that the kappas in these AV and CV studies are comparing individual (or algorithm) to a consensus GT of three pathologists, not to another individual. Inflammation presents the highest for inter-pathologist variability, as indicated by the lower kappas achieved, across the literature vs. any other scoring component. This also illustrates the importance of finding not only an accurate method of scoring, but one that is reliably consistent over time and across pathologists. Lastly, the published values are for overall histologic component agreements only and to our knowledge, no agreement data exists for each individual score levels within each feature. AIM-NASH/AIM-NASH-assisted kappas are similar to or higher than IMR kappas in these categories and superior repeatability/reproducibility of AIM-NASH (compared to that demonstrated by experts in the literature), leads us to believe that AIM-NASH-assisted scoring in clinical trials will results in more accurate and reproducible results.

On inter-site reproducibility, since precision studies are AIM-NASH only, without the pathologist review, we believe that the differences in QC protocols and staining quality between different sites can be mitigated by a pathologist review, where the pathologist has an option to request for restain and rescan, and all inter-site agreements achieved by AIM-NASH were higher than inter-pathologist agreements in this study and across the literature.

The impact on clinical trials was also discussed in question #3, indicating that we included trial relevant categories in our original data analyses for both analytical and clinical validation analysis. These include F0&F1 vs other fibrosis stages (Briefing Document Table 49 for AV and Table 87 for CV), F2&F3 vs other fibrosis stages (Briefing Document Table 50 for AV and Table 92 for CV), F4 vs other fibrosis stages (Briefing Document Table 49 for AV and Table 87 for CV), NAS ≥4 and ≥1 for each component vs. Other (Briefing Document Table 50 for AV and Table 92 for CV) and NASH resolution (Briefing Document Table 50 for AV and Table 92 for CV); for each of these categories (including the "extremes"), AIM-NASH and pathologist assisted by AIM-NASH perform similar or better than an IMR. Additionally, the composite score agreement (inclusion criteria such as NAS>=4, NASH resolution, etc.) for AIM-NASH compared to GT relative to IMRs compared to GT should, in part, illustrate the potential impact since validation trial datasets used here were representative of current NASH trial screened and enrolled populations. Kappa results demonstrate non-inferiority for all features and superiority for NAS >=4, as well as NASH resolution, and % agreement tables included here (Appendix A, **Table** 24, **Table** 25, **Table** 26 demonstrate superiority for AIM-NASH in identifying F2, 3 populations, NAS>=4, and NASH resolution across the entire CV dataset.

In summary, 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 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
2. the superior repeatability/reproducibility of AIM-NASH compared to manual pathology (intra- and inter-) should result in more accurate, standardized and consistent enrollment and detection of steatosis grade change or fibrosis stage change for a patient in a trial.

9. Standalone analytical verification: Neither the briefing document nor the submitted study report in the Appendix X contains a statistical evaluation on how the differences between the different kappa's complied with the pre-defined requirements. Also, results for the accuracy evaluation per NAS component appears to be missing. Therefore, currently, the step to move forward with analytical and clinical validation cannot be followed. It also needs to be clarified why there were 19 and 29 exclusions from the evaluation, and whether these were related to the manual reading, or the inability of the AIM NASH to evaluate these. It also appears to be counter logic to state in the study report that no re-evaluation is necessary based on the new Version 1.1.0 but afterwards conduct the IAV study, which has the verification of the new version defined as one of its purposes. Please comment.

PathAI Response:
The acceptance criteria in the protocol (Briefing Document Appendix Xa, section 6.1.4), defines the acceptance criteria as the lower 2.5% confidence interval of the linearly weighted kappa of the ML scores (AIM-NASH) vs. the reference standard median consensus scores be at least as good as 0.1 below the mean pairwise linearly weighted kappa among network pathologists. Even though not clearly stated, this acceptance criteria applies to the accuracy statement in protocol section 6.1.4.1 (Accuracy will be assessed separately for each NAS component (ballooning, steatosis and lobular inflammation) and CRN fibrosis score.) The results for this analysis per NASH component are listed in Briefing Document Table 22, we have clarified each column in Appendix A Table 46. Each of the histologic features was within the 0.1 non-inferiority acceptance criteria and therefore the step to move forward with analytical and clinical validation was justified.

The 19 H&E and 29 trichrome slides excluded from the analysis were deemed non-evaluable by consensus reads and not excluded due to AIM-NASH inability to evaluate these.

Integrated analytical verification was performed due to AIM-NASH algorithm integration to the AISight Clinical Trial platform as the Standalone Analytical Verification was performed on the development platform, and not due to the version change.

Clinical validation

10. In the AV and CV studies the Applicant compared WKs for AIM-NASH/AIM-NASH-assisted vs GT and WKs for IMR vs GT. This analysis is of interest but does not say anything about the overall accuracy of the tool and could be considered supportive only. The comparison of AIM-NASH-assisted score with GT may be considered the most relevant analysis instead, as AI assisted scoring is going to replace the GT in the clinical trial. Please comment.

    PathAI Response:
    The main approach aligned upon in the qualification plan and advice for this tool, was designed to test the accuracy of AIM-NASH or pathologists assisted by AIM-NASH, compared to that achieved on average by expert NASH pathologists who score manually, relative to each modality's agreement with the gold standard or "ground truth." The overarching goal is to provide an accurate, standardized, and reproducible scoring method to be used in the enrollment and endpoint contexts. Therefore, in addition to providing a high level of accuracy, the tool must also be highly repeatable and reproducible across readers or panels of readers (and relative to that achieved by pathologists both inter- and intra in this study and in the literature) for consistent enrollment and accurate detection of change in score over time. Although the consensus gold standard helps to reduce the influence of only one pathologist's scoring style and variability over time (e.g., ballooning identification and quantification), gold standard panels are still subject to lack of standardization (see inter-panel agreement in (Sanyal et al. 2021), demonstrating similar inter-agreement to that achieved between expert CRN individuals in the literature), consistency, and are also still subject to enrollment bias, encountered when biopsies are borderline for inclusion criteria and scores tend to trend towards inclusion (e.g., border of 0 vs. 1 for ballooning or F3 vs. F4 for a non-cirrhotic trial). Therefore, the goal was to provide a tool that is just as accurate as the expert, unassisted readers, but also to standardize and make the scoring more reliable over time and read conditions, which is essential in enrolling trials consistently and accurately capturing change over time and across phases and drug candidates.

    This being said, we agree it would be relevant to compare the performance of a pathologist assisted by AIM-NASH to that of a consensus of pathologists. Recently, many pathology read strategies in trials have utilized a statistical consensus method, 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. Utilizing this consensus methodology, we executed an additional analysis to address this topic.

    In order to compare AIM-NASH-assisted scoring to a current gold standard workflow with 3 pathologists, we assessed non-inferiority of AIM-NASH-assisted reads vs. a statistical GT consensus using mode/median scores; we then compared the algorithm

agreement to that achieved by a statistical consensus derived from IMR reads using the same mode/median approach for each case to that of the GT mode/median. Although this comparison may be a higher bar than when comparing two panels who reach consensus after discussion (the statistical consensus for the two groups would potentially be more consistent than for two ground truth panels establishing consensus through subjective discussion), this should adequately provide evidence to compare AIM-NASH-assisted read performance to a group of experts, rather than an average expert individual.

In this mode/median analysis, the primary endpoint of non-inferiority was met for all histologic components for AIM-NASH-assisted reads compared to statistical consensus reads. In addition, AIM-NASH-assisted read agreement for hepatocellular ballooning with GT was superior to IMR statistical consensus compared to GT. For steatosis, the average WK for AIM-NASH-assisted vs GT was 0.676 and for manual median was 0.743, with a difference of –0.067; for lobular inflammation the WK for AIM-NASH-assisted vs GT was 0.422 and for manual median was 0.443, with a difference of –0.021; for hepatocellular ballooning the WK for AIM-NASH-assisted vs GT was 0.563 and for manual median was 0.522, with a difference of 0.041; and for fibrosis the WK for AIM-NASH-assisted vs GT was 0.655 and for manual median was 0.719, with a difference of –0.064. The non-inferiority margin, pre-defined at 0.1, was met statistically for each component (Table 45). Reduction of time-point (e.g,, enrollment or temporal) bias and consistency over time is essential in standardizing scoring and accurately capturing change in grade/stage for primary histologic endpoints within and across trials, AIM-NASH is also expected to improve upon the gold standard, which is still subject to these types of challenges. As demonstrated in (Sanyal et al. 2021), inter- and intra-panel variability is largely equivalent to the inter- and intra-pathologist variability for individual expert CRN pathologists. demonstrating continued challenges in standardization across gold standard panels. The AIM-NASH algorithm is 100% repeatable on the same whole slide image and demonstrates above 90% repeatability across different scans at the same site over the course of three non-consecutive days, and 85% or above across 3 external lab sites. Given that the challenges targeted to be addressed include the need for an accurate, standardized, consistent enrollment and measurement of scores over time for endpoint analysis, both accuracy and reproducibility must be part of the solution. For accuracy Pathologists assisted by AIM-NASH are superior to individual pathologists for multiple histologic component scores, and non-inferior to a consensus, demonstrating high levels of accuracy, comparable to the gold standard, for scoring assessment of a biopsy. This performance characteristic combined with superior repeatability/reproducibility of AIM-NASH vs. manual readers (both individuals and panels, (Sanyal et al. 2021)), should allow for more accurate and reproducible scoring in NASH clinical trials.

11. Please elaborate on the kappa values achieved during the analytical and clinical validation exercises, and how these compare to the values reported in the literature. Please especially elaborate on the following:

a. Different "types" of kappa-values have been reported (e.g. Sanyal 2021) and the applicant is requested to display which type of kappa has been used in which studies, including their own analytical and clinical validation exercises.
b. Please elaborate on the fact that – when applying an inter-study comparison between the analytical and clinical validation – the algorithm appears to perform similar irrespective of the proposed "intervention" or "help" of a histopathologist (and the overlays). While it is agreed that histopathologist involvement has "face value", the usefulness of the manual supervision should also be justified on data.
c. Please elaborate on the clinical relevance of the comparisons to the study by Davison (2020) which appears to represent a "worst case scenario", and whether other reports with regard to variability (and repeatability) in NASH clinical trials can be found (apart from the two NASH-CRN studies, and the one by Sanyal et al (2021)).
d. Please elaborate on the clinical meaningfulness of the results achieved during analytical and clinical validation studies with respect to the "external comparison" to studies with "optimal" variability (Kleiner et al 2005 and 2019, Sanyal 2021).

PathAI Response:

a. The kappa values reported in the analytical and clinical validation studies are linearly weighted kappas (Cicchetti-Allison), and the same linearly weighted kappas are also reported in (Davison et al. 2020), (Kleiner et al. 2005) and (Kleiner et al. 2019) papers and are therefore directly comparable as far as agreement methodology is concerned. In (Sanyal et al. 2021) (and Briefing Document Table 7), the kappas from (Kleiner et al. 2005) and (Kleiner et al. 2019) are incorrectly labeled as quadratic weighted kappas (Shrout-Fleiss), and this was confirmed with direct correspondence with the NASH CRN's Data Coordinating Center, who confirmed methods used in the Kleiner papers (Cicchetti and Allison 1971).

b. While the results of the algorithm only on these large and extensive AV and CV datasets are very promising and do indicate that the algorithm alone is performing just as well or better than a pathologist alone, the biopsies analyzed prospectively in the "real world" clinical trial setting could potentially present new challenges for the algorithm (e.g., a new drug candidate with novel mechanism of action, a different inclusion criteria or biopsy timepoint schedule, a different staining protocol), and having a pathologist serve as a "control" or QC for the algorithm, we feel, is an important risk measure. Due to the need for standardization and consistency and given the performance demonstrated by the algorithm alone in CV, we do feel as though it is necessary to keep the controls placed on how the score can be changed. Additionally, it's essential during enrolling and for follow-up timepoints that the pathologist assesses the sample adequacy and evaluability (number of portal tracts per protocol, etc.) confirms the overall diagnosis and identifies any other or additional findings, which the algorithm is not trained to do.

c. and d. The Davison study (Davison et al. 2020) utilizes a NASH clinical trial ph2b dataset, with common inclusion criteria, containing both placebo and drug-treated biopsies, whereas the Kleiner studies (Kleiner et al. 2005; 2019) contained clinical samples and the

2005 had an extremely low n. Additionally, the Davison paper compares pair-wise kappas among three pathologists which could be low or high depending on the pairs being examined and the similarity of their reading styles (especially for ballooning, inflammation). Kleiner averages across the whole group of up to nine (9) CRN pathologists. The Sanyal dataset (Sanyal et al. 2021) was a relevant comparison to the studies included in AV/CV, as it was a subset of the REGENERATE phase 3 trial. However, AV and CV datasets contained more variability by including several other trials and samples coming from multiple central labs used to stain and scan (for CV, scans were collected from multiple original trial labs). Additionally, the Sanyal study involved comparing across two different panels of three pathologists (not focusing only on kappas between individuals) as for Davison or Kleiner or the individual to panel comparison as in AV/CV here. That being said, as discussed in the response for question 10, the kappas achieved by the panels in the Sanyal publication (inter-) were similar to the CRN kappas achieved between individuals in the literature, providing evidence that the gold standard panel workflow still experiences intra- and inter- variability, similar to that seen with expert individual CRN pathologists. Therefore, consistency and standardization are still challenges with the gold standard today, and AIM-NASH addresses some of those challenges.

