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 Second List of Issues (LoI) issued after EMA review of the LoI written response submitted 16 November 2023 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 LoI to be addressed – October 13, 2023
- Submitted written response for LoI – November 16, 2023
- Discussion meeting to discuss LoI responses – November 29, 2023
- EMA Delivers Second LoI to be addressed – January 17, 2024
- Submitted written response for Second LoI – January 29, 2024

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 only by 29 January 2024

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:

1. The previous issue 1 on the Context of Use statement has not been satisfactorily addressed. The applicant should re-implement some of the previous wording, and also define the term "pathologist", as well as the use within the trial more clearly. The following is suggested:

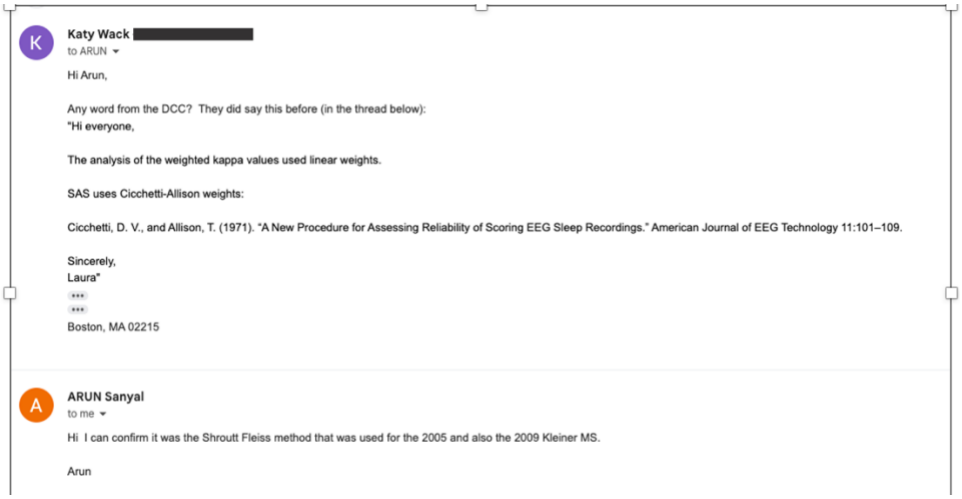
   "A tool which determines a disease activity biomarker based on NAS component scores (steatosis, hepatocellular ballooning, lobular inflammation) and fibrosis stage in biopsies in MASH clinical trials. The tool is an aid to a single central pathologist that is to be used for enrolment/inclusion of patients into clinical phase 2 and phase 3 trials in MASH as well as for the evaluation of the study outcome (primary or secondary) in case this is intended to be based on histology evaluation. The pathologist will review the output of AIM-NASH and take an active role in its interpretation by accepting or rejecting each of the NAS components and fibrosis stage, after confirming sample evaluability and determining the presence of any additional findings. AIM-NASH should be used with slides scanned with the Aperio AT2 scanner."

   In case any deviations or additions are proposed, these should be thoroughly justified. The applicant is also made aware that the term "MASH" should be used throughout going forward (unless it is referred to "historical" data/facts) instead of "NASH".

   PathAI Response:
   We agree with the suggested Context of Use statement:
   "A tool which determines a disease activity biomarker based on NAS component scores (steatosis, hepatocellular ballooning, lobular inflammation) and fibrosis stage in biopsies in MASH clinical trials. The tool is an aid to a single central pathologist that is to be used for enrolment/inclusion of patients into clinical phase 2 and phase 3 trials in MASH as well as for the evaluation of the study outcome (primary or secondary) in case this is intended to be based on histology evaluation. The pathologist will review the output of AIM-NASH and take an active role in its interpretation by accepting or rejecting each of the NAS components and fibrosis stage, after confirming sample evaluability and determining the presence of any additional findings. AIM-NASH should be used with slides scanned with the Aperio AT2 scanner."

2. The previous issue 11 has been addressed satisfactorily. However, since there are now conflicting statements on the table with regard to the use of which kappa type in the NASH CRN trials (Kleiner 2005 and 2019), the applicant is requested to provide a written statement by the authors of the two trials, describing the statistical methodology used.

