









ORIGINAL ARTICLE

OPEN

Utility of pathologist panels for achieving consensus in NASH histologic scoring in clinical trials: Data from a phase 3 study

Arun J. Sanyal¹  | Rohit Loomba²  | Quentin M. Anstee³  | Vlad Ratziu⁴  |
 Kris V. Kowdley⁵  | Mary E. Rinella⁶ | Stephen A. Harrison⁷  |
 Murray B. Resnick⁸  | Thomas Capozza⁹ | Sangeeta Sawhney⁹ |
 Nirav Shelat⁹ | Zobair M. Younossi¹⁰ 

¹Department of Internal Medicine, Division of Gastroenterology, Hepatology and Nutrition, Virginia Commonwealth University, Richmond, Virginia, USA

²Division of Gastroenterology and Hepatology, Department of Medicine, University of California San Diego, La Jolla, California, USA

³Translational & Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, UK

⁴Sorbonne Université, Institute of Cardiometabolism and Nutrition, Pitié Salpêtrière University Hospital, Paris, France

⁵Liver Institute Northwest, Seattle, Washington, USA

⁶Pritzker School of Medicine, University of Chicago, Chicago, Illinois, USA

⁷Pinnacle Clinical Research, San Antonio, Texas, USA

⁸Department of Pathology and Laboratory Medicine, Brown University, Providence, Rhode Island, USA

⁹Intercept Pharmaceuticals, Inc., Morristown, New Jersey, USA

¹⁰Inova Medicine, Inova Healthy System, Falls Church, Virginia, USA

Correspondence

Arun J. Sanyal, Department of Internal Medicine, Division of Gastroenterology, Hepatology and Nutrition, Virginia Commonwealth University, 907 Floyd Avenue, Richmond, VA 23284, USA.
 Email: arun.sanyal@vcuhealth.org

Abstract

Background: Liver histopathologic assessment is the accepted surrogate endpoint in NASH trials; however, the scoring of NASH Clinical Research Network (CRN) histologic parameters is limited by intraobserver and interobserver variability. We designed a consensus panel approach to minimize variability when using this scoring system. We assessed agreement between readers, estimated linear weighted kappas between 2 panels, compared them with published pairwise kappa estimates, and addressed how agreement or disagreement might impact the precision and validity of the surrogate efficacy endpoint in NASH trials.

Methods: Two panels, each comprising 3 liver fellowship-trained pathologists who underwent NASH histology training, independently evaluated scanned whole slide images, scoring fibrosis, inflammation, hepatocyte

Abbreviations: CRN, Clinical Research Network; FDA, US Food and Drug Administration; WSI, whole slide image.

ClinicalTrials.gov Identifier: NCT02548351

Supplemental Digital Content is available for this article. Direct URL citations are provided in the HTML and PDF versions of this article on the journal's website, www.hepcommjournal.com.

This is an open access article distributed under the terms of the Creative Commons Attribution-Non Commercial-No Derivatives License 4.0 (CCBY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

Copyright © 2023 The Author(s). Published by Wolters Kluwer Health, Inc. on behalf of the American Association for the Study of Liver Diseases.

ballooning, and steatosis from baseline and month 18 biopsies for 100 patients from the precirrhotic NASH study REGENERATE. The consensus score for each parameter was defined as agreement by ≥ 2 pathologists. If consensus was not reached, all 3 pathologists read the slide jointly to achieve a consensus score.

Results: Between the 2 panels, the consensus was 97%–99% for steatosis, 91%–93% for fibrosis, 88%–92% for hepatocyte ballooning, and 84%–91% for inflammation. Linear weighted kappa scores between panels were similar to published NASH CRN values.

Conclusions: A panel of 3 trained pathologists independently scoring 4 NASH CRN histology parameters produced high consensus rates. Inter-panel kappa values were comparable to NASH CRN metrics, supporting the accuracy and reproducibility of this method. The high concordance for fibrosis scoring was reassuring, as fibrosis is predictive of liver-specific outcomes and all-cause mortality.