12. The WKs provided in the CV validation study give an indication of agreement but do not allow an assessment of the type of disagreement. False positive and negative rates compared to the ground truth would help assessing whether the disagreement is 'balanced' or whether there is a systematic deviation (AI-assisted vs GT).

    PathAI Response:
    Please reference discussion in response #3, which describes the additional agreement tables per component score generated for each feature (Appendix A Table 4, Table 7, Table 10, Table 13, Table 16, Table 18, Table 20, Table 22, Table 5, Table 8, Table 11, Table 14, Table 17, Table 19, Table 21, Table 23, Table 6, Table 9, Table 12, Table 15) and for the binary composite clinical trial endpoints for AIM-NASH assist and IMR vs GT (Appendix A, **Table** 25, **Table 28**, **Table** 31 for overall agreement (OPA), positive percent agreement (PPA) and negative percent agreement (NPA) for both AIM-NASH and IMRs compared to GT. AIM-NASH achieves high levels of overall agreement and positive percent agreement being superior to average IMR for multiple measurements, with NPA being somewhat similar. Importantly, this ability to accurately and sensitively detect inclusion criteria populations and assess for endpoints will be applied reproducibly for AIM-NASH throughout enrollment periods and for follow-up timepoints, which should allow for more accurate capture of change in score over time compared to manual reads (individual or consensus).

13. Please elaborate on the clinical meaningfulness of the evaluation of reproducibility in the analytical validation which failed to demonstrate reproducibility when using different

sites/operators at the pre-defined 85% level, and whether this had any consequences on operating instructions and/or equipment use.

PathAI Response:
The AV study tested many aspects of reproducibility, including algorithm reproducibility when the algorithm was run multiple times on the same image (achieving 100% agreement, compared to lower intra-pathologist agreement demonstrated in the literature), algorithm reproducibility across scans from the same site but different days (called scanner repeatability), and algorithm reproducibility across scans from three different sites (called scanner reproducibility). Scanner repeatability is a relevant analysis in utilizing one central lab for a trial for scanning, as is often the case in NASH clinical trials. It is important to note that the AV dataset represented stained slides from multiple, original trial labs, including samples which were up to 5 years old, which provided additional pre-analytical variability as a challenging factor which could, theoretically, affect a tool's intra- and inter-site measurements.

For the reproducibility study, where scanning occurred across three external sites with different operators and scanners, the AIM-NASH inter-site reproducibility % agreement achieved in AV, which was approximately 85% or above, is higher than inter-pathologist agreements demonstrated in the literature (Davison showed overall % agreement for intra- and inter- reader) and within this study (Briefing Document Table 66).

Importantly, in terms of clinical trial impact, a post-hoc AIM-NASH AV reproducibility analysis of borderline fibrosis populations (F0+1, F4) as well as NAS >=4 inclusion criteria (where steatosis, inflammation, and ballooning scores contribute) was performed (Briefing Document Figure 35, Table 65), and AIM-NASH average agreement across the three sites was above 93% for all three measures.

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. It could be beneficial for the sponsor and lab(s) to use a control slide set, as is common in CAP/CLIA procedures, to ensure consistent output across scanners that will be used in studies. Nevertheless, the reproducibility achieved by the AIM-NASH tool alone in these studies, represents a significant improvement to that which has been achieved by expert manual readers thus far. Additionally, 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.

14. Please also elaborate on the usefulness of the overlay features of the AIM-NASH, referring to the failure of the overlay validation study to demonstrate success rates at the pre-defined

level for ballooning, as well as the evaluation of clinical utility of the overlays in the clinical validation study, which appeared rather unconvincing.

PathAI Response:
During the overlay validation study, all acceptance criteria were passed, except in the case of ballooning where it was narrowly missed. Upon examination, results from 2 out of 3 of the study pathologists' evaluations of the ballooning overlays were significantly higher than the acceptance criteria for both sensitivity and specificity measures. It appears as though the 3rd pathologist identified the presence of ballooning on significantly more frames, where the other two did not indicate presence. Additionally, it appears as though 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. This is to be expected, as the lack of an agreed upon definition of the ballooning feature is one of the largest challenges in scoring for NASH. Indeed, in the recent Brunt et al paper (Brunt et al. 2022), kappas across CRN experts on the presence and absence of ballooning was only 0.197 and, within one representative WSI, out of the approximately one hundred potential ballooned cells that were annotated, the entire group of pathologists could only agree unanimously on one cell as being ballooned. Therefore, we expect this variability, but the overall results for ballooning in AV/CV accuracy across all ballooning scores, together with the high repeatability/reproducibility results indicate that AIM-NASH can help to reduce this lack of standardization and consistency. Since the AIM-NASH-assisted reads in CV did indeed achieve superiority compared to unassisted reads, and since the pathologist has the choice to toggle on/off the overlays depending on preference, we believe the overlays to be useful overall in model interpretability and review by the pathologist.

Issues to be addressed in writing only by 19 October 2023

General Questions including context of use and utilisation of the tool:

15. Please comment on the recently published consensus change of nomenclature for NAFLD/NASH into MASLD/MASH and whether it has any implications for the potential labelling including the context of use statement.

    PathAI Response:
    Regarding the implications to the context of use, we do anticipate changing the naming references as appropriate.

16. In general, the role of the trial pathologist in the AIM-NASH in the NASH Clinical Trial workflow is supported. However, it needs to be clarified in detail what will be the steps if no consensus can be reached between the consensus pathologist and the trial pathologist within the Rejection Workflow.

    PathAI Response:

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.

17. No instructions/guidance for use could be found in the submission documents. The Applicant is asked to comment what information will be provided in the guidance document and if possible, provide the document itself. In addition, the applicant might consider incorporating explainability measures into the web-based platforms to improve interpretability of the results.

    PathAI Response:
    The AIM-NASH Instructions for Use are provided in Appendix D and these provide an overview of the tool, the overlays, CRN scores produced, accepting or rejecting the scores and report generation. These instructions are also available on the AISight Clinical Trials platform for the users to review at any time during the workflow.

    The instructions provided are for AIM-NASH version 1.5, in Briefing Document Appendix VIII, the AIM-NASH version 1.4 is the latest referenced. The change from v1.4 to v1.5 will include the removal of any exploratory features that were added in v1.4, but there will be no changes to the algorithm itself. Change assessment will be performed prior to changes and any necessary verification will be performed.

18. The applicant has applied strict rules for the recruitment of histopathologists. However, the evaluation exercise has been undertaken in the US only, and it remains to be fully demonstrated how it will be assured that sufficiently and similarly qualified labs and histopathologists will use the new tool in an European (or other region of the world) environment, considering that usually, NASH trials are conducted as global trials.

    PathAI Response:
    The pathologists utilizing this tool are recruited based on their qualification and experience. The experience criteria are applicable to any pathologist in any country and location (years of experience, number of NASH cases signed out, previous clinical trial reading experience and any NASH related publications). As defined in Briefing Document Section 4.8, for qualification, we require Board certification in pathology as evidenced by documentation of The American Board of Medical Specialties (ABMS) certification or equivalent certification in the country for which they are practicing. Therefore, we believe that this process is generalizable to any country in the world. In fact, we included a pathologist from the United Kingdom as one of the ground truth pathologists.

    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.

19. Please provide a “lay-language” description of the technical equipment used during development and proposed to be necessary when in final operation (e.g. AWS, ISMS and other features).

PathAI Response:
The following technical equipment were used during development and will be necessary when in final operation:

- Amazon Web Services (AWS): AWS is a commercial cloud computing provider. The AIM-NASH algorithm is run on servers provided by AWS.
- Amazon Simple Storage Service (S3): S3 is a cloud storage service provided by AWS. The inputs required to run the AIM-NASH algorithm are stored in S3 as well as the algorithm outputs displayed on the AISight Clinical Trials platform.
- Amazon Cognito: This is a AWS service that manages user authentication via single sign-on in order to login to the AISight Clinical Trials platform using external credentials.
- Convolutional neural networks: A specialized artificial neural network architecture, commonly used for image classification tasks.
- Graph neural networks: An artificial neural network architecture used to process data that can be represented as mathematical graphs (i.e., not restricted to images).
- Information security management system (ISMS): ISMS describes all processes and tools used to ensure data and information security in accordance with applicable laws.
- Kubernetes: Supports the continuity of the AISight Clinical Trials platform and the AIM-NASH algorithm workflows initiated by the platform.
- Postgres: Database used to store all the metadata uploaded to the AISight Clinical Trials platform.

20. The use of the different platforms during the CV and planned use in the real trial setting is not entirely clear. For example, it is not clear what role the OpenClinica eDC platform will play outside the validation. Please describe and characterise the “OpenClinica” electronic data platform used during the clinical validation study, which was used in the earlier studies for the data management of the glass slides only, but now seems to be part of the final data management algorithm.

PathAI Response:
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. OpenClinica was used to enter data for glass slide reads in the platform validation studies and data collection for AIM-NASH-assisted reads in the clinical validation study as 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). AISight Clinical Trials Platform now has the same data entry capabilities as the OpenClinica eDC platform and will be the only platform used for AIM-NASH in clinical trials.

21. Please discuss the potential to minimize inter-site variability of WSI scanning procedures, considering plans to evaluate different scanner models in a future exercise. Please further describe the post-hoc analysis for trial influence on the analysis of reproducibility.

    PathAI Response:
    Importantly, a mix of three trials was used to generate the reproducibility dataset, which was powered overall for study endpoints, but not powered for each trial.

    Please see relevant response for question 13 which discusses possible sources of variability in the study overall and implications/recommendations for prospective use once qualified. Additionally, the samples for each trial were stained in different labs and contained varying disease activity and drug efficacy (e.g. there was a cirrhotic ph2b trial included which did not meet its study endpoints and included follow-up timepoints where disease had advanced even further). Again, no pathologist review of algorithm is included in this AV study but would be important in a clinical trial use case to capture and resolve sample/stain/scan inadequacy which could impact score.

Model development

22. It is not clear if 'N/A samples' for fibrosis stage for trichrome image are samples with unknown fibrosis stage or these are non-NASH non-fibrotic samples. Please clarify.

    PathAI Response:
    In Briefing Document Table 12, the fibrosis samples with N/A are slides used for model development from trials for which fibrosis stage was not available (e.g., non-NASH samples).

23. It is not clear if the front-end is guaranteed to display whole slide images at their optimal size and resolution. Monitor sizes and resolution change, and perhaps physicians may use their phones to access the interface. Please comment or provide system specifications in the platform documentation.

    PathAI Response:
    The AISight Clinical Trials Platform is a cloud-based software that may be accessed from a networked workstation (either Microsoft Windows PC or Apple Macintosh) that is running a Google Chrome web browser. When using the platform, PathAI recommends the following minimum requirements:

Component	Minimum Requirements
Workstation	Microsoft Windows PC: • Operating System: Windows , 10 or later • Processor: Intel Pentium 4, AMD Athlon 64 or later SSE2-capable processor Apple Macintosh: • Operating System: OS X Yosemite 10.10 or later
Web Browser	Google Chrome v.81 or later
Display	• 27" monitor • Pixel resolution: 2560w x 1440h
Component	**Minimum Requirements**
Workstation	Microsoft Windows PC: • Operating System: Windows , 10 or later • Processor: Intel Pentium 4, AMD Athlon 64 or later SSE2-capable processor Apple Macintosh: • Operating System: OS X Yosemite 10.10 or later
Web Browser	Google Chrome v.81 or later
Display	• 27" monitor • Pixel resolution: 2560w x 1440h

The platform does not support mobile phone devices.

This information is provided to pathologists in the AISight Clinical Trials platform IFU, along with the AIM-NASH IFU (Appendix D), during training.

24. The annotation process appears to be performed quite thoroughly based on the information provided by the Applicant. However, the Applicant states that *"During initial rounds for each contributor, roughly 10% of annotations were randomly selected and reviewed for quality by PathAI pathologists. If an annotator had a large number of poor-quality (as defined by incorrect identification of substances by internal expert pathologists) annotations for a particular substance, their annotations for that substance were removed from the dataset."* Based on this, it seems that PathAI pathologists could overrule the

annotation of the external pathologists, that defeats the purpose of collecting the data from a large sample of pathologists to prevent model overfitting. Please comment why PathAI pathologists appear to have had preference over external ones and implications of this choice.

PathAI Response:
Some annotations were collected from qualified surgical pathologists (not necessarily NASH experts), where appropriate (e.g., general tissue architecture or areas of fibrosis) and, to be diligent, our internal NASH expert pathologists wanted to perform an extra QC check to ensure that there were no large issues with a particular pathologists' annotations. Internal pathologists though, did not, for example, reject ballooning annotations form a particular pathologist if they were in the range of what pathologists identify as potential ballooned cells (e.g. they were generous with their QC of these types of features and were more so looking for potential clearly incorrect annotations (e.g. not circling a feature region with enough precision) or mistakes affecting many of a particular pathologist's annotations across slides. In other words, our NASH expert pathologists were QCing for general quality, and we don't believe there's a large risk of over-fitting. In addition, we believe that our extensive AV, overlay, and CV studies demonstrate this is not an issue.

