Overview of comments received on the draft Qualification Opinion for Artificial Intelligence-Based Measurement of Non-alcoholic Steatohepatitis Histology (AIM-NASH) tool

Comments from:

Name of organisation or individual
1. **Mark DeLegge**, MD - Head GI/Hepatology Center of Excellence at IQVIA USA
2. **Michelle T. Long**, MD, MSc; Associate Professor of Medicine Boston University Chobanian & Avedisian School of Medicine and International Medical Vice President, MASH, Novo Nordisk
3. Takeda
4. **Piotr Krzeski**, MD, PhD, FFPM

Official address Domenico Scarlattilaan 6 ● 1083 HS Amsterdam ● The Netherlands
Address for visits and deliveries Refer to www.ema.europa.eu/how-to-find-us
Send us a question Go to www.ema.europa.eu/contact **Telephone** +31 (0)88 781 6000 An agency of the European Union

1. General comments

Line number(s) of the relevant text	Stakeholder number	General comment	Outcome
n/a	Mark DeLegge	I strongly support the use of assisted AI in liver pathology assessment	Thank you. Endorsement is noted.
7-15	Piotr Krzeski	The context of use could more clearly stipulate that AIM-NASH is meant to assist a pathologist in a clinical trial setting but not influence the selection of the overall central pathology reading paradigm as the data presented in the opinion does not provide evidence for selection of a specific central pathology reader paradigm whether assisted by AIM-NASH or not.	The tool should be used by a single central pathologist of the trial. In this way, the usual central pathology reading paradigm of the NASH trials (2 independent pathologists + 1 tiebreaker) will be replaced by an AIM-NASH-assisted central single pathologist read. This is reflected in lines 9-11. Hopefully this clarifies.

2. Specific comments on text

Line number(s) of the relevant text	Stakeholder number	General comment	Outcome
1-5	Michelle T Long *(personal comments, not representing Novo Nordisk)*	I would like to provide general comments as a hepatologist caring for patients with MASH for over 15 years and also as a drug developer in the MASH area working for Novo Nordisk, though the views I express are my own. There is a large unmet need for patients living with MASH, despite all the efforts over the last 20 years to develop effective therapies. A major challenge to the field relates to the grading and staging of MASH disease activity and fibrosis, which are critical factors evaluated within clinical trials. Additionally, we rely on liver biopsy and the assessment of the disease activity grade and stage (now with human pathologists, unassisted) for entry into clinical trials. Clinical practice has advanced such that liver biopsy is not so often used for diagnosis, so most patients undergoing liver biopsy for clinical trials primarily do this for research purposes. Over the last several years, it has become clear that variability between pathologists, even expert pathologists, leads to a high screen failure into MASH trials and also challenges with assessing the histology-based endpoints. It is not uncommon for a patient to be rejected from one study, based on the interpretation of the liver biopsy, only to be accepted by another study considering the same criteria. With such variability, sponsors often choose to rely on a small number of	Thank you for this supportive comment. Arguments are supported and noted.

Line number(s) of the relevant text	Stakeholder number	General comment	Outcome
		expert pathologists within their studies, which can be ongoing for several years. The reliance on specific individuals for screening into clinical trials and assessing endpoints for regulatory approval poses significant risks to the trial programs. Tools such as AIM-NASH provide much needed innovation to the assessment of liver biopsy samples for both study entry and throughout the study. This tool has the potential to not only simplify the operational complexities of clinical trials, but also assist in addressing some of the variability within and between pathologists evaluating the liver biopsy samples during the study. Of course, the AIM-NASH tool is a supplement to the review of the pathologist within the study; however, it is expected that providing the AIM-NASH tool to the pathologist in the study will also allow for more consistency, should different pathologists be involved with the study. It is my opinion that the availability of the AIM-NASH tool will significantly improve the histology setup within trials, improving the efficiency and reducing the complexity of the trial program. I also think that it will greatly improve the screening process and there will be fewer cases of patients failing screening based on differences of opinion between pathologists. This is important, especially since patients undergo biopsies, and face the risk of the procedure, usually only for research purposes. I am in full support of AIM-NASH tool and	

Line number(s) of the relevant text	Stakeholder number	General comment	Outcome
		thank the CHMP for their careful consideration of the AIM-NASH proposal.	
33-35	Takeda	Please clarify if it means that a single central pathologist, assisted by AIM-NASH, will be sufficient to conduct histology evaluation at baseline and follow-ups for the study histological outcomes (primary or secondary) for all Phase 2 and Phase 3 MASH clinical trials where histologic evaluation of liver tissue is used as part of the inclusion criteria, and/or efficacy evaluation.	That is correct.
42-43	Takeda	Please clarify whether it is necessary to monitor the single central pathologist's (or pathologist assisted by AIM-NASH) performance during the study after approval of their qualification prior to the study.	The question is not entirely clear. It is assumed that the commentor asks whether the performance of the qualified single pathologist (using AIM-NASH) should be monitored during the study. If so, please see line 284-306 in the opinion document on the monitoring plan. The objectives of the monitoring plan will include among others tracking AIM-NASH 1-point and 2-point discordance rates. Discordance rates will be collected, and significant shifts will be investigated. Additionally, it is recommended that the central clinical trial lab or other responsible data management party overseeing the pathology reads with AIM-NASH should also monitor according to their own pre-specified plans outlined and agreed upon with the Sponsor of the trial.
369-370	Takeda	Tables 3 and 4: Please clarify whether the ground truth of each case established by random selection of	Tables 3 and 4 in the Draft QO list the datasets used for AIM-MASH training and testing and were