INTRODUCTION

Current regulatory guidance calls for liver histology assessment as a composite of 4 parameters of NASH (steatosis, inflammation, ballooned hepatocytes, and fibrosis) for study entry and as a surrogate efficacy endpoint in phase 2b and phase 3 clinical trials in NASH.^[1–4] The presence of steatohepatitis is critical for diagnosing NASH and drives fibrogenesis; the fibrosis stage determines clinical outcomes.^[5] However, there are known limitations to liver histology assessments, and the subjectivity in liver histology interpretation, even among expert pathologists, can lead to low inter-reader and intra-reader concordance.^[6–9] Furthermore, the composite endpoints, as defined by regulatory guidance for NASH clinical trials, require multiple criteria beyond the fibrosis stage to be achieved simultaneously, and all 3 elements of steatohepatitis are required to remain stable for a patient with an improved fibrosis stage to be considered a therapeutic responder.^[1,2] Thus, clinical trial design and power calculations based on these composite endpoints may underestimate the true treatment effect.^[4,7]

The NASH Clinical Research Network (CRN) approach is the most widely accepted, semiquantitative histology scoring method supported by rigorous performance data.^[10] Despite different populations and vastly different subject numbers, the weighted kappa scores for the fibrosis stage and each of the Nonalcoholic Fatty Liver Disease Activity Score parameters were remarkably similar between 2 published NASH CRN studies.^[10,11] In both studies, the mean kappa scores were highest for fibrosis stage (0.84 and 0.75) and steatosis (0.79 and 0.77) and lowest for lobular inflammation (0.45 and 0.46) and ballooning (0.56 and 0.54).^[10,11] We report a

consensus panel approach,^[6] which provides a robust method for achieving high concordance for fibrosis and steatosis. However, the poor concordance for lobular inflammation^[7] and ballooning^[5] can still have a significant impact on the assessment of the primary endpoint for a registrational trial; for example, a patient with a full stage of fibrosis improvement is not considered a responder if hepatocyte ballooning is worse than at the baseline biopsy, even though the primary determinant of clinical disease progression is the change in fibrosis over time^[12,13] and hepatocyte ballooning has not been correlated with clinical outcomes.^[14]

Many clinical trials in NASH have used a single-reader approach, which can be subject to temporal bias and intra-reader variability.^[7] The use of 2 independent readers has been recognized as suboptimal, with poor inter-reader concordance observed across NASH parameters.^[7] The US Food and Drug Administration's (FDA) Division of Hepatology and Nutrition recently published a manuscript and presented a webinar with several suggestions to reduce the discordance in liver histology interpretation,^[6,15] including the following:

- (1) A standardized procedure for processing biopsy slides.
- (2) Prespecified details of liver biopsy interpretation.
- (3) Improvement in pathologists' training both before and during the study.
- (4) Recommendation of a minimum of 2 pathologists, with a third if there is disagreement; the same slide is read by all pathologists.

Herein, we present a consensus panel methodology aimed at following the FDA's recommendations and

generating data to support the original REGENERATE results and interpretations. The method employs liver fellowship-trained board-certified pathologists who have undergone standardized proficiency testing specific to reading NASH CRN fibrosis stage and Non-alcoholic Fatty Liver Disease Activity Score parameters. The goal was to determine whether the interpanel and intrapanel kappas from this analysis are comparable to the published NASH CRN data.

METHODS

This was a methodology substudy of the ongoing phase 3 REGENERATE trial of obeticholic acid in precirrhotic NASH, with previously published results.^[16] REGENERATE is being conducted in accordance with the principles of the Declaration of Helsinki and Istanbul and the Good Clinical Practice guidelines of the International Council for Harmonization. All clinical sites participating in the trial obtained approval from the institutional review board, and all patients were provided with written informed consent before enrollment in the trial. All selected study pathologists were trained in the use of the Intercept Central Histology Manual (including the predefined, study-specific scoring criteria) to align with the NASH CRN scoring system definitions, and harmonization sessions consisting of representative and edge cases were conducted by Murray B. Resnick. Steatosis was scored on a 0-to-3 scale; lobular inflammation, 0–3;

hepatocyte ballooning, 0–2; and fibrosis staging, 0–4. Pathologists were also trained to view and score images using the vendor platform.^[10]

Our proposed consensus panel methodology was tested using 2 separate panels (panel A and panel B), each with 3 pathologists who analyzed scanned whole slide images (WSIs) from 100 subjects with liver fibrosis due to NASH who had baseline and 18-month biopsies from the REGENERATE study (ie, 200 biopsies total). Subjects were randomly selected based on fibrosis stage (F1, 16.67%; F2, 33.33%; F3, 50%) and treatment arm to ensure representation across study treatment arms. The scanned images were hosted on a dedicated online platform (PathAI), which could be accessed simultaneously by multiple pathologists to allow for independent reads. To ensure that the WSIs and the reading of the images were of high quality, we used a rigorous process to validate the glass-to-digital transition, including validation of the WSI viewer by each pathologist with predefined acceptance criteria. Simultaneous viewing of specific WSIs was also used for joint panel reads when an agreement between at least 2 of the 3 pathologists could not be reached in phase 1 (Figure 1). The PathAI platform was used only to view slide images and record scores; no artificial intelligence tools were used to score or to assist in scoring the WSIs. Each subject contributed 2 liver biopsies, one at baseline and one at month 18, with 2 WSIs (hematoxylin/eosin to score steatosis, inflammation, and hepatocyte ballooning grades, and trichrome to score fibrosis stage) from each time point, for a total of 400 WSIs from 200 biopsies.