25. The Applicant indicated that pathologists involved in NASH CRN models 4 and 5 were presented with the WSI alone and were unable to view any model outputs or scores from other annotators. It is, however, not entirely clear if this means that overlay function was also not available for them to use. Please clarify.

PathAI Response:
No overlays were presented to the pathologists in training the NASH CRN models 4 and 5, WSIs only. The aim was to collect slide levels scores manually from pathologists as a separate input to the GNN models, in addition to the model's prediction heatmaps (overlays).

26. Integrated analytical verification: Please describe the objectives and analyses undertaken of this part of the validation exercise more clearly and clarify the discrepancies between the submitted study report and the summary included in the Briefing Document (e.g., regarding which magnification was used). Please also discuss the use of 20 slides only and whether this allows drawing robust conclusions, and why the full hold-out test set was used.

PathAI Response:
The purpose of integrated analytical verification (IAV) is to ensure that the locked algorithm model yields the same results with defined tolerance on the AISight Clinical Trials platform as it did in the development environment where standalone analytical verification (SAV) was performed, using a subset of the same held-out test set, and to

verify that the platform integration, functional user workflow and reporting requirements defined for the algorithm product are met.

The 20 slides used in IAV were all scanned at 40x magnification. The IAV also included software verification for 2 slides scanned on 20x, however 20x magnification is not part of the AIM-NASH specifications currently.

Since SAV was powered for the accuracy endpoint, no changes to the algorithm were made and the goal of IAV was to verify platform integration, a representative subset of 20 slides was chosen to confirm that no algorithm outputs were affected when the algorithm was integrated onto the platform. For example, NASH CRN scores provided by the algorithm for these 20 slides were required to be identical before and after integration into the CTS platform. In addition, functional software requirements are tested during IAV, and we determined that 20 cases were sufficient to confirm requirements were met.

27. According to the model development scheme provided, before testing the held-out sets, the system was supposed to be tested on internal dataset. No information could be found on this step, including acceptance criteria, details on the dataset and results. It remains unclear whether this step has erroneously not been included. Please comment on this and provide required data.

    PathAI Response:
    Internal verification with a held-out test set, designated as "the internal test set," was performed to evaluate model performance on a dataset from a phase 2b NASH trial evaluating a novel insulin sensitizer. 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 weighted kappa statistics. For reference, pairwise pathologist agreements were computed. For all histologic features, agreement of AIM-NASH read-outs with consensus reads was greater than the pairwise agreement between pathologists performing manual reads (Appendix A Table 47). Even though the acceptance criteria was not documented prior to the internal testing, the results met our standard acceptance criteria of 0.1 non-inferiority. Additionally, the algorithm was tested using the SAV held-out test set (derived from separate trials) where it was evaluated against pre-defined acceptance criteria, before moving on to validation studies.

28. According to page 61 of the briefing document, changes are "aligned with relevant regulations and internationally recognized consensus standards". The applicant should provide references.

    PathAI Response:
    Per PathAI's Change Management Procedure, the references are the following:

- 21 CFR 820 Quality System Regulation
- ISO 13485:2016 Medical Devices – Quality Management Systems – Requirements for regulatory purposes
- IEC 62304:2015 Medical device software – Software life cycle processes
- ICH GCP E6 R2 Good Clinical Practice Guidelines

Validation of the AISight Trials Platform or Translational platform

29. Validation of the AISight clinical trials platform: Please clarify the number of cases and slides used during this exercise from the different sources and how NASH features were evaluated based on the inclusion of 48 cases from the Precision of Medicine source with presumed non-NASH disease, or whether these were also cases of NASH.

    PathAI Response:
    The AISight Clinical Trials platform validation included 18 cases (outlined in Briefing Document section 4.3.4) from Precision of Medicine. These cases were all non-NASH samples that included normal liver and other liver indications (e.g. 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.

30. As understood, the GT and IMRs were performed on AISight Translational platform and AIM-NASH was run on the AISight Clinical Trials platform during validation studies. The Applicant is asked to comment whether only AISight Clinical Trials platform will be used in the real trial setting. Also, since some differences were noted in the use of two platforms, the Applicant needs to discuss the implications of using both of them in the same study.

    PathAI Response:
    AISight Clinical Trials platform will be the only platform utilized in the prospective NASH trial setting. The GT and IMR reads were performed on the AISight Translational platform because the stage of the AISight Clinical Trials Platform development at the time of these studies. The differences in the validation data between different platforms is confounded in the differences in the reading panels as also demonstrated in (Sanyal et al. 2021), showing that inter-panel agreement is similar to individual pathologist agreement between CRN pathologists. As both platforms were validated for manual reads of digital images for NASH and only manual reads were performed on the Translational platform, we do not believe there are any implications of using both platforms in the same study.

31. In general, the validation results on the Translational platform appear to perform worse compared to the Clinical Trial platform for the secondary endpoints. The Applicant argues that the variation can likely be attributed to the utilization of different panel readers for each validation study and the known variability between readers, as demonstrated in the literature. This argument is not fully understood, as the same pathologists were involved in

both validation studies. Please clarify and discuss if other differences between two platforms could have accounted for observed differences in the validation results.

PathAI Response:
In the platform validation studies, only the ground truth panel was common between the two studies, however, the ground truth panel only read slides on glass and did not utilize either of the platforms. The study pathologists were platform specific and there was no overlap between the 2 panels. The differences in absolute kappa values between study groups are within the expected range of variability demonstrated by different NASH experts.

We don't anticipate differences between the platforms themselves could account for observed differences in the validation results as the platforms met acceptance criteria and are both validated for NASH.

32. Six fewer slides were used to validate the clinical trials platform than the translational platform. The applicant is invited to comment on the difference.

PathAI Response:
159 cases were enrolled in the AISight Clinical Trial Platform study and 156 cases were enrolled in the AISight Translational Platform study. For the AISight Translational Platform validation study, 3 cases that were deemed evaluable by the GT panel pathologists were inadvertently not sent out to the AISight Translational Platform study pathologists for evaluation. This is also documented in the study report and deviation log.

33. The applicant should elaborate on how vulnerabilities in Identity and Access Management are addressed (eg. Do they use Multi-Factor Authentication?).

PathAI Response:
PathAI has a robust vulnerability management program in place to detect, monitor, and remediate all applicable vulnerabilities including those targeting identity and access management solutions in our environment. In addition to compensating controls such as multi-factor authentication and the use of Single Sign-On (SSO), PathAI enforces stringent device posture assessments prior to allowing access to applications and actively blocks high-risk countries.

Analytical validation

34. When looking at differences between trials, WKs AI-assisted vs. GT were similar between different trials, aside for inflammation and especially fibrosis score where WKs were (much) lower in Falcon 1 and 2 trials compared to Regenerate trial. Please comment on possible reasons.

PathAI Response:

In analytical validation, the datasets were not powered for overarching conclusions for each individual trial level. As demonstrated in Briefing Document Table 42, for the two histological features mentioned (lobular inflammation and fibrosis) the sample size for 90% power is 540 cases and 420 cases respectively. Therefore, this study was powered for overall accuracy, which was met, but not powered per individual trial. Performance of the AIM-NASH-assisted generally trended with performance of IMR for inflammation and fibrosis in the trials cited, further supporting the use of AIM-NASH in these settings.

Clinical validation

35. If understood correctly, IMR in AV and CV were obtained in the same manner. In this respect, WKs of IMR vs GT were similar in AV and CV studies, except inflammation, WKs of which were lower in CV validation (0.297) compared to the AV (0.402). The Applicant is asked to comment on these differences, as the lower WKs for IMR vs GT in CV study resulted in the superiority claims for Lobular Inflammation.

    PathAI Response:
    In addition to the dataset differences below, which could potentially impact IMR performance (both due to dataset composition and due to additional case number in CV), the overall % agreement analyses that were performed here (Appendix A Table 1, Table 2) for AV and CV demonstrate that the IMR-GT OPA does not change substantially from AV to CV (56.4% vs. 54.1%), whereas the AIM-NASH -GT agreement increases more substantially from AV to CV (57% vs. 63.5%); therefore, the superiority achieved in CV for inflammation is not solely due to a decreased IMR performance.

    AV/CV Dataset Differences: CV had a significantly larger overall sample size (over twice as many biopsies), and inclusion of an additional drug candidate (semaglutide; those glass slides were not available for AV, but WSIs were available for CV). Therefore, the accuracy results from CV may be thought of as more representative of the NASH trial landscape. Additionally, AV samples were evaluated using WSIs prospectively scanned at one external lab. CV samples were read using the original trial WSIs scanned across several labs. It is promising to see that AIM-NASH demonstrated higher agreement with GT for inflammation from AV to CV, even with the increased variability in drug candidates included, number of scanning sites, and largely increased sample size in CV, a more robust and representative dataset.

36. Clinical and Analytical Validation: Please display all the results of the "categorical" evaluation of accuracy for the "inclusion criteria" and endpoint categories (such as F2/3 vs. other, NAS ≥4 vs. <4, and NASH resolution), the evaluation of statistical significance, and elaborate on their clinical implications (out of the analytical and clinical validation).

    PathAI Response:
    Please see tables (Table 24, Table 27, Table 30, Table 25, Table 28, Table 31, Table 26, Table 29, and Table 32) in the Appendix A for these requested analyses.

As discussed in the response for #3, overall 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 included here (Appendix A Table 25, Table 26) 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 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
2. the superior repeatability/reproducibility of AIM-NASH compared to manual pathology (intra- and inter-) should result in more accurate, standardized and consistent enrollment and detection of steatosis grade change or fibrosis stage change for a patient in a trial

Please see supportive efficacy analyses in the case studies in Appendix C, which add strong supportive evidence across trials and drug candidates of the robust performance of AIM-NASH in measuring trial endpoints from manually enrolled NASH trials.

37. Please clarify the discrepant description for the inclusion of data sources (studies) in the clinical validation study, in the study protocol and the final evaluation conducted.

PathAI Response:
The data sources described in the clinical validation study protocol are 4 datasets that were potentially available for the study at the time of protocol writing with an estimated number of patient samples available based on study enrollment. The data sources described in Section 4.7.4 in the briefing document, should have matched the estimated numbers in the protocol but were erroneously copied. In the final study report, we specifically describe all available cases that were transferred to us for the use of the study by the trial sponsor. One of the phase 2 pegbelfermin (Falcon 1) studies was not included in the clinical validation study. Overall, for Falcon 2, we received and enrolled 154 patients with 284 unique samples. For Regenerate, we enrolled 470 unique subjects with 694 samples, representative of screened and enrolled populations. For Novo Nordisk study, the subject information was not available, however 523 unique samples from the enrolled trial population (both baseline and follow-up time points) were available and all were enrolled.

38. The big difference in WKs between the analysis of F0+F1 compared to F4 are noted (in a range of 0.5 for F0+F1 and in a range of 0.7 for F4), which were not seen during the AV test. Please comment on possible reasons.

PathAI Response:
Even though the kappas in these sub-populations were somewhat lower for the CV dataset than they were for the AV dataset, for both analyses, the AIM-NASH/AIM-NASH-assisted reads met the non-inferiority criteria and were all acceptable compared to IMRs. Regardless of any differences in the datasets, AIM-NASH performed equivalently to the IMRs, meeting study endpoints.

Additionally, see relevant responses for questions 34 and 35 for possible reasons why kappas achieved (by IMRs, GT, and/or AIM-NASH) could be somewhat different in AV vs. CV study populations. Using the above n breakdown of fibrosis scores represented, and knowing the dataset differences, the range of kappas could be due to differences in F0/1 relative proportions from AV to CV.

39. In general, WKs AI-assisted vs GT were similar between different trials, aside for fibrosis score where WKs were much lower in BMS trial compared to Intercept and NovoNordisk. Please comment on possible reasons.

PathAI Response:
Similar to the response in question 34, the available BMS samples (and overall enrolled n) is lower than for the other two trials in CV and therefore not as well powered to make substantial conclusions when comparing to the other trial agreement rates for particular features, given the inherent variability in manual reads. 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.