PathAI Response:
The information sent to us in 2020 from authors and the NASH CRN's Data Coordinating Center, stating Cicchetti-Allison weights were utilized in the Kleiner 2005, 2019 manuscripts, has now been corrected and Dr. Sanyal has confirmed, via a recent communication with the CRN DCC, that the analyses in those manuscripts actually utilized the Shrout-Fleiss Kappa method (as stated in the Sanyal 2021 AASLD publication), which tends to produce higher values for multiple binned, ordinal scoring systems. These AIM-NASH validation studies, as well as the Davison and Newsome 2021 studies, utilized the Cicchetti-Allison method of linear weights in calculating Kappa statistics.



K **Katy Wack** [redacted]
to ARUN

Hi Arun,

Any word from the DCC? They did say this before (in the thread below):
"Hi everyone,

The analysis of the weighted kappa values used linear weights.

SAS uses Cicchetti-Allison weights:

Cicchetti, D. V., and Allison, T. (1971). "A New Procedure for Assessing Reliability of Scoring EEG Sleep Recordings." American Journal of EEG Technology 11:101–109.

Sincerely,
Laura"
***
***
Boston, MA 02215

A **ARUN Sanyal**
to me

Hi I can confirm it was the Shroutt Fleiss method that was used for the 2005 and also the 2009 Kleiner MS.

Arun

3. The response to the previous issue 27 is not fully clear. From the answer provided, it seems that the terms "held-out test-set" and "internal test-set" are used interchangeably, while Figure 11 of the briefing document clearly pointed to a separate analysis with separate data. The applicant should clarify the issue further.

PathAI Response:
In the response to the previous issue 27, the term held-out test set was used generally to describe a dataset that was not used in the training and development of the models. As depicted in Figure 11 in the briefing document (also see below), 2 separate steps with different datasets, both not used in training, are utilized for verification purposes. The internal test set and held-out test set datasets are described in lines 10-12 of Table 10 in the briefing document (also see below).

Figure 11: AIM-NASH Iterative Model Development

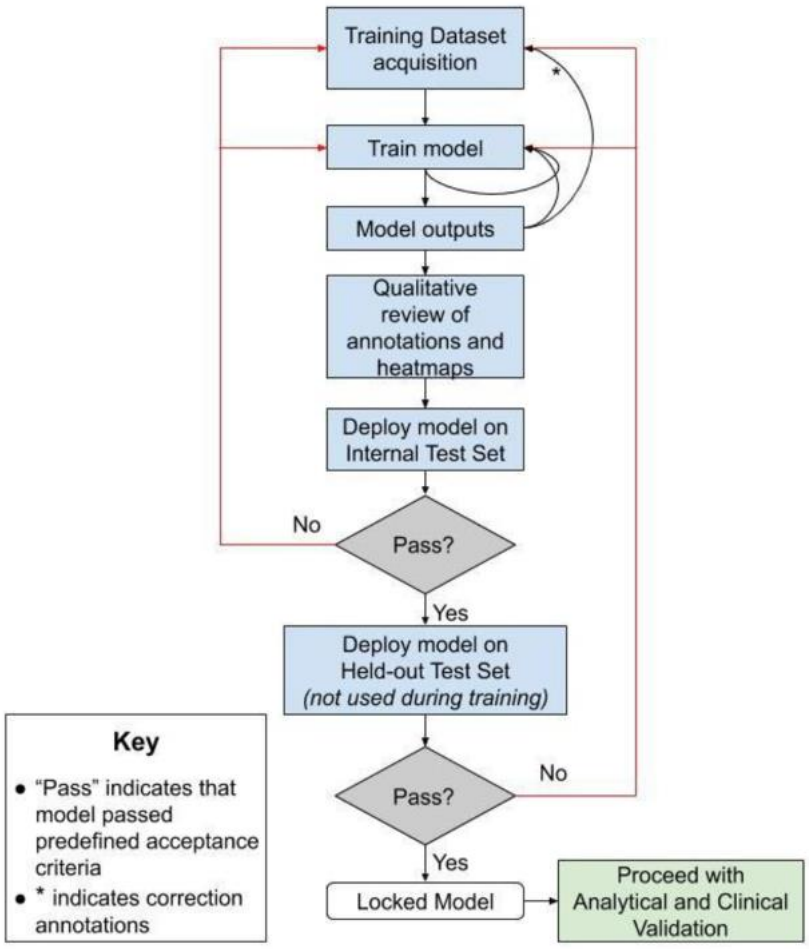


Table 10: Overview of Available Datasets for Developing ML-Based Image Segmentation and CRN Scoring Models