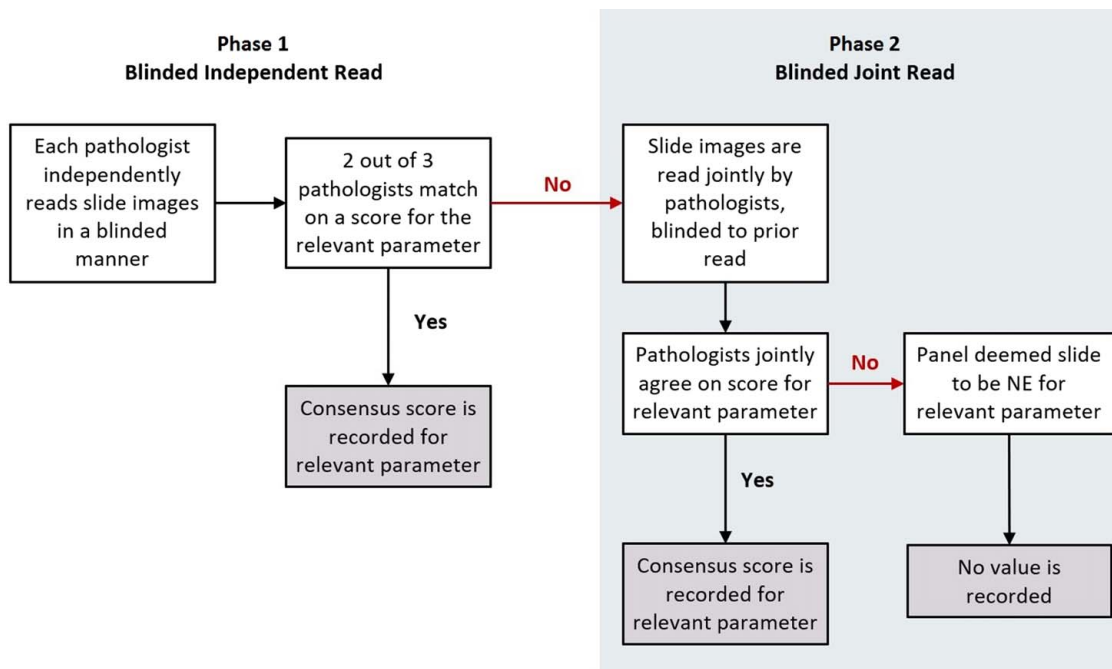


FIGURE 1 Flow diagram of slide histopathology assessment using proposed consensus panel 2-stage approach. Parameters assessed by pathologists: fibrosis, lobular inflammation, ballooning, and steatosis. Abbreviation: NE, nonevaluable.

In phase 1, all 3 pathologists from each of the 2 panels read each subject's slide images and entered the scores for fibrosis stage, lobular inflammation, hepatocyte ballooning, and steatosis into a database (Figure 1). Readers were blinded to treatment assignment, time point, subject identification, and each other's reads. After all images had been processed through phase 1, the scores for the 4 parameters from the 3 pathologists within a panel were compared. For each parameter, if at least 2 of the 3 pathologists agreed on a score (ie, if consensus was reached), it was deemed the consensus score. If consensus was not reached, the slide was considered discordant for that parameter.

Discordant slides were reevaluated in phase 2 (see Figure 1) by a joint panel consisting of the same 3 pathologists who previously reviewed the slides. Only the discordant parameter was rescored by the joint panel, and all readers were blinded to prior reads. The 3 pathologists from the joint panel convened virtually, and all WSIs marked for phase 2 were read jointly by the panel. At this stage, the following rules were applied for the WSIs:

- If the panel reached a consensus on a parameter that was discordant in phase 1, that consensus score was entered.
- If the panel could not reach a consensus following the joint read, the slide was deemed not evaluable for the discordant parameter(s), and no consensus score was entered.