Appendix A

Appendix A.1

Table 1: Overall AIM-NASH and IMR Agreement with Ground Truth for NASH Features (AV)

				Agreement Evaluation[1]		AIM-NASH - Average IMR	
Feature	**Modality**	**Number of reads**	**Number of unique cases**	**%**	**95% CI[2]**	**Difference (95% CI)[2]**	**P-value[2,3]**
Steatosis	AIM-NASH vs GT	597	597	66.5%	62.9%, 69.8%	-5.4% (-10.5%, -0.8%)	0.030
	IMR vs GT	1902	597	71.9%	68.3%, 75.1%		
Lobular inflammation	AIM-NASH vs GT	593	593	57.0%	50.8%, 62.4%	0.5% (-5.3%, 6.2%)	0.819
	IMR vs GT	1886	593	56.4%	53.1%, 59.3%		
Hepatocellular ballooning	AIM-NASH vs GT	597	597	71.0%	67.8%, 73.7%	12.4% (7.5%, 17.2%)	<0.001
	IMR vs GT	1902	597	58.6%	54.7%, 62.4%		
Fibrosis	AIM-NASH vs GT	583	583	60.9%	55.4%, 65.3%	1.8% (-4.4%, 7.9%)	0.446
	IMR vs GT	1870	583	59.1%	55.0%, 63.6%		

[1] 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 2: Overall AIM-NASH-Assisted and IMR Agreement with Ground Truth for NASH Features (CV)

				Agreement Evaluation[1]		AIM-NASH-Assisted - Average IMR	
Feature	**Modality**	**Number of Reads**	**Number of Unique Cases**	**%**	**95% CI[2]**	**Difference (95% CI)[2]**	**P-value[2,3]**
Steatosis	AIM-NASH vs GT	1467	1467	71.1%	69.2%, 73.9%	1.6% (-0.9%, 4.7%)	0.195
	IMR vs GT	4521	1481	69.5%	67.6%, 71.5%		
Lobular inflammation	AIM-NASH vs GT	1465	1465	63.5%	60.7%, 66.0%	9.4% (6.1%, 12.3%)	<0.001
	IMR vs GT	4509	1478	54.1%	52.4%, 56.0%		
Hepatocellular ballooning	AIM-NASH vs GT	1465	1465	68.8%	65.3%, 71.3%	10.9% (7.6%, 13.7%)	<0.001
	IMR vs GT	4506	1476	57.9%	56.0%, 59.9%		
Fibrosis	AIM-NASH vs GT	1429	1429	65.2%	62.3%, 67.5%	2.7% (-0.7%, 5.6%)	0.122
	IMR vs GT	4506	1453	62.5%	60.5%, 64.4%		

[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 3: Overall AIM-NASH and IMR Agreement with Ground Truth for NASH Features (CV)

				Agreement Evaluation[1]		AIM-NASH - Average IMR	
Feature	**Modality**	**Number of reads**	**Number of unique cases**	**%**	**95% CI[2]**	**Difference (95% CI)[2]**	**P-value[2,3]**

Steatosis	AIM-NASH vs GT	1480	1480	71.0%	68.8%, 73.3%	1.5% (-1.5%, 4.6%)	0.340
	IMR vs GT	4524	1481	69.5%	67.6%, 71.5%		
Lobular inflammation	AIM-NASH vs GT	1477	1477	63.3%	60.9%, 65.7%	9.2% (6.0%, 12.1%)	<0.001
	IMR vs GT	4512	1478	54.1%	52.4%, 56.0%		
Hepatocellular ballooning	AIM-NASH vs GT	1475	1475	68.7%	66.4%, 71.0%	10.8% (7.7%, 13.8%)	<0.001
	IMR vs GT	4509	1476	57.9%	56.0%, 59.9%		
Fibrosis	AIM-NASH vs GT	1452	1452	64.2%	61.6%, 66.6%	1.7% (-1.5%, 4.8%)	0.296
	IMR vs GT	4509	1453	62.5%	60.5%, 64.4%		

[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.

Appendix A.2

Table 4: AIM-NASH and IMR Agreement with Ground Truth per Score for Steatosis (AV)

Steatosis			AIM-NASH					Average IMR			
	Grade	N	0	1	2	3	N	0	1	2	3
Ground Truth	0	93	61.29 %	38.71 %	0.00%	0.00%	93	63.85 %	35.87 %	2.22%	0.00%
	1	229	5.68%	66.38 %	26.2%	1.75%	229	5.37%	83.30 %	13.67 %	2.77%
	2	184	0.00%	8.70%	57.61 %	33.7%	184	0.00%	12.93 %	59.22 %	29.46%
	3	91	0.00%	0.00%	9.89%	90.11%	91	0.00%	2.27%	23.45 %	79.19%

Table 5: AIM-NASH-Assisted and IMR Agreement with Ground Truth per Score for Steatosis (CV)

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 6: AIM-NASH and IMR Agreement with Ground Truth per Score for Steatosis (CV)

Steatosis			AIM-NASH					Average IMR			
	Grade	N	0	1	2	3	N	0	1	2	3
Ground Truth	0	124	42.74%	57.26%	0.00%	0.00%	124	62.43%	37.57%	0.00%	0.00%
	1	644	2.33%	72.83%	23.76%	1.09%	644	3.99%	79.21%	16.51%	1.25%
	2	511	0.00%	8.41%	68.88%	22.70%	511	0.00%	16.56%	54.81%	28.63%
	3	201	0.00%	0.00%	11.94%	88.06%	202	0.00%	1.65%	19.33%	79.85%

Table 7: AIM-NASH Agreement with Ground Truth per Score for Lobular Inflammation (AV)

Inflammation			AIM-NASH					Average IMR			
	Grade	N	0	1	2	3	N	0	1	2	3
Ground Truth	0	15	86.67 %	13.33 %	0.00%	0.00%	15	61.01 %	49.55 %	18.33 %	0.00%
	1	339	25.96 %	60.18 %	13.86%	0.00%	339	14.65 %	59.91 %	18.67 %	11.47%
	2	207	3.86%	36.23 %	53.62%	6.28%	207	2.81%	29.21 %	49.02 %	38.98%
	3	32	0.00%	3.12%	65.62%	31.25 %	32	4.76%	22.74 %	39.6%	71.18%

Table 8: AIM-NASH-Assisted and IMR Agreement with Ground Truth per Score for Lobular Inflammation (CV)

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 9: AIM-NASH and IMR Agreement with Ground Truth per Score for Lobular Inflammation (CV)

Inflammation			AIM-NASH					Average IMR			
	Grade	N	0	1	2	3	N	0	1	2	3
Ground Truth	0	12	66.67%	33.33%	0.00%	0.00%	12	54.46%	54.22%	25.00%	14.29%
	1	937	8.78%	61.37%	19.74%	0.11%	938	12.54%	58.59%	22.78%	11.91%
	2	498	1.00%	25.10%	69.68%	4.22%	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 10: AIM-NASH and IMR Agreement with Ground Truth per Score for Hepatocellular Ballooning (AV)

Ballooning			AIM-NASH			Average IMR			
	Grade	N	0	1	2	N	0	1	2
Ground Truth	0	76	73.68 %	25.00 %	1.32%	76	62.88 %	43.06%	12.86%
	1	295	13.56 %	62.37 %	24.07 %	295	28.78 %	55.32%	19.49%
	2	226	0.88%	17.70 %	81.42 %	226	7.19%	33.46%	62.04%

Table 11: AIM-NASH-Assisted and IMR Agreement with Ground Truth per Score for Hepatocellular Ballooning (CV)

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 12: AIM-NASH and IMR Agreement with Ground Truth per Score for Hepatocellular Ballooning (CV)

Ballooning			AIM-NASH			Average IMR			
	Grade	N	0	1	2	N	0	1	2
Ground Truth	0	164	75.61%	23.78%	0.61%	164	62.70%	39.03%	5.05%
	1	693	16.16%	57.58%	26.26%	694	24.95%	56.5%	18.55%
	2	618	1.94%	18.77%	79.29%	618	5.24%	35.46%	59.96%

Table 13: AIM-NASH and IMR Agreement with Ground Truth per Score for Fibrosis (AV)

Fibrosis			AIM-NASH						Average IMR				
	Stage	N	0	1	2	3	4	N	0	1	2	3	4
Ground Truth	0	21	71.43 %	28.57 %	0.00%	0.00%	0.00%	21	44.55 %	57.16 %	10.85 %	0.00%	0.00%
	1	78	25.64 %	34.62 %	29.49 %	10.26 %	0.00%	78	12.35 %	57.11 %	26.92 %	5.51%	3.23%
	2	110	7.27%	11.82 %	31.82 %	47.27 %	1.82%	110	12.94 %	27.28 %	44.73 %	24.85 %	3.14%
	3	197	0.00%	1.02%	8.63%	73.10 %	17.26 %	197	0.00%	2.91%	19.92 %	63.37 %	16.6%
	4	177	0.00%	0.00%	0.56%	23.73 %	75.71 %	177	0.00%	0.00%	0.00%	33.97 %	66.03%

Table 14: AIM-NASH-Assisted and IMR Agreement with Ground Truth per Score for Fibrosis (CV)

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%

Table 15: AIM-NASH and IMR Agreement with Ground Truth per Score for Fibrosis (CV)

Fibrosis			AIM-NASH						Average IMR				
	Stage	N	0	1	2	3	4	N	0	1	2	3	4
Ground Truth	0	13	76.92%	7.69%	15.38 %	0.00%	0.00 %	13	60.65%	37.35%	26.67%	0.00%	0.00%
	1	175	20.57 %	33.14%	34.86 %	10.86 %	0.57%	175	18.96%	60.01%	18.19%	4.28%	1.28%

	2	396	3.03%	17.42%	44.7%	34.09 %	0.76 %	396	4.78%	31.15%	49.3%	15.58%	1.52%
	3	564	0.00%	1.42%	9.75%	79.08%	9.75%	565	0.44%	4.02%	18.67%	66.62%	12.6%
	4	304	0.00%	0.00%	0.66%	20.07%	79.28%	304	0.00%	0.00%	6.67%	24.23%	74.94%

Table 16: Overall Agreement for Steatosis for IMRs (AV)[1]

Steatosis		**Ground Truth**			
	Grade	**0**	**1**	**2**	**3**
Reader 1	0	88.10%	13.33%	0.00%	0.00%
	1	11.90%	80.95%	16.67%	0.00%
	2	0.00%	4.76%	42.86%	18.60%
	3	0.00%	0.95%	40.48%	81.40%
	N (GT)	42	105	84	43
Reader 2	0	45.00%	0.00%	0.00%	0.00%
	1	55.00%	90.48%	15.52%	0.00%
	2	0.00%	9.52%	82.76%	79.31%
	3	0.00%	0.00%	1.72%	20.69%
	N (GT)	40	84	58	29
Reader 3	0	69.09%	1.83%	0.00%	0.00%
	1	30.91%	83.49%	11.11%	0.00%
	2	0.00%	14.68%	78.89%	15.91%
	3	0.00%	0.00%	10.00%	84.09%
	N (GT)	55	109	90	44
Reader 4	0	64.15%	1.79%	0.00%	0.00%
	1	35.85%	68.75%	5.15%	2.27%
	2	0.00%	26.79%	36.08%	2.27%
	3	0.00%	2.68%	58.76%	95.45%
	N (GT)	53	79	97	44
Reader 5	0	96.15%	8.62%	0.00%	0.00%
	1	3.85%	84.48%	19.39%	0.00%
	2	0.00%	6.90%	69.39%	25.00%
	3	0.00%	0.00%	11.22%	75.00%
	N (GT)	52	116	98	48
Reader 6	0	40.00%	0.00%	0.00%	0.00%
	1	57.78%	77.48%	12.50%	0.00%
	2	2.22%	18.92%	36.36%	11.54%
	3	0.00%	3.60%	51.14%	88.46%
	N (GT)	45	111	88	52
Reader 7	0	75.0%	1.28%	0.00%	0.00%
	1	25.0%	80.77%	10.17%	0.00%
	2	0.00%	14.1%	47.46%	11.54%
	3	0.00%	3.85%	42.37%	88.46%
	N (GT)	24	78	59	26

[1] Only IMRs who read at least 10 cases were included.

Table 17: Overall Agreement for Steatosis for IMRs (CV)[1]

Steatosis		Ground Truth			
	Grade	**0**	**1**	**2**	**3**
Reader 1	0	82.26%	4.90%	0.00%	0.00%
	1	17.74%	83.57%	17.07%	0.88%
	2	0.00%	10.09%	36.18%	4.42%
	3	0.00%	1.44%	46.75%	94.69%
	N (GT)	62	347	246	113
Reader 2	0	68.12%	2.00%	0.00%	0.00%
	1	31.88%	89.6%	32.54%	2.47%
	2	0.00%	8.40%	66.99%	61.73%
	3	0.00%	0.00%	0.48%	35.80%
	N (GT)	69	250	209	81
Reader 3	0	68.57%	2.76%	0.00%	0.00%
	1	31.43%	80.31%	13.3%	0.00%
	2	0.00%	16.54%	75.53%	23.44%
	3	0.00%	0.39%	11.17%	76.56%
	N (GT)	35	254	188	64
Reader 4	0	9.52%	0.00%	0.00%	0.00%
	1	90.48%	83.83%	15.29%	0.00%
	2	0.00%	16.17%	63.06%	25.00%
	3	0.00%	0.00%	21.66%	75.00%
	N (GT)	42	167	157	64
Reader 5	0	67.92%	3.94%	0.00%	0.00%
	1	32.08%	73.12%	7.21%	1.27%
	2	0.00%	22.22%	46.15%	1.27%
	3	0.00%	0.72%	46.63%	97.47%
	N (GT)	53	279	208	79
Reader 6	0	81.03%	6.48%	0.00%	0.00%
	1	18.97%	66.76%	7.84%	0.00%
	2	0.00%	25.63%	71.64%	8.65%
	3	0.00%	1.13%	20.52%	91.35%
	N (GT)	58	355	268	104
Reader 7	0	61.97%	4.41%	0.00%	0.00%
	1	38.03%	77.21%	22.58%	1.96%
	2	0.00%	15.81%	45.56%	17.65%
	3	0.00%	2.57%	31.85%	80.39%

	N (GT)	71	272	248	102
Reader 8	0	60.00%	3.45%	0.00%	0.00%
	1	40.0%	79.31%	16.67%	0.00%
	2	0.00%	17.24%	33.33%	12.50%
	3	0.00%	0.00%	50.00%	87.50%
	N (GT)	10	29	24	16

[1] Only IMRs who read at least 10 cases were included.