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

The Internal Test Set was a Phase 2 NASH trial evaluating a novel insulin sensitizer (Table 10, Dataset #10 in the briefing document). Six-hundred and thirty-two (632) cases were

utilized and the dataset contained WSIs from a population with a range of fibrosis stage and NAS≥4 with a score of at least 1 in each component of NAS. Slide level scores for NAS components and fibrosis were collected from 3 expert liver pathologists. Agreement of AIM-NASH read-outs with mean consensus pathologist reads was assessed using linearly weighted kappa statistics. For reference, pairwise pathologist agreements were also 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 for internal test set verification (Appendix A, Table 47 in the LoI response document submitted 06 Dec 2023, also see below).

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

After acceptable performance was demonstrated on the Internal Test Set, the model was deployed on the Held-out Test Set (Figure 11 in the briefing document; also referred to as Standalone Analytical Verification). For the Held-out Test Set, 231 H&E slides and 220 trichrome slides from two (2) Phase 2 clinical trials (different from Internal Test Set) were utilized (Table 10, Datasets #11 and 12 in the briefing document). These slides were selected based on fibrosis stage derived from AIM-NASH scores, with approximately 50 cases being represented per fibrosis stage. As with the Internal Test Set, slide level scores 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. AIM-NASH met the pre-defined acceptance criteria for standalone analytical verification (the lower 2.5% confidence interval of the linearly weighted Kappa of the AIM-NASH scores vs. the reference standard median consensus scores be at least as good as 0.1 below the mean pairwise linearly weighted Kappa among network pathologists assessed separately for each NAS component and fibrosis stage; Appendix A, Table 46 in the LoI response document submitted 06 Dec 2023, also see below) and therefore analytical and clinical validation studies were performed.

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

4. The previous issue on the categorical evaluation (such as F2/3 vs. other, NAS , ≥4 vs. <4, and NASH resolution) for accuracy (OPA, PPA, and NPA), while having provided some insight on the potential influence on patient inclusion, should be further extended to address the endpoint evaluation regarding fibrosis stage. The applicant is therefore requested to present additional analyses with presentation of OPA, PPA, and NPA results for the single fibrosis stages 2 and 3 separately, and (although potentially available only in low numbers) for the stages 0 and 1. The applicant should also discuss the potential consequences for the evaluation of the interim endpoint evaluation based on the reduction of fibrosis stage.

PathAI Response:

A. OPA, PPA, NPA for each individual fibrosis stage

As described in Table 42 in the briefing document, to be adequately powered for accuracy for fibrosis, the sample size would need to be approximately 420. As stated in Issue #4, the following per stage analyses may not be adequately powered (results described in the tables below), and this should be considered when making observations. Additionally, with an "imperfect" or "variable" gold standard, it is difficult to interpret positive percent agreement (PPA) and negative percent agreement (NPA) the same way as true sensitivity and specificity; rather they simply describe the level of positive or negative agreement with a reference in binary comparisons (e.g. F1 vs. other scores). Overall percent agreement (OPA), in addition to the kappa analyses (which consider agreement by chance) in the original submission, should be the primary metrics for overall evaluation of accuracy. Importantly, the other part of the challenge in accurately detecting score change over time and across trials is a lack of standardization across readers and consistency within readers in enrolling accurate populations and detecting true change over time for histologic endpoints. Accuracy should therefore never be considered in isolation without also considering reproducibility, standardization across readers, and reduction of bias (e.g. enrollment bias), etc. The reverse is also true: reproducibility should also be paired with accuracy in considering solutions to meet the challenges with scoring for this context of use.