Line number(s) of the relevant text	Stakeholder number	General comment	Outcome
		2 expert liver pathologists with a third pathologist serving as tiebreaker or by the average/median scores of the 10 NASH pathologists from PathAI expert contributor network.	performed before any validation (AV, CV) studies were conducted. The first column in tables 3, 4 describes the training dataset characteristics, derived from the original central reader scores from the trial, not from the prospectively collected scores for training. Further, as the process for training is different from validation, the approach for utilizing scores for each part of the process was different. For training: The scoring models were trained with overlays and the individual scores from multiple pathologists, the selection of which was based on MASH expertise and availability. Each of the 10 selected pathologists scored a random set of slides from the training set. So, each slide had an overlay and multiple scores from pathologists drawn from the pool of 10 to train the GNN scoring model and to calculate the statistical bias for a single pathologist. For testing (standalone analytical validation; SAV): The second column in tables 3, 4 describes the test dataset characteristics, as determined by a statistical consensus (mode/median), another gold standard method being used in trials, from prospectively collected, individual manual reads supplied by 3 separate MASH pathologists. These

Line number(s) of the relevant text	Stakeholder number	General comment	Outcome
			pathologists were qualified by PathAI in addition to having MASH expertise and reading experience and were not involved in training.
147-148	Takeda	Please provide information on how and when this tool will be applicable for images obtained on additional scanners such as GT450, as AT2 may be discontinued and replaced by GT450 (also from Leica). This will impact adoption of the tool for long Ph3 trials.	The Applicant expressed plans to validate additional scanners in the near future. Equipment changes will be part of the life-cycle management. According to the Applicant's proposal, any changes (including equipment changes) that impact the AIM-NASH tool's safety and effectiveness will be evaluated and the necessary verification and/or validation will be documented and are intended to be submitted to EMA before release. Please see 'Life-cycle management' section in the Qualification Opinion.
1241-46	Piotr Krzeski	The "Ground Truth" in the clinical validation stage consisted of two panels of 2 pathologists and an adjudicator (2+1 paradigm). This central pathology reader paradigm is frequently used in MASH trials. I would welcome metrics of the "GT" performance such as WKs for intra- and inter-reader variability (between the two "2+1" panels) to allow external validity estimation.	In the CV study, the agreement of the test method of reading (in this case, AI-assisted) with a reference "Ground Truth" or gold standard read is compared to the agreement of an unassisted, standard read with that same Ground Truth. As is common in AI or digital pathology validation studies, the dataset can be derived from multiple subsets with different readers, and the results pooled (e.g., https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN200080.pdf, https://www.fda.gov/drugs/resources-information-approved-drugs/intellisite-pathology-solution-pips-philips-medical-systems). In this study, two sets of Ground Truth panels were selected based on pre-

Line number(s) of the relevant text	Stakeholder number	General comment	Outcome
			defined requirements of MASH expertise and trial reading experience and a defined qualification procedure, and these panels each read a different subset of the overall CV dataset (please see line 669-674). Although each GT panel did not read the same slide set (and therefore, we cannot calculate inter-panel agreement from them), the overall kappas with the study reads (assisted or unassisted) were monitored and determined to be within the range of kappas described in the literature from other MASH experts. These readers have MASH expertise and previous MASH trial reading experience and represent the various consensus panels used in multiple MASH trials. Additionally, the read method and pathologist selection requirements were the same as in some earlier studies (e.g. https://doi.org/10.1097/HC9.0000000000000325). Also, no inter-reader variability was provided for a single GT panel. But due to the reasons mentioned above, it is expected that these values would be similar to values already reported from the literature. Nevertheless, the lack of these data remains a limitation of this study. In summary, the selection process of the expert readers, the use of the gold standard panel methodology, and the monitoring of the various read comparison (weighted kappas) using similar

Line number(s) of the relevant text	Stakeholder number	General comment	Outcome
			literature values allow to some extent for external validation estimation. However, the absence of the information on the variability among the readers within each GT panel remains a limitation of this study.
184	Piotr Krzeski	Consider changing "Trial Site" to "Trial Site or vendor with appropriate capability" to avoid confusion with investigative site.	Thank you for the suggestion. For clarity 'Trial Laboratory' was added to express that scanning is not per se bound to a clinical trial site. "Qualified WSI Scanner at Trial Site/ ***Trial Laboratory***: Slides must be scanned at a CAP/CLIA (or European equivalent, ISO 15189) compliant laboratory with the validated Aperio AT2 scanner at 40X magnification."
779	Piotr Krzeski	Scores are missing in Table 10.	Many thanks for spotting this. The scores were added.
219-232	Piotr Krzeski	In the proposed workflow it is not clear how single-point deviations for individual histology components are treated (accepted, rejected for consensus?).	In case of disagreement by +/- 1 point, the score produced by AIM-NASH will be accepted. This has been clarified in the sentence. "If the pathologist accepts these scores (***within +/- 1 point per individual feature***), they will record their agreement and sign-out the case."