Table 18: Overall Agreement for Lobular Inflammation for IMRs (AV)[1]

Inflammation		Ground Truth			
	Grade	**0**	**1**	**2**	**3**
Reader 1	0	55.56%	20.59%	1.23%	0.00%
	1	44.44%	49.41%	22.22%	0.00%
	2	0.00%	28.24%	48.15%	18.18%
	3	0.00%	1.76%	28.40%	81.82%
	N (GT)	9	170	81	11
Reader 2	0	0.00%	6.25%	0.00%	0.00%
	1	100.00%	78.12%	33.82%	10.00%
	2	0.00%	14.84%	66.18%	40.00%
	3	0.00%	0.78%	0.00%	50.00%
	N (GT)	3	128	68	10
Reader 3	0	33.33%	2.03%	0.00%	0.00%
	1	50.00%	33.78%	9.84%	4.76%
	2	16.67%	47.97%	45.08%	23.81%
	3	0.00%	16.22%	45.08%	71.43%
	N (GT)	6	148	122	21
Reader 4	0	60.00%	22.01%	1.68%	0.00%
	1	20.00%	10.06%	4.20%	0.00%
	2	20.00%	20.75%	5.04%	0.00%
	3	0.00%	47.17%	89.08%	100.0%
	N (GT)	5	159	119	21
Reader 5	0	100.00%	25.32%	5.51%	4.76%
	1	0.00%	62.03%	29.13%	4.76%
	2	0.00%	12.03%	38.58%	9.52%
	3	0.00%	0.63%	26.77%	80.95%
	N (GT)	6	158	127	21
Reader 6	0	60.00%	20.67%	0.00%	0.00%
	1	40.00%	63.13%	34.44%	0.00%
	2	0.00%	13.97%	60.0%	57.14%
	3	0.00%	2.23%	5.56%	42.86%
	N (GT)	10	179	90	14

Reader 7	0	57.14%	5.69%	0.00%	0.00%
	1	42.86%	91.87%	70.83%	71.43%
	2	0.00%	2.44%	29.17%	28.57%
	3	0.00%	0.00%	0.00%	0.00%
	N (GT)	7	123	48	7

[1] Only IMRs who read at least 10 cases were included.

Table 19: Overall Agreement for Lobular Inflammation for IMRs (CV)[1]

Inflammation		**Ground Truth**			
	Grade	**0**	**1**	**2**	**3**
Reader 1	0	25.00%	7.48%	0.36%	0.00%
	1	50.00%	48.50%	17.33%	6.25%
	2	25.00%	37.39%	52.35%	43.75%
	3	0.00%	6.62%	29.96%	50.00%
	N (GT)	4	468	277	16
Reader 2	0	20.00%	7.29%	0.00%	0.00%
	1	80.00%	68.59%	27.08%	23.08%
	2	0.00%	23.87%	70.83%	46.15%
	3	0.00%	0.25%	2.08%	30.77%
	N (GT)	5	398	192	13
Reader 3	0	33.33%	2.52%	0.00%	0.00%
	1	66.67%	27.36%	5.39%	0.00%
	2	0.00%	42.14%	42.16%	18.75%
	3	0.00%	27.99%	52.45%	81.25%
	N (GT)	3	318	204	16
Reader 4	0	0.00%	5.06%	0.00%	0.00%
	1	100.0%	81.01%	44.76%	0.00%
	2	0.00%	13.92%	55.24%	66.67%
	3	0.00%	0.00%	0.00%	33.33%
	N (GT)	5	316	105	3
Reader 5	0	71.43%	34.75%	20.18%	6.25%
	1	14.29%	16.98%	15.14%	0.00%
	2	0.00%	15.12%	9.17%	6.25%
	3	14.29%	33.16%	55.5%	87.5%
	N (GT)	7	377	218	16
Reader 6	0	71.43%	22.83%	1.49%	0.00%
	1	28.57%	64.24%	39.93%	0.00%
	2	0.00%	11.52%	38.43%	14.29%
	3	0.00%	1.41%	20.15%	85.71%
	N (GT)	7	495	268	14

Reader 7	0	60.00%	14.61%	0.43%	0.00%
	1	40.0%	67.81%	31.33%	0.00%
	2	0.00%	15.53%	60.52%	64.29%
	3	0.00%	2.05%	7.73%	35.71%
	N (GT)	5	438	233	14
Reader 8	0	100.0%	5.77%	0.00%	0.00%
	1	0.00%	94.23%	81.82%	50.0%
	2	0.00%	0.00%	18.18%	50.0%
	3	0.00%	0.00%	0.00%	0.00%
	N (GT)	1	52	22	2

[1] Only IMRs who read at least 10 cases were included.

Table 20: Overall Agreement for Hepatocellular Ballooning for IMRs (AV)[1]

Ballooning		Ground Truth		
	Grade	0	1	2
Reader 1	0	65.96%	20.92%	6.76%
	1	31.91%	45.75%	18.92%
	2	2.13%	33.33%	74.32%
	N (GT)	47	153	74
Reader 2	0	30.00%	2.65%	0.00%
	1	70.00%	87.61%	42.65%
	2	0.00%	9.73%	57.35%
	N (GT)	30	113	68
Reader 3	0	40.62%	6.20%	1.46%
	1	56.25%	86.82%	48.91%
	2	3.12%	6.98%	49.64%
	N (GT)	32	129	137
Reader 4	0	58.62%	11.11%	1.41%
	1	41.38%	74.07%	32.39%
	2	0.00%	14.81%	66.20%
	N (GT)	29	135	142
Reader 5	0	100.00%	61.48%	13.61%
	1	0.00%	33.33%	36.73%
	2	0.00%	5.19%	49.66%
	N (GT)	32	135	147
Reader 6	0	74.51%	20.37%	0.00%
	1	25.49%	67.28%	34.94%
	2	0.00%	12.35%	65.06%
	N (GT)	51	162	83
Reader 7	0	100.00%	78.72%	12.70%
	1	0.00%	19.15%	36.51%

	2	0.00%	2.13%	50.79%
	N (GT)	30	94	63

[1] Only IMRs who read at least 10 cases were included.

Table 21: Overall Agreement for Hepatocellular Ballooning for IMRs (CV)[1]

Ballooning		Ground Truth		
	Grade	**0**	**1**	**2**
Reader 1	0	64.38%	15.22%	7.12%
	1	31.51%	45.65%	17.03%
	2	4.11%	39.13%	75.85%
	N (GT)	73	368	323
Reader 2	0	40.74%	3.01%	0.44%
	1	58.02%	79.26%	29.07%
	2	1.23%	17.73%	70.48%
	N (GT)	81	299	227
Reader 3	0	44.74%	2.98%	0.37%
	1	55.26%	87.23%	56.93%
	2	0.00%	9.79%	42.70%
	N (GT)	38	235	267
Reader 4	0	19.74%	0.51%	0%
	1	63.16%	41.54%	10.13%
	2	17.11%	57.95%	89.87%
	N (GT)	76	195	158
Reader 5	0	60.94%	14.55%	0.72%
	1	37.5%	74.55%	37.18%
	2	1.56%	10.91%	62.09%
	N (GT)	64	275	277
Reader 6	0	91.25%	55.40%	10.00%
	1	7.50%	39.20%	39.14%
	2	1.25%	5.40%	50.86%
	N (GT)	80	352	350
Reader 7	0	79.78%	31.85%	3.75%
	1	20.22%	62.8%	51.31%
	2	0.00%	5.36%	44.94%
	N (GT)	89	336	267
Reader 8	0	100.0%	76.09%	14.29%
	1	0.00%	21.74%	42.86%
	2	0.00%	2.17%	42.86%
	N (GT)	12	46	21

[1] Only IMRs who read at least 10 cases were included.

Table 22: Overall Agreement for Fibrosis for IMRs (AV)[1]

Fibrosis		Ground Truth				
	Grade	**0**	**1**	**2**	**3**	**4**
Reader 1	0	57.14%	10.20%	4.65%	0.00%	0.00%
	1	35.71%	65.31%	25.58%	3.75%	0.00%
	2	7.14%	22.45%	55.81%	17.5%	0.00%
	3	0.00%	2.04%	13.95%	57.5%	24.05%
	4	0.00%	0.00%	0.00%	21.25%	75.95%
	N (GT)	14	49	43	80	79
Reader 2	0	33.33%	22.73%	3.03%	0.00%	0.00%
	1	66.67%	59.09%	51.52%	3.75%	0.00%
	2	0.00%	18.18%	33.33%	26.25%	0.00%
	3	0.00%	0.00%	12.12%	62.50%	18.42%
	4	0.0.0%	0.00%	0.00%	7.50%	81.58%
	N (GT)	3	22	33	80	76
Reader 3	0	36.36%	9.09%	0.00%	0.00%	0.00%
	1	63.64%	48.48%	16.67%	1.83%	0.00%
	2	0.00%	36.36%	33.33%	8.26%	0.00%
	3	0.00%	6.06%	48.48%	78.90%	30.00%
	4	0.00%	0.00%	1.52%	11.01%	70.00%
	N (GT)	11	33	66	109	90
Reader 4	0	33.33%	2.94%	0.00%	0.00%	0.00%
	1	55.56%	41.18%	7.69%	0.00%	0.00%
	2	11.11%	55.88%	58.46%	11.93%	0.00%
	3	0.00%	0.00%	32.31%	62.39%	11.83%
	4	0.00%	0.00%	1.54%	25.69%	88.17%
	N (GT)	9	34	65	109	93
Reader 5	0	0.00%	9.68%	0.00%	0.00%	0.00%
	1	100.0%	45.16%	10.45%	0.87%	0.00%
	2	0.00%	32.26%	38.81%	2.61%	0.00%
	3	0.00%	9.68%	43.28%	59.13%	8.42%
	4	0.00%	3.23%	7.46%	37.39%	91.58%
	N (GT)	10	31	67	115	95
Reader 6	0	57.14%	14.81%	4.08%	0.00%	0.00%
	1	42.86%	70.37%	40.82%	4.88%	0..00%
	2	0.00%	14.81%	46.94%	28.05%	0.00%
	3	0.00%	0.00%	6.12%	56.1%	37.5%
	4	0.00%	0.00%	2.04%	10.98%	62.5%
	N (GT)	14	54	49	82	80
Reader 7	0	50.00%	17.02%	2.94%	0.00%	0.00%
	1	35.71%	70.21%	38.24%	2.38%	0.00%

	2	14.29%	8.51%	41.18%	4.76%	0.00%
	3	0.00%	4.26%	17.65%	90.48%	61.54%
	4	0.00%	0.00%	0.00%	2.38%	38.46%
	N (GT)	14	47	34	42	26

[1] Only IMRs who read at least 10 cases were included.

Table 23: Overall Agreement for Fibrosis for IMRs (CV)[1]

Fibrosis		Ground Truth				
	Grade	**0**	**1**	**2**	**3**	**4**
Reader 1	0	60.00%	17.95%	2.48%	0.00%	0.00%
	1	40.0%	55.13%	18.32%	2.94%	0.00%
	2	0.00%	23.08%	60.40%	13.73%	0.00%
	3	0.00%	3.85%	18.81%	75.49%	23.87%
	4	0.00%	0.00%	0.00%	7.84%	76.13%
	N (GT)	5	78	202	306	155
Reader 2	0	50.00%	28.79%	4.73%	0.00%	0.00%
	1	50.0%	59.09%	46.62%	2.33%	0.00%
	2	0.00%	12.12%	45.95%	32.56%	0.00%
	3	0.00%	0.00%	2.70%	57.67%	20.89%
	4	0.00%	0.00%	0.00%	7.44%	79.11%
	N (GT)	4	66	148	215	158
Reader 3	0	80.0%	9.43%	0.00%	0.00%	0.00%
	1	0.00%	73.58%	33.1%	2.46%	0.00%
	2	20.0%	16.98%	45.52%	12.32%	0.00%
	3	0.00%	0.00%	20.69%	73.40%	20.74%
	4	0.00%	0.00%	0.69%	11.82%	79.26%
	N (GT)	5	53	145	203	135
Reader 4	0	85.71%	37.5%	18.25%	0.61%	0.00%
	1	14.29%	55.00%	51.82%	11.52%	0.00%
	2	0.00%	5.00%	25.55%	41.82%	6.67%
	3	0.00%	2.50%	4.38%	43.64%	43.33%
	4	0.00%	0.00%	0.00%	2.42%	50.0%
	N (GT)	7	80	137	165	30
Reader 5	0	66.67%	11.54%	1.21%	0.39%	0.00%
	1	33.33%	42.31%	11.52%	1.97%	0.00%
	2	0.00%	39.74%	65.45%	14.96%	0.00%
	3	0.00%	5.13%	21.82%	61.02%	3.76%
	4	0.00%	1.28%	0.00%	21.65%	96.24%
	N (GT)	6	78	165	254	133
Reader 6	0	42.86%	17.58%	0.94%	0.31%	0.00%

	1	57.14%	49.45%	25.47%	5.64%	0.00%
	2	0.00%	24.18%	42.92%	10.03%	0.00%
	3	0.00%	8.79%	28.3%	58.62%	8.90%
	4	0.00%	0.00%	2.36%	25.39%	91.10%
	N (GT)	7	91	212	319	146
Reader 7	0	66.67 %	15.56%	1.08 %	0.00%	0.00%
	1	33.33%	72.22%	29.03%	1.29%	0.00%
	2	0.00%	11.11%	58.6%	13.73%	0.00%
	3	0.00%	1.11%	11.29%	73.39%	34.81%
	4	0.00%	0.00%	0.00%	11.59%	65.19%
	N (GT)	6	90	186	233	158
Reader 8	0	33.33%	13.33%	0.00%	0.00%	0.00%
	1	33.33%	73.33%	33.33%	0.00%	0.00%
	2	33.33%	13.33%	50.00%	10.26%	0.00%
	3	0.00%	0.00%	16.67%	89.74%	37.5%
	4	0.00%	0.00%	0.00%	0.00%	62.5%
	N (GT)	3	30	48	39	8

[1] Only IMRs who read at least 10 cases were included.