As requested, the OPA, NPA, and PPA values are described in **Table *1***, **Table *2***, and **Table *3***, respectively, comparing AIM-NASH-assisted read agreement with ground truth (GT) to average Independent Manual Read (IMR) agreement with GT for each individual fibrosis stage. The OPA is significantly higher for AIM-NASH-assisted at F1 compared to average IMR. In general, OPA is comparable or higher for AIM-NASH-assisted reads to GT vs. average IMR to GT, for all fibrosis stages in clinical validation (CV) (**Table *1***). The PPA at F1 is significantly lower for AIM-NASH-assisted (**Table *2***), compared to the average IMR, meaning AIM-NASH-assisted reads agree less with GT on what is considered to be an F1, compared to manual reading.

Table 1: AIM-NASH-Assisted Overall Percentage Agreement with Ground Truth by Level for Fibrosis Stage (CV)

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

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

Table 2: AIM-NASH-Assisted Positive Percentage Agreement with Ground Truth by Level for Fibrosis Stage (CV)

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

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

Table 3: AIM-NASH-Assisted Negative Percentage Agreement with Ground Truth by Level for Fibrosis Stage (CV)

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

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

Of these cases being called differently by AIM-NASH-assisted reads, some cases are being called down to F0, and others up to F2 (& F3, less often), as indicated in **Table 4**. Conversely, NPA at F1 is higher for AIM-NASH-assisted reads than for average IMR reads. NPA is comparable elsewhere except for F3, where NPA is lower than IMR, but PPA is higher than for average IMR reads.

Table 4: AIM-NASH-Assisted Agreement with Ground Truth (GT) per Fibrosis Stage

Fibrosis Stage		AIM-NASH-Assisted					N (GT)
		0	1	2	3	4	
GT	0	10	1	1	0	0	**12**
	1	34	59	62	15	0	**170**
	2	8	70	177	128	1	**384**
	3	0	6	55	444	54	**559**
	4	0	0	1	61	242	**304**
	N (AIM-NASH)	**52**	**136**	**296**	**648**	**297**	-

In the proposed AIM-NASH workflow, pathologist readers may change a score if they disagree by more than 1 for fibrosis or any other component score. In addition, they can also request a recut, restain, or rescan if they believe any of these inputs are not optimal, using the validated overlays to guide their review. However, in order to further explore what the impact of disagreements (1pt or more) between AIM-NASH and manual reading could be on enrollment or endpoint determinations, the OPA, PPA, and NPAs for inclusion criteria (NAS >= 4 with >=1 in each feature, F2/3, F4) and endpoint evaluation (NASH resolution based on score definition, F4 as a measure of cirrhosis endpoint) were presented in the LoI response document submitted 06 Dec 2023. These are further discussed below and additional analyses were performed to explore potential impact on detection of fibrosis response (stage improvement for a subject vs. no change or increase), as requested.

B. Clarification for statement in EMA LoI issue #2 introductory paragraph

"The only drawback in the analyses presented (and as discussed in the discussion meeting) is the inferior performance for the NPA for fibrosis scores 2-3, leading to the identification of less patients eligible for enrolment in clinical trials in the non-cirrhotic population."

Relative to GT for identification of F2/3 biopsies, in CV AIM-NASH identifies approximately the same number of biopsies (**Table 4**, highlighted sample numbers, F2/3 by GT, n=943, by AIM-NASH assisted reads, n=944, for biopsies scored by both GT and AIM-NASH-assisted).

Compared to the average IMR reads, AIM-NASH should potentially enroll more patients as F2/3, considering the significantly larger PPA for F2, lower NPA (Tables 28, 31 in the LoI response document submitted 06 Dec 2023), and the individual reader n's compared to GT. These patients could potentially improve (or not) and may add to response rates, though this possibility is not evaluable here in the manually enrolled and treated CV datasets. Potential impact on response rates is discussed further below in part C.

C. Potential clinical impact

Response Study Purpose and Design:
Because of the observed, lower PPA for AIM-NASH at F1, and since AIM-NASH calls some of the GT F1's', F2's, which is on the border of inclusion for most F2/3 trials (although some trials do enroll limited F1 populations; e.g., MAESTRO, REGENERATE), the following response analyses were performed to explore potential impact on detecting fibrosis improvement as a part of one of the composite histologic endpoints used in phase 2b, and phase 3 NASH trials. In order to explore any impact on response rates in a trial (where AIM-NASH should enroll similar numbers of patients, but perhaps a different subset, compared to GT), paired biopsies from CV, where timepoint data was available, were analyzed. Fibrosis improvement response rates for AIM-NASH-assisted reads, GT reads, and IMR reads for fibrosis improvement were determined for Falcon 2 (n=129 subjects) and REGENERATE (n=224 subjects) trials. In this analysis, fibrosis "response" indicates any reduction in stage from baseline to follow-up, and "non-response" indicates no change or increase in stage. The ground truth score distribution of paired biopsies for the two (2) clinical trials is described in **Table 5**.

Table 5: Ground Truth Score Distribution of Paired Biopsies in Response Analysis

	Clinical Trial			
	Falcon 2 (n=129, 1 biopsy per timepoint)		**REGENERATE (n=224, 1 biopsy per timepoint)**	
Fibrosis Stage	**Baseline %(n)**	**End %(n)**	**Baseline %(n)**	**End %(n)**
0	0.0% (0)	0.0% (0)	0.4% (1)	1.8% (4)
1	0.0% (0)	0.0% (0)	13.8% (31)	11.6% (26)
2	0.8% (1)	1.6% (2)	33.9% (76)	29.9% (67)
3	9.3% (12)	23.3% (30)	44.6% (100)	47.8% (107)
4	89.1% (115)	73.6% (95)	4.5% (10)	7.1% (16)
N/A	0.8% (1)	1.6% (2)	2.7% (6)	1.8% (4)
Total	129	129	224	224

Response rates for AIM-NASH-assisted, IMR, and GT reads were compared overall and for cases where AIM-NASH-assisted called GT F1s, F2s (since AIM-NASH-assisted would have included these patients in an F2/3 trial, if used for prospective enrollment). Overall percent agreement with GT for subject-level response was described for AIM-NASH-assisted and average IMR. As a note, REGENERATE was primarily an F2/3 enrolled trial as scored by a single reader, but also included F1 biopsies with designated comorbidities (e.g., Type-2 Diabetes), so it was possible to make some observations about a small subset of F1 (by GT) patients who were treated in the REGENERATE study. The Falcon 2 study was a compensated cirrhosis, or F4 study, also enrolled by a single reader's scores in the original trial. This population is blinded for treatment arm and, therefore, response rates include both placebo and treated groups and all biopsies were prospectively scored for CV by GT panel, AIM-NASH-assisted readers, and IMRs. It should be noted, as

mentioned above, that this analysis leaves out those that would have been screened in by AIM-NASH-assisted reads and includes some subjects who would have been screened out by AIM-NASH-assisted reads (per the histologic inclusion criteria).

Results and Observations:
The response rates for AIM-NASH-assisted, IMR, and GT reads compared overall for Falcon 2 and REGENERATE, and by each clinical trial are described in **Table *6***, **Table *7***, and **Table *8***. Overall, AIM-NASH-assisted response rate was similar to GT. In particular, for all F1 cases by GT where AIM-NASH-assisted staged as F2 (and would be included in an F2/3 study), there were no cases where GT detected a response and AIM-NASH-assisted did not. In fact, AIM-NASH-assisted detected more responders in that subpopulation.

Table 6: Response Rates for Falcon 2 and REGENERATE

	N (Responders/Total)	**Response Rate (CI Lower, Upper)**	**OPA**
Ground Truth	72/340	21.2% (17.2%, 25.8%)	N/A
AIM-NASH-Assisted	81/332	24.4% (20.1%, 29.3%)	77.7%
IMR			
Reader 1	45/186	24.2% (18.6%, 30.8%)	79.0%
Reader 2	27/81	33.3% (24.0%, 44.1%)	77.8%
Reader 3	43/188	22.9% (17.4%, 29.4%)	79.3%
Reader 4	1/8	12.5% (2.2%, 47.1%)	87.5%
Reader 5	28/94	29.8% (21.5%, 39.7%)	78.7%
Reader 6	56/190	29.5% (23.4%, 36.3%)	72.6%
Reader 7	23/110	20.9% (14.4%, 29.4%)	80.9%
Reader 8	8/36	22.2% (11.7%, 38.1%)	77.8%
Average IMR	N/A	24.4%	79.2%