Appendix A.3

Table 24: Overall AIM-NASH Agreement with Ground Truth for Aggregate Component Scores (AV AIM-NASH only)

				Agreement Evaluation[1]		AIM-NASH - Average IMR	
Feature	**Modality**	**Number of reads**	**Number of unique cases**	**%**	**95% CI[2]**	**Difference (95% CI)[2]**	**P-value[2,3]**
NAS Score ≥ 4 with ≥1 in each score category	AIM-NASH vs GT	593	593	85.8%	82.8%, 88.3%	8.6% (4.1%, 12.9%)	<0.001
	IMR vs GT	1886	593	77.2%	73.6%, 80.8%		
Fibrosis Score 2 or 3 vs other	AIM-NASH vs GT	583	583	77.2%	73.2%, 81.1%	2.8% (-2.2%, 9.0%)	0.246
	IMR vs GT	1870	583	74.4%	69.9%, 78.9%		
Fibrosis Score 4 vs other	AIM-NASH vs GT	583	583	86.4%	82.7%, 89.5%	2.0% (-2.8%, 7.3%)	0.369
	IMR vs GT	1870	583	84.4%	79.9%, 88.4%		
NASH Resolution	AIM-NASH vs GT	593	593	90.2%	87.8%, 92.0%	8.1% (4.5%, 11.8%)	<0.001
	IMR vs GT	1886	593	82.1%	78.9%, 84.8%		

[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 25: 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 26: Overall Agreement with Ground Truth for Aggregate Component Scores (CV AIM-NASH only)

			Agreement Evaluation[2]		AIM-NASH - 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 vs GT	1473	84.0%	82.1%, 85.8%	6.4% (3.8%, 9.0%)	<0.001
	IMR vs GT	4500	77.5%	75.8%, 79.3%		
Fibrosis Score 2 or 3 vs other	AIM-NASH vs GT	1452	79.9%	77.8%, 81.9%	2.8% (0.1%, 5.7%)	0.044
	IMR vs GT	4509	77.1%	75.2%, 78.8%		
Fibrosis Score 4 vs other	AIM-NASH vs GT	1452	91.6%	90.1%, 93.1%	0.1% (-1.7%, 1.9%)	0.935
	IMR vs GT	4509	91.5%	90.4%, 92.6%		
NASH Resolution	AIM-NASH vs GT	1473	89.0%	87.4%, 90.5%	6.5% (4.1%, 8.8%)	<0.001
	IMR vs GT	4500	82.5%	80.7%, 84.3%		

[1]N represents total AIM-NASH 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.

Table 27: AIM-NASH and IMR Positive Percent Agreement with Ground Truth for Aggregate Component Scores (AV)

			Positive Percent Agreement Evaluation[2]		AIM-NASH - 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 vs GT	377	85.7%	79.9%, 88.7%	5.2% (0.3%, 9.2%)	0.043
	IMR vs GT	1185	80.4%	77.9%, 82.8%		
Fibrosis Score 2 or 3 vs other	AIM-NASH vs GT	307	80.8%	77.1%, 86.6%	3.1% (-2.4%, 10.9%)	0.256
	IMR vs GT	981	77.7%	72.7%, 82.1%		
Fibrosis Score 4 vs other	AIM-NASH vs GT	177	75.7%	67.6%, 81.4%	9.7% (-4.2%, 15.5%)	0.235
	IMR vs GT	544	66.0%	62.7%, 76.7%		
NASH Resolution	AIM-NASH vs GT	71	76.1%	67.2%, 84.7%	16.1% (6.1%, 27.3%)	0.001
	IMR vs GT	236	59.9%	54.3%, 65.7%		

[1] N represents total AIM-NASH assessments where the GT met the binary score definition and total of all IMR assessments where the GT met the binary score definition.
[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 28: AIM-NASH-Assisted and IMR Positive Percent Agreement with Ground Truth for Aggregate Component Scores (CV)

			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: AIM-NASH and IMR Positive Percent Agreement with Ground Truth for Aggregate Component Scores (CV)

			Positive Percent Agreement Evaluation[2]		AIM-NASH - 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 vs GT	1011	87.6%	85.6%, 89.6%	9.8% (6.7%, 12.8%)	<0.001
	IMR vs GT	3072	77.8%	75.6%, 80.1%		
Fibrosis Score 2 or 3 vs other	AIM-NASH vs GT	960	84.7%	82.3%, 86.8%	8.7% (5.5%, 11.9%)	<0.001
	IMR vs GT	2977	75.9%	73.7%, 78.1%		
	AIM-NASH vs GT	304	79.3%	74.8%, 83.7%	4.3% (-3.0%, 11.7%)	0.243

Fibrosis Score 4 vs other	IMR vs GT	923	74.9%	69.1%, 81.1%		
NASH Resolution	AIM-NASH vs GT	158	75.9%	69.5%, 82.2%	18.8% (10.8%, 26.5%)	<0.001
	IMR vs GT	492	57.2%	53.0%, 61.4%		

[1] N represents total AIM-NASH 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 30: AIM-NASH and IMR Negative Percent Agreement with Ground Truth for Aggregate Component Scores (AV)

			Negative percent Agreement Evaluation[2]		AIM-NASH - 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 vs GT	216	86.1%	81.2%, 92.7%	11.3% (2.9%, 17.8%)	0.009
	IMR vs GT	701	74.8%	70.5%, 81.2%		
Fibrosis Score 2 or 3 vs other	AIM-NASH vs GT	276	73.2%	68.1%, 80.4%	3.5% (-9.2%, 9.4%)	0.936
	IMR vs GT	889	69.7%	66.9%, 79.9%		
Fibrosis Score 4 vs other	AIM-NASH	406	91.1%	89.1%, 94.6%	-1.3% (-4.1%, 2.4%)	0.641
	IMR vs GT	1326	92.4%	90.8%, 94.0%		
NASH Resolution	AIM-NASH	522	92.1%	89.3%, 94.6%	6.0% (2.9%, 8.5%)	<0.001
	IMR vs GT	1650	86.1%	84.3%, 87.8%		

[1] N represents total AIM-NASH 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 31: AIM-NASH-Assisted and IMR Negative Percent Agreement with Ground Truth for Aggregate Component Scores (CV)

			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.

Table 32: AIM-NASH and IMR Negative Percent Agreement with Ground Truth for Aggregate Component Scores (CV)

			Negative Percent Agreement Evaluation[2]		AIM-NASH - 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 vs GT	462	76.0%	71.9%, 79.8%	0.8% (-3.9%, 5.4%)	0.720
	IMR vs GT	1428	75.1%	72.4%, 77.6%		
Fibrosis Score 2 or 3 vs other	AIM-NASH vs GT	492	70.5%	66.5%, 74.3%	-9.3% (-14.2%, -4.4%)	<0.001
	IMR vs GT	1532	79.8%	76.8%, 82.7%		
Fibrosis Score 4 vs other	AIM-NASH vs GT	1148	94.9%	93.6%, 96.1%	0.5% (-1.2%, 2.0%)	0.550
	IMR vs GT	3586	94.4%	93.4%, 95.3%		
NASH Resolution	AIM-NASH vs GT	1315	90.6%	88.9%, 92.1%	4.8% (2.3%, 7.1%)	<0.001
	IMR vs GT	4008	85.8%	84.1%, 87.5%		

[1] N represents total AIM-NASH 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.

Appendix A.4

Table 33: Weighted Kappa Worst Case Imputation Analysis for AIM-NASH and IMR (AV)

				Weighted Kappa Evaluation		AIM-NASH - Average IMR		
Feature	**Modality**	**Number of Reads**	**Number of Unique Cases**	**Estimate**	**95% CI[1]**	**Difference (95% CI)[1]**	**P-value[1,2]**	**P-value[1,3]**
Steatosis	AIM-NASH vs GT	597	597	0.679	0.634, 0.711	0.051 (-0.010, 0.112)	<0.001	0.049
	IMR vs GT	1997	597	0.628	0.578, 0.674			
Inflammation	AIM-NASH vs GT	593	593	0.412	0.365, 0.479	0.093 (0.012, 0.186)	<0.001	0.011
	IMR vs GT	1980	593	0.319	0.254, 0.382			
Ballooning	AIM-NASH vs GT	597	597	0.597	0.548, 0.651	0.235 (0.160, 0.321)	<0.001	<0.001
	IMR vs GT	1997	597	0.363	0.297, 0.423			
Fibrosis	AIM-NASH vs GT	583	583	0.654	0.612, 0.702	0.113 (0.058, 0.175)	<0.001	<0.001
	IMR vs GT	1939	583	0.541	0.499, 0.579			

[1] 95% CI based on bootstrap analysis resampling cases.
[2] P-value for non-inferiority hypothesis AIM-NASH – IMR < -0.1.

[3] P-value for two-sided null hypothesis that difference in agreement is equal to 0.0.

Table 34: Weighted Kappa Worst Case Imputation Analysis for AIM-NASH-Assisted and IMR (CV)

			Weighted Kappa Evaluation		AIM-NASH-Assisted - Average IMR		
Feature	**Modality**	**Number of reads**	**Estimate**	**95% CI[1]**	**Difference (95% CI)[1]**	**P-value[1,2]**	**P-value[1,3]**
Steatosis	AIM-NASH-Assisted vs GT	1480	0.661	0.652, 0.703	0.185 (0.164, 0.236)	<0.001	<0.001
	IMR vs GT	4951	0.476	0.448, 0.502			
Lobular Inflammation	AIM-NASH-Assisted vs GT	1477	0.405	0.354, 0.454	0.246 (0.201, 0.302)	<0.001	<0.001
	IMR vs GT	4938	0.159	0.137, 0.183			
Hepatocellular Ballooning	AIM-NASH-Assisted vs GT	1475	0.553	0.516, 0.599	0.253 (0.215, 0.304)	<0.001	<0.001
	IMR vs GT	4936	0.300	0.273, 0.326			
Fibrosis	AIM-NASH-Assisted vs GT	1452	0.628	0.612, 0.660	0.103 (0.074, 0.150)	<0.001	<0.001
	IMR vs GT	4848	0.525	0.498, 0.550			

[1] 95% CI based on bootstrap analysis resampling cases.
[2] P-value for non-inferiority hypothesis AIM-NASH-Assisted – IMR < -0.1.
[3]P-value for two-sided null hypothesis that difference in agreement is equal to 0.0.

Table 35: Weighted Kappa Worst Case Imputation Analysis for AIM-NASH and IMR (CV)

			Weighted Kappa Evaluation		AIM-NASH – Average IMR		
Feature	**Modality**	**N[1]**	**Estimate**	**95% CI[2]**	**Difference (95% CI)[2]**	**P-value[2,3]**	**P-value[2,4]**
Steatosis	AIM-NASH vs GT	1481	0.674	0.648, 0.701	0.198 (0.160, 0.238)	<0.001	<0.001
	IMR vs GT	4954	0.476	0.449, 0.502			
Lobular Inflammation	AIM-NASH vs GT	1478	0.415	0.382, 0.449	0.256 (0.215, 0.297)	<0.001	<0.001
	IMR vs GT	4941	0.159	0.137, 0.183			
Hepatocellular Ballooning	AIM-NASH vs GT	1476	0.561	0.526, 0.596	0.261 (0.218, 0.305)	<0.001	<0.001
	IMR vs GT	4939	0.300	0.273, 0.326			
Fibrosis	AIM-NASH vs GT	1453	0.635	0.607, 0.660	0.110 (0.070, 0.147)	<0.001	<0.001
	IMR vs GT	4851	0.525	0.498, 0.550			

[1] N represents total AIM-NASH 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 – IMR < -0.1.
[4] P-value for two-sided null hypothesis that difference in agreement is equal to 0.0.