Table 7: Response Rates for Falcon 2

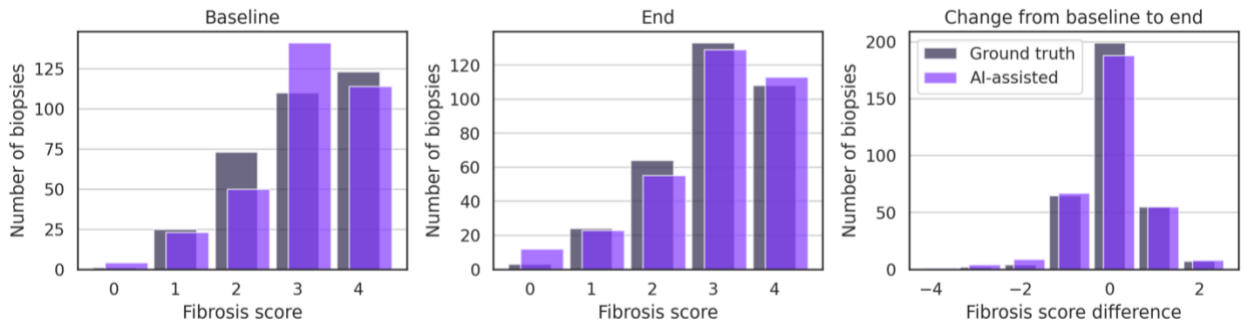
	N (Responders/Total)	**Response Rate (CI Lower, Upper)**	**OPA**
Ground Truth	25/126	19.8% (13.8%, 27.7%)	N/A
AIM-NASH-Assisted	31/126	24.6% (17.9%, 32.8%)	77.8%
IMR			
Reader 1	16/66	24.2% (15.5%, 35.8%)	77.3%
Reader 2	23/66	34.8% (24.5%, 46.9%)	78.8%
Reader 3	10/59	16.9% (9.5%, 28.5%)	81.4%
Reader 4	0/0	N/A	N/A
Reader 5	6/59	10.2% (4.7%, 20.5%)	84.7%
Reader 6	8/59	13.6% (7.0%, 24.5%)	81.4%
Reader 7	15/66	22.7% (14.3%, 34.2%)	75.8%
Reader 8	0/0	N/A	N/A
Average IMR	N/A	20.4%	79.9%

Table 8: Response Rates for REGENERATE

	N (Responders/Total)	Response Rate (CI Lower, Upper)	OPA
Ground Truth	47/214	22.0% (16.9%, 28.0%)	N/A
AIM-NASH-Assisted	50/206	24.3% (18.9%, 30.6%)	77.7%
IMR			
Reader 1	29/120	24.2% (17.4%, 32.6%)	80.0%
Reader 2	4/15	26.7% (10.9%, 52.0%)	73.3%
Reader 3	33/129	25.6% (18.8%, 33.7%)	78.3%
Reader 4	1/8	12.5% (2.2%, 47.1%)	87.5%
Reader 5	22/35	62.9% (46.3%, 76.8%)	68.6%
Reader 6	48/131	36.6% (28.9%, 45.2%)	68.7%
Reader 7	8/44	18.2% (9.5%, 32.0%)	88.6%
Reader 8	8/36	22.2% (11.7%, 38.1%)	77.8%
Average IMR	N/A	28.6%	77.9%

Distribution of change in fibrosis between timepoints, looks to be similar for AIM-NASH-assisted reads vs. GT, although score distributions are different at baseline as depicted in **Figure 1**. One concern could be that, if there's a directional bias of a 1pt difference, where readers cannot override a score, this could lead to a reduction in detected responders. Lowered fibrosis response rates from AIM-NASH-assisted compared to GT were not observed.

Figure 1: Fibrosis Score and Change Distributions for AIM-NASH-Assisted Reads and Ground Truth



In summary, these results indicate that the differences in fibrosis staging by AIM-NASH-assisted compared to manual reads does not result in a noticeable impact on fibrosis response rates compared to GT and does not affect the ability to detect improvement as compared to average IMR in these datasets (one concentrated around F1-F3, and the other concentrated around F3, 4).

Finally, in regard to addressing the substantial issues in inconsistent, non-standardized scoring resulting in inconsistently enrolled populations and inaccurate detection of true

score change over time, the superior reproducibility of AIM-NASH compared to manual reads (Davison et al; Tables 41, 66 in the briefing document) should again be emphasized. In the referenced 2021 Sanyal publication (now published in Journal of Hepatology Communications; 2023), it was demonstrated that lack of standardization between two gold standard panels is still an issue, with inter-panel Kappas being similar to those demonstrated by individual CRN pathologists (Kleiner 2005, 2019). Together, this total body of evidence demonstrates that AIM-NASH, with the currently proposed accept/reject workflow, can help pathologists to accurately score NASH biopsies (increasing accuracy for hepatocellular ballooning, lobular inflammation, and composite scoring involving both NAS components and fibrosis for inclusion and endpoint evaluations), in a reproducible, less biased manner. Additionally, as demonstrated in the additional analyses here, AIM-NASH can effectively detect fibrosis stage improvement over time in a subset of CV samples from an F2/3 and F4 trial. Together, this body of evidence demonstrates a significant improvement in addressing the unmet needs for accurate and reliable NASH clinical trial biopsy scoring for the proposed context of use.

D. Post qualification monitoring in real-world trial setting

It is recognized that F0,1 cases are less represented in training and validation, and the analyses here and in the original submission explore fibrosis performance in a more granular manner and to explore potential clinical-trial impact (accuracy for detecting inclusion criteria, NAS-related and fibrosis endpoints, and fibrosis response rates). As will be described in the response to Issue #5, this highlights the importance of monitoring performance with prospective use in general as well as at these less represented stages, in the "real world" clinical trial landscape.

5. It is acknowledged that the training set used in the model development represents the population currently enrolled in the clinical trials (e.g. patients with fibrosis stages 2-4). However, the treatment strategies and paradigms might change in the future and patients with lower fibrosis stages/disease activity are already being enrolled in the trials (for now in small additional "sub"-groups, but it might change in the future). Therefore, confirmation of the satisfactory performance of the tool in poorly studied extremes is needed. The Applicant is asked to discuss and possibly present more concrete plans on the additional tests that can be performed (potentially post-qualification) to confirm the acceptable performance of the tool in these populations.

PathAI Response:

We acknowledge that treatment strategies and paradigms may change in the future. Confirmation of satisfactory performance of the tool in currently underrepresented stages is included in our monitoring plan for AIM-NASH, the purpose of which is to define how we will collect and analyze data in prospective clinical trials. Data will be collected through agreements with biopharma partners and will be used to determine, implement, and monitor any necessary preventive and corrective actions. The monitoring plan will include, at a minimum:

- Details of the data to be collected and utilized in performing monitoring;

- Effective and appropriate methods and processes to assess the data;
- Benefit-risk analysis;
- Reference to procedures that detail the methods and protocols to communicate effectively with EMA;
- Systematic procedures to identify and initiate appropriate measures

Data will be reviewed relative to the context of use of the tool and analysis will be conducted to address questions such as whether there is: an impact to the benefit/risk of the tool; a need to update design information or context of use; improvement necessary to the usability, performance, or safety of the tool.

A monitoring report will be generated, summarizing the results and conclusions of all analyses, and any impact to the tool's safety and effectiveness will be reported directly to the EMA.

The objectives of the monitoring plan will include detection of population changes; monitoring underrepresented score categories in current clinical trial populations; and tracking AIM-NASH 1-point and 2-point discordance rates. To achieve these objectives, cases will be collected from both the screening and follow-up timepoints, spanning all score categories and levels, and compared to consensus ground truth scores. Further, for the extremes (fibrosis 0 and 1 and lobular inflammation 0), additional cases will be collected, and any discrepancies will be evaluated. Discordance rates will be collected, and significant shifts will be investigated.

Data collected during monitoring will be used to continuously inform PathAI of the performance and continued safety and effectiveness of the tool. Changes to the safety and effectiveness of the tool will be communicated to the EMA. Any necessary verification and/or validation of the tool will be documented and also submitted to EMA.

6. The life-cycle management of the tool is of high importance. The approach of the Applicant on defining major and minor changes to the tool is in principle supported. However, it is still not fully clear how any updates and changes to the tool will be reported to the agency. Also, as mentioned before, small changes might accumulate overtime and require re-validation. It is not clear who will control for that and communicate to the agency. Please provide a clearer proposal for post-validation steps with respect to surveillance of tool performance and reporting of the results of this surveillance.