Table 36: Weighted Kappa Worst Case Imputation for AIM-NASH and Best Case Imputation for IMR (AV)

				Weighted Kappa Evaluation		AIM-NASH – Average IMR		
Feature	**Modality**	**Number of Reads**	**Number of Unique Cases**	**Estimate**	**95% CI[1]**	**Difference (95% CI)[1]**	**P-value[1,2]**	**P-value[1,3]**

Steatosis	AIM-NASH vs GT	597	597	0.679	0.634, 0.711	-0.061 (-0.109, -0.010)	0.052	0.992
	IMR vs GT	1997	597	0.740	0.700, 0.769			
Lobular Inflammation	AIM-NASH vs GT	593	593	0.412	0.365, 0.479	-0.015 (-0.088, 0.080)	0.012	0.592
	IMR vs GT	1980	593	0.427	0.360, 0.477			
Hepatocellular Ballooning	AIM-NASH vs GT	597	597	0.597	0.548, 0.651	0.141 (0.072, 0.225)	<0.001	<0.001
	IMR vs GT	1997	597	0.456	0.390, 0.512			
Fibrosis	AIM-NASH vs GT	583	583	0.654	0.612, 0.702	-0.003 (-0.054, 0.059)	<0.001	0.426
	IMR vs GT	1939	583	0.658	0.613, 0.695			

[1] 95% CI based on bootstrap analysis resampling cases.
[2] P-value for non-inferiority hypothesis AIM-NASH – IMR < -0.1.
[3] P-value for two-sided null hypothesis that difference in agreement is equal to 0.0.

Table 37: Weighted Kappa Worst Case Imputation for AIM-NASH-Assisted and Best Case Imputation for IMR (CV)

			Weighted Kappa Evaluation		AIM-NASH-Assisted – Average IMR		
Feature	**Modality**	**Number of reads**	**Estimate**	**95% CI[1]**	**Difference (95% CI)[1]**	**P-value[1,2]**	**P-value[1,3]**
Steatosis	AIM-NASH-Assisted vs GT	1480	0.661	0.652, 0.703	-0.053 (-0.068, -0.008)	<0.001	0.991
	IMR vs GT	4951	0.713	0.694, 0.731			
Lobular Inflammation	AIM-NASH-Assisted vs GT	1477	0.405	0.354, 0.454	0.029 (-0.013, 0.087)	<0.001	0.058
	IMR vs GT	4938	0.376	0.349, 0.401			
Hepatocellular Ballooning	AIM-NASH-Assisted vs GT	1475	0.553	0.516, 0.599	0.069 (0.034, 0.120)	<0.001	<0.001
	IMR vs GT	4936	0.484	0.458, 0.510			
Fibrosis	AIM-NASH-Assisted vs GT	1452	0.628	0.612, 0.660	-0.041 (-0.064, -0.000)	<0.001	0.976
	IMR vs GT	4848	0.669	0.647, 0.687			

[1] 95% CI based on bootstrap analysis resampling cases.
[2] P-value for non-inferiority hypothesis AIM-NASH-Assisted – IMR < -0.1.
[3] P-value for two-sided null hypothesis that difference in agreement is equal to 0.0.

Table 38: Weighted Kappa Worst Case Imputation for AIM-NASH and Best Case Imputation for IMR (CV)

			Weighted Kappa Evaluation		AIM-NASH - Average IMR		
Feature	**Modality**	**N[1]**	**Estimate**	**95% CI[2]**	**Difference (95% CI)[2]**	**P-value[2,3]**	**P-value[2,4]**
Steatosis	AIM-NASH vs GT	1481	0.674	0.648, 0.701	-0.040 (-0.071, -0.006)	0.001	0.987
	IMR vs GT	4954	0.713	0.694, 0.732			
Lobular Inflammation	AIM-NASH vs GT	1478	0.415	0.382, 0.449	0.038 (-0.004, 0.083)	<0.001	0.039
	IMR vs GT	4941	0.376	0.349, 0.401			
Hepatocellular Ballooning	AIM-NASH vs GT	1476	0.561	0.526, 0.596	0.077 (0.034, 0.121)	<0.001	<0.001
	IMR vs GT	4939	0.484	0.458, 0.510			
Fibrosis	AIM-NASH vs GT	1453	0.635	0.607, 0.660	-0.034 (-0.067, -0.001)	<0.001	0.978

	IMR vs GT	4851	0.669	0.647, 0.687			

[1] N represents total AIM-NASH 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 – IMR < -0.1.
[4] P-value for two-sided null hypothesis that difference in agreement is equal to 0.0.

Table 39: Weighted Kappa Best Case Imputation for AIM-NASH and Worst Case Imputation for IMR (AV)

				Weighted Kappa Evaluation		Average AIM-NASH - IMR		
Feature	**Modality**	**Number of Reads**	**Number of Unique Cases**	**Estimate**	**95% CI[1]**	**Difference (95% CI)[1]**	**P-value[1,2]**	**P-value[1,3]**
Steatosis	AIM-NASH vs GT	597	597	0.679	0.634, 0.711	0.051 (-0.010, 0.112)	<0.001	0.049
	IMR vs GT	1997	597	0.628	0.578, 0.674			
Lobular Inflammation	AIM-NASH vs GT	593	593	0.412	0.365, 0.479	0.093 (0.012, 0.186)	<0.001	0.011
	IMR vs GT	1980	593	0.319	0.254, 0.382			
Hepatocellular Ballooning	AIM-NASH vs GT	597	597	0.597	0.548, 0.651	0.235 (0.160, 0.321)	<0.001	<0.001
	IMR vs GT	1997	597	0.363	0.297, 0.423			
Fibrosis	AIM-NASH vs GT	583	583	0.654	0.612, 0.702	0.113 (0.058, 0.175)	<0.001	<0.001
	IMR vs GT	1939	583	0.541	0.499, 0.579			

[1] 95% CI based on bootstrap analysis resampling cases.
[2] P-value for non-inferiority hypothesis AIM-NASH – IMR < -0.1.
[3] P-value for two-sided null hypothesis that difference in agreement is equal to 0.0.

Table 40: Weighted Kappa Best Case Imputation for AIM-NASH-Assisted and Worst Case Imputation for IMR (CV)

			Weighted Kappa Evaluation		AIM-NASH-Assisted - Average IMR		
Feature	**Modality**	**Number of reads**	**Estimate**	**95% CI[1]**	**Difference (95% CI)[1]**	**P-value[1,2]**	**P-value[1,3]**
Steatosis	AIM-NASH-Assisted vs GT	1480	0.680	0.652, 0.703	0.204 (0.168, 0.239)	<0.001	<0.001
	IMR vs GT	4951	0.476	0.448, 0.502			
Lobular Inflammation	AIM-NASH-Assisted vs GT	1477	0.423	0.362, 0.461	0.265 (0.209, 0.308)	<0.001	<0.001
	IMR vs GT	4938	0.159	0.137, 0.183			
Hepatocellular Ballooning	AIM-NASH-Assisted vs GT	1475	0.567	0.520, 0.602	0.267 (0.220, 0.308)	<0.001	<0.001
	IMR vs GT	4936	0.300	0.273, 0.326			
Fibrosis	AIM-NASH-Assisted vs GT	1452	0.660	0.633, 0.683	0.135 (0.094, 0.169)	<0.001	<0.001

	IMR vs GT	4848	0.525	0.498, 0.550			

[1] 95% CI based on bootstrap analysis resampling cases.
[2] P-value for non-inferiority hypothesis AIM-NASH-Assisted – IMR < -0.1.
[3]P-value for two-sided null hypothesis that difference in agreement is equal to 0.0.

Table 41: Weighted Kappa Best Case Imputation for AIM-NASH and Worst Case Imputation for IMR (CV)

			Weighted Kappa Evaluation		AIM-NASH - Average IMR		
Feature	**Modality**	**N[1]**	**Estimate**	**95% CI[2]**	**Difference (95% CI)[2]**	**P-value[2,3]**	**P-value[2,4]**
Steatosis	AIM-NASH vs GT	1481	0.676	0.649, 0.703	0.200 (0.163, 0.240)	<0.001	<0.001
	IMR vs GT	4954	0.476	0.449, 0.502			
Lobular Inflammation	AIM-NASH vs GT	1478	0.416	0.383, 0.450	0.257 (0.216, 0.299)	<0.001	<0.001
	IMR vs GT	4941	0.159	0.137, 0.183			
Hepatocellular Ballooning	AIM-NASH	1476	0.562	0.526, 0.597	0.262 (0.219, 0.305)	<0.001	<0.001
	IMR vs GT	4939	0.300	0.273, 0.326			
Fibrosis	AIM-NASH vs GT	1453	0.636	0.608, 0.661	0.111 (0.071, 0.149)	<0.001	<0.001
	IMR vs GT	4851	0.525	0.498, 0.550			

[1] N represents total AIM-NASH 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 – IMR < -0.1.
[4] P-value for two-sided null hypothesis that difference in agreement is equal to 0.0.

Table 42: Weighted Kappa Best Case Imputation Analysis for AIM-NASH and IMR (AV)

				Weighted Kappa Evaluation		AIM-NASH - Average IMR		
Feature	**Modality**	**Number of Reads**	**Number of Unique Cases**	**Estimate**	**95% CI[1]**	**Difference (95% CI)[1]**	**P-value[1,2]**	**P-value[1,3]**
Steatosis	AIM-NASH vs GT	597	597	0.679	0.634, 0.711	-0.061 (-0.109, -0.010)	0.052	0.992
	IMR vs GT	1997	597	0.740	0.700, 0.769			
Lobular Inflammation	AIM-NASH vs GT	593	593	0.412	0.365, 0.479	-0.015 (-0.088, 0.080)	0.012	0.592
	IMR vs GT	1980	593	0.427	0.360, 0.477			
Hepatocellular Ballooning	AIM-NASH vs GT	597	597	0.597	0.548, 0.651	0.141 (0.072, 0.225)	<0.001	<0.001
	IMR vs GT	1997	597	0.456	0.390, 0.512			
Fibrosis	AIM-NASH vs GT	583	583	0.654	0.612, 0.702	-0.003 (-0.054, 0.059)	<0.001	0.426
	IMR vs GT	1939	583	0.658	0.613, 0.695			

[1] 95% CI based on bootstrap analysis resampling cases.
[2] P-value for non-inferiority hypothesis AIM-NASH – IMR < -0.1.
[3] P-value for two-sided null hypothesis that difference in agreement is equal to 0.0.

Table 43: Weighted Kappa Best Case Imputation Analysis for AIM-NASH-Assisted and IMR (CV)

	Weighted Kappa Evaluation	AIM-NASH-Assisted - Average IMR

Feature	Modality	Number of reads	Estimate	95% CI[1]	Difference (95% CI)[1]	P-value[1,2]	P-value[1,3]
Steatosis	AIM-NASH-Assisted vs GT	1480	0.680	0.652, 0.703	-0.034 (-0.065, -0.005)	<0.001	0.990
	IMR vs GT	4951	0.713	0.694, 0.731			
Lobular Inflammation	AIM-NASH-Assisted vs GT	1477	0.423	0.362, 0.461	0.047 (-0.007, 0.093)	<0.001	0.044
	IMR vs GT	4938	0.376	0.349, 0.401			
Hepatocellular Ballooning	AIM-NASH-Assisted vs GT	1475	0.567	0.520, 0.602	0.083 (0.038, 0.124)	<0.001	<0.001
	IMR vs GT	4936	0.484	0.458, 0.510			
Fibrosis	AIM-NASH-Assisted vs GT	1452	0.660	0.633, 0.683	-0.009 (-0.044, 0.020)	<0.001	0.777
	IMR vs GT	4848	0.669	0.647, 0.687			

[1] 95% CI based on bootstrap analysis resampling cases.
[2] P-value for non-inferiority hypothesis AIM-NASH-Assisted – IMR < -0.1.
[3]P-value for two-sided null hypothesis that difference in agreement is equal to 0.0.

Table 44: Weighted Kappa Best Case Imputation Analysis for AIM-NASH and IMR (CV)

			Weighted Kappa Evaluation		AIM-NASH – Average IMR		
Feature	Modality	N[1]	Estimate	95% CI[2]	Difference (95% CI)[2]	P-value[2,3]	P-value[2,4]
Steatosis	AIM-NASH vs GT	1481	0.676	0.649, 0.703	-0.037 (-0.069, -0.004)	<0.001	0.982
	IMR vs GT	4954	0.713	0.694, 0.732			
Lobular Inflammation	AIM-NASH vs GT	1478	0.416	0.383, 0.450	0.040 (-0.003, 0.085)	<0.001	0.035
	IMR vs GT	4941	0.376	0.349, 0.401			
Hepatocellular Ballooning	AIM-NASH vs GT	1476	0.562	0.526, 0.597	0.078 (0.035, 0.121)	<0.001	<0.001
	IMR vs GT	4939	0.484	0.458, 0.510			
Fibrosis	AIM-NASH vs GT	1453	0.636	0.608, 0.661	-0.033 (-0.066, 0.001)	<0.001	0.970
	IMR vs GT	4851	0.669	0.647, 0.687			

Appendix A.5

Table 45: Weighted Kappa Analysis for NASH Components AIM-NASH-Assisted and Mode/Median Panel Comparison (CV)

			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	0.676	0.648, 0.703	-0.067 (-0.096, -0.039)	0.0095
	IMR vs GT	1470	0.743	0.717, 0.766		
Lobular Inflammation	AIM-NASH-Assisted vs GT	1467	0.422	0.385, 0.456	-0.021 (-0.064, 0.024)	<0.00001
	IMR vs GT	1470	0.443	0.406, 0.478		
Hepatocellular Ballooning	AIM-NASH-Assisted vs GT	1467	0.563	0.529, 0.596	0.041 (0.001, 0.081)	<0.00001
	IMR vs GT	1469	0.522	0.486, 0.557		
Fibrosis	AIM-NASH-Assisted vs GT	1434	0.655	0.629, 0.681	-0.064 (-0.092, -0.036)	0.006

	IMR vs GT	1450	0.719	0.697, 0.741

[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 at a 5% alpha.

Table 46: Agreement of AIM-NASH and Pathologist Mean Pairwise Comparison from Standalone Verification

Feature	AIM - Consensus WK (95% CI)	Pathologist Mean Pairwise WK (95% CI)	N Consensus	Difference (>0 for acceptance)
Steatosis	0.68 (0.62, 0.75)	0.55 (0.5, 0.6)	231	0.17
Lobular Inflammation	0.5 (0.42, 0.58)	0.45 (0.37, 0.51)	231	0.07
Hepatocellular Ballooning	0.49 (0.41, 0.56)	0.39 (0.32, 0.45)	231	0.12
Fibrosis	0.7 (0.65, 0.74)	0.65 (0.62, 0.69)	220	0.1

Table 47: 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

Appendix B References

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.

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.

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.

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.

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.

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.

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*.

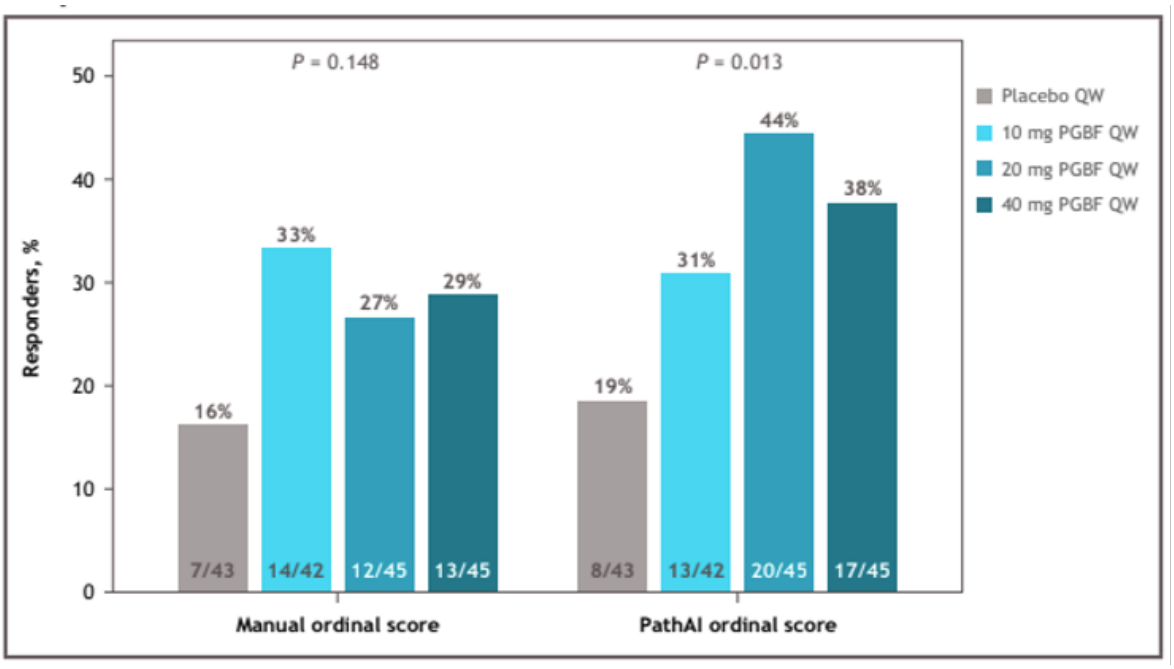
Appendix C Evidence from Published Literature

CASE STUDY 1:

Lead Author/Title: Diane Shevell, Comparison of manual vs machine learning approaches to liver biopsy scoring for NASH and fibrosis: a post hoc analysis of the FALCON 1 study.
Conference: Poster presentation at the American Association for the Study of Liver Diseases Meeting, 2021
Clinical trial (Sponsor): NCT03486899 (Bristol Myers Squibb)
Method: Whole slide images of H&E- and Masson's trichrome-stained liver biopsy tissue from subjects enrolled in the FALCON 1 NASH clinical trial (Ph2b study of pegbelfermin for treatment of NASH in adults with Stage 3 Liver Fibrosis) were read by the study's Central Pathologist and by AIM-NASH. Response rates per treatment group were recorded and compared across the histologic evaluation methodologies. Primary endpoint responders were patients with ≥1-stage improvement in NASH CRN fibrosis stage without NASH worsening, or NASH improvement with no worsening of fibrosis, at Week 24.
Key results:

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

Figure 1: AIM-NASH vs. Central Pathologist detection of primary endpoint response in a Ph2 study of pegbelfermin for treatment of NASH with CRN Fibrosis Stage 3. Primary endpoint responders were patients with ≥ 1 stage NASH CRN fibrosis improvement without NASH worsening or NASH improvement with no worsening of fibrosis at week 24. Cochran-Armitage test for trend was used to compare PGBF vs placebo. NASH, nonalcoholic steatohepatitis; PGBF, pegbelfermin; QW, once weekly.



CASE STUDY 2:

Lead Author/Title: Stephen Harrison Retrospective AI-based Measurement of NASH Histology (AIM-NASH) analysis of biopsies from Phase 2 study of resmetirom confirms significant treatment-induced changes in histologic features of nonalcoholic steatohepatitis.

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

Clinical Trial (Sponsor): NCT02912260 (Madrigal Pharmaceuticals)

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

Key results:

AIM-NASH detected a statistically significant difference between response rates in the resmetirom vs. placebo subject groups **(**Table 48**)**.

Both the Central Reader and the independent expert NASH pathologist ("Reader 2") also detected statistically significant differences between response rates in the resmetirom vs. placebo subject groups **(**Table 48**)**.

Table 48: AIM-NASH vs. Pathologist detection of endpoint response in Ph2 study of MGL-3196 for treatment of NASH with CRN Fibrosis Stages 1-3. Consistent with the Central Pathologist and Reader 2, AIM-NASH detected a significantly greater treatment response in the resmetirom-treated group relative to placebo.

Endpoint	Scorer	Resmetirom	Placebo response rate	p-value

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

CASE STUDY 3:

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

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

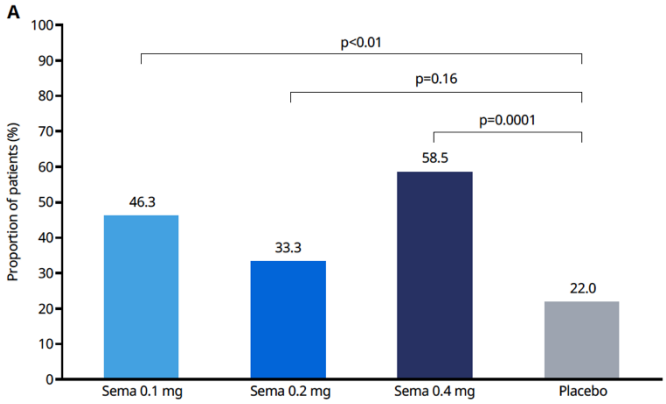
Clinical Trial: NCT02970942 (Novo Nordisk)

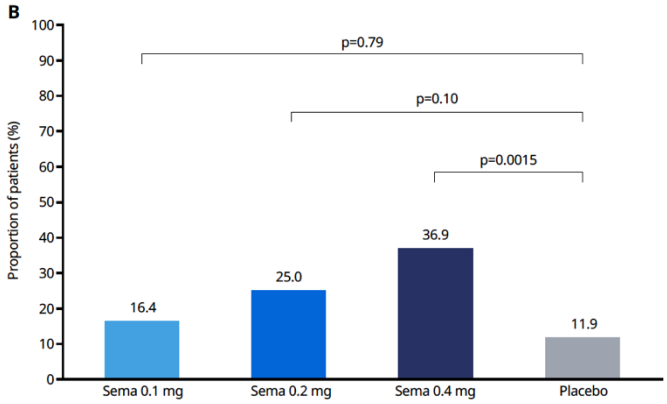
Method: Whole slide images of H&E- and Masson's trichrome-stained liver biopsy tissue from subjects enrolled in the Ph2 study of semaglutide for treatment of NASH in patients with NASH CRN Fibrosis stages 1-3 were read by the study's Central Pathologists (N=2) and AIM-NASH. Response rates per treatment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

AIM-NASH detected a dose-related response in subjects treated with semaglutide, where increasing dosages of semaglutide resulted in increasingly improved drug response **(Figure 2, Panel B).**

Central Pathologist scoring did NOT detect a dose-related drug response in subjects treated with semaglutide **(Figure 2, Panel A)**.

Figure 2: Dose-related drug response detected via Central Pathologists vs. AIM-NASH in Ph2 study of semaglutide for treatment of NASH with CRN Fibrosis Stages 1-3. (A) Dose-related drug response is not detected by Central Pathologist scoring. (B) Dose-related dr





CASE STUDY 4:

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

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

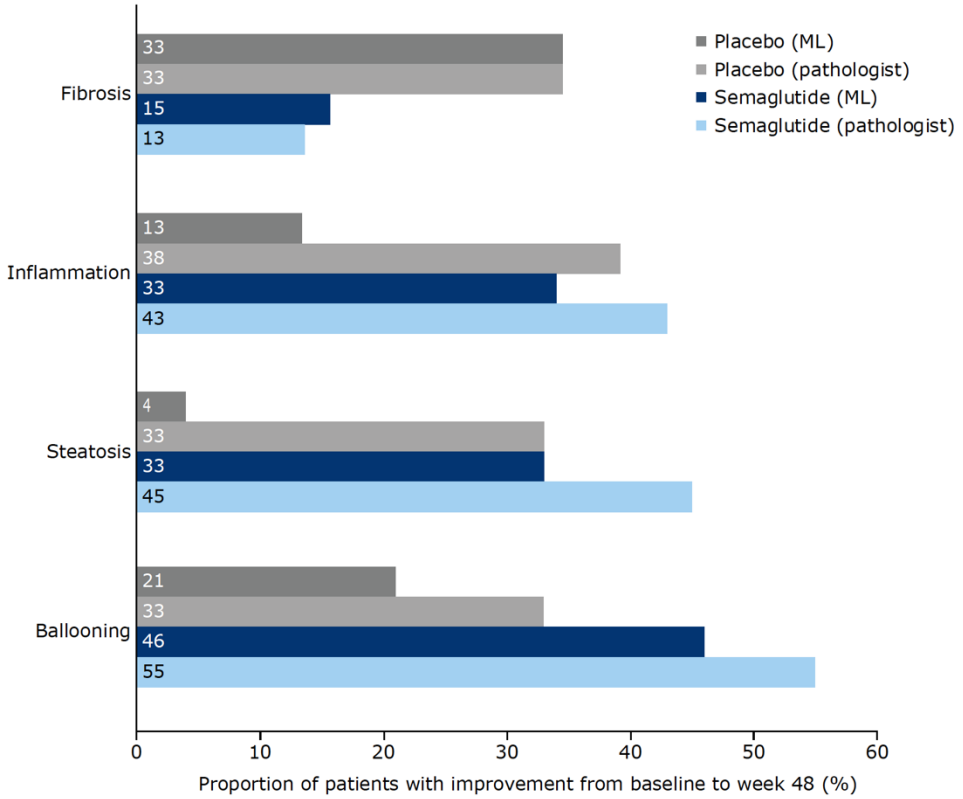
Clinical Trial (Sponsor): NCT03987451 (Novo Nordisk)

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

Key result:

AIM-NASH detected significantly lower placebo response than the Central Pathologist for steatosis (4% vs. 33%), ballooning (21% vs. 33%), and lobular inflammation (13% vs. 38%) **(**Figure 3**)**.

Figure 3: 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



CASE STUDY 5:

Lead Author/Title: Janani S. Iyer, 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.

Conference: Poster presentation at the American Association for the Study of Liver Diseases Meeting, 2023

Clinical trial (Sponsor): NCT03900429 (Madrigal)

Method: Participants with paired Baseline and End-of-Study (Week 52) biopsies were assessed by the study's two Central Pathologists and by AIM-NASH (N=782 and N=777, respectively). A 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).

Key results:

- Overall, 966 patients were in the primary analysis (80mg resmetirom [n=322], 100mg resmetirom [n=323], or placebo [n=321]) of which a total of 782 had paired baseline and Week 52 liver biopsies.
- As reported previously, both primary endpoints were achieved with both resmetirom 80 and 100mg ($p<0.0002$ vs placebo for all) by Central Pathologist review. Consensus analyses between the two pathologists confirmed the results.
- Both the Central Pathologists and AIM-NASH showed a clear effect on histologic response (Figure 1).
- Placebo response rates measured by the Central Pathologists vs. AIM-NASH for both primary endpoints were numerically similar (Figure 1).
- The difference in response rates between treated and placebo subjects was statistically significant for both the 80mg and 100mg treatment groups, as measured by Pathologists and AIM-NASH ($p<0.02$ for all comparisons).

Fi0.17gure 1: Comparison of response rates for primary endpoints measured by Central Pathologists and AIM-NASH

	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**

Appendix D AIM-NASH IFU