PathAI Response:

Updates and changes to the tool will be assessed from a risk-based perspective prior to implementation. Any changes that impact the AIM-NASH tool's safety and effectiveness will be evaluated and the necessary verification and/or validation will be documented and submitted directly to EMA before release.

Changes will be evaluated through PathAI's SOP-070, Software Control Change Management Procedure. This procedure defines the general requirements for product, system, and documentation changes where documented evidence of change control is required. When a change is proposed, consideration is given to the potential impact of the change to function, performance, usability, safety and risk, and applicable regulatory requirements for the device and its context of use. Per PathAI procedure, regulatory assessment of proposed changes is required and includes evaluation of the impact of current proposed modifications considered together with any previous modifications since last submitted to a regulatory body to ensure any accumulation has not resulted in significant changes or modifications that require premarket notification.

In addition to evaluating intentional updates and changes made to the tool, PathAI will also monitor performance of the tool as deployed in clinical trials. Agreements with biopharma partners will include language allowing us to surveille AIM-NASH performance once deployed in trials. Tool performance will be compared to the validated tool and significant deviation from validated performance will be investigated and understood by PathAI. Performance changes that could impact the safety and effectiveness of AIM-NASH will be reported directly to EMA via email initially, then continue as appropriate based on discussions.

7. In response to the question 34, the Applicant simply stated that the validation studies were not powered to detect the differences between trials. This is understood. However, clear differences raise questions whether the differences observed might have been, for example, due to trial population or other reasons. The Applicant is asked to discuss any possible reasons that could have contributed to the differences observed.

PathAI Response:
Differences observed in weighted Kappas for fibrosis and lobular inflammation in different trial populations could be attributed to various reasons, including differences in the underlying score distributions and drug candidate effectiveness. However, performance goals were still met for all 3 clinical trials (Table 52 in the briefing document) as performance of the AIM-NASH generally trended with performance of IMR.
The Falcon 1 and Falcon 2 clinical trials have a greater concentration of high fibrosis stages (F3 and F4) compared to the REGENERATE trial, which is expected as REGENERATE enrolled both F2 and F3 patients, Falcon 1 enrolled only F3 patients and Falcon 2 enrolled only F4 patients. The most significant difference in weighted Kappas were observed for fibrosis in the Falcon 2 clinical trial dataset, which included cirrhotic patients (enrollment criteria fibrosis stage 4). The original drug trial did not meet fibrosis-related endpoints, whereas the REGENERATE trial did meet the fibrosis endpoint. It is possible that the morphology within the Falcon 2 trial samples was different and more severe due to the inclusion of cirrhotic patients (e.g., burned-out NASH, severe scarring and necrosis could play a role in challenging pathology evaluations), and, additionally, there could have been an enrichment of borderline F3/4 cases (where increase in inter-reader variability is expected).

In order to investigate the differences, a PathAI internal NASH pathologist was given a sample to assess manually from the Falcon 2 dataset, including all F3 and F4 cases where GT and AIM-NASH disagreed, as well as a random sample of cases where they agreed on F4. Observations made by the pathologist support the statements above, where they concluded that this is a challenging dataset with advanced fibrosis, which could make scoring difficult. The dataset also includes some sub-optimal and/or faint staining and is enriched (32% in this sample, per manual assessment by internal NASH expert pathologist), with challenging cases that are concentrated around the border of fibrosis 3 and 4. The presence of challenging/borderline cases increases the frequency of disagreement and therefore negatively impacts measured Kappas. This investigation supports our hypotheses on possible reasons why the Kappas for Falcon 2 are on the lower range, for both IMR and AIM-NASH reads, compared to other trials. Additionally, the normal architecture of the liver can be significantly changed when there is cirrhosis present, and therefore, determining which foci of inflammatory cells should be included is more difficult, possibly contributing to the lower lobular inflammation Kappas overall.

As indicated in the response to question 34 in LoI response document submitted 06 Dec 2023, the Falcon 2 study was significantly lower in sample size than the other two included trials, however, it still met the performance goal of non-inferiority to the IMRs.