To compare the magnitude of agreement between the 2 panels (interpanel concordance), kappa scores were determined using both the Shrout–Fleiss method^[17] with a 2-way mixed model and the Cicchetti–Allison method.^[18] The Shrout–Fleiss methodology was employed to match

the Kleiner analysis.^[10] The Cicchetti–Allison methodology was employed to match the kappa scores published by Davison et al.^[7] Intrapanel concordance (ie, agreement between at least 2 pathologists within a panel) was also assessed for each parameter using the Shrout–Fleiss methodology. To assess the extent to which the proposed process was responsible for enhancing the panel's performance (vs. that of a single pathologist), kappas were also calculated for all pairs among the 6 pathologists' phase 1 independent reads.

RESULTS

Slide quality

Overall, ~5%–10% of the trichrome slides (for scoring fibrosis stage) were deemed nonevaluable by at least one pathologist, mostly due to stain color quality. For hematoxylin/eosin slides (for scoring steatosis, inflammation, and ballooning grades), the percentage of nonevaluable slides for any parameter was <5% across both panels.

Intrapanel agreement

Using the consensus panel approach, the fundamental reader unit is the panel and not the individual pathologists within a panel. Following blinded independent reads by the pathologists, consensus between at least 2 of the 3 pathologists within a panel was reached in ~90% of slides at phase 1 (Figure 2) for each of the 4 parameters in both panels. In phase 1 (ie, independent blinded read), steatosis had the highest agreement rates across the 2 panels (range, 97%–99%) followed by fibrosis stage (range, 91%–93%). The range of agreement across the 2 panels was 88%–92% for hepatocyte ballooning and 84%–91% for inflammation. Examples of slides that went to the joint panel review are shown in Supplemental Figure S1, <http://links.lww.com/HC9/A667>. Panels allow for multiple chances for

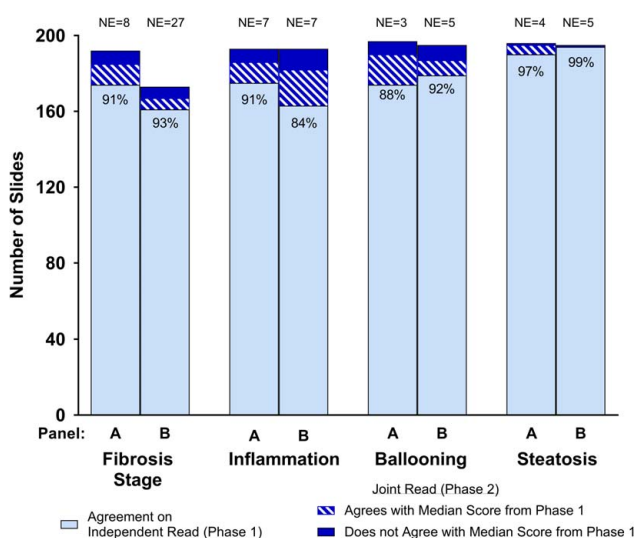


FIGURE 2 Concordance within panels A and B at phase 1 of the consensus panel approach. Denominator for slides with agreement on independent read is based on the number of evaluable slides.

Abbreviation: NE, nonevaluable.

TABLE 1 Concordance (agreement between ≥ 2 readers) and pairwise kappas within panels

Parameter	Methodology substudy		Kleiner et al ^[10] (N = 32)
	Panel A ^a (N = 100)	Panel B ^a (N = 100)	
Fibrosis	0.61–0.75	0.63–0.71	0.85
Lobular inflammation	0.23–0.61	0.38–0.57	0.60
Ballooning	0.25–0.75	0.44–0.64	0.66
Steatosis	0.69–0.81	0.79–0.87	0.83

^aRanges from pairwise kappas from pairs within panel.

^bValues represent the average intra-reader kappa.

TABLE 2 Comparison of interpanel concordance with published literature

Parameter	Shrout–Fleiss weighted kappa			Cicchetti–Allison weighted kappa	
	Methodology substudy; panels A vs. B (N = 100)	Kleiner et al ^a . ^[11] (N = 446)	Kleiner et al ^a . ^[10] (N = 32)	Methodology substudy panels A vs. B (N = 100)	Davison et al ^b . ^[7] (N = 339)
Fibrosis	0.82	0.75	0.84	0.71	0.48
Lobular inflammation	0.60	0.46	0.45	0.46	0.33
Ballooning	0.62	0.54	0.56	0.51	0.52
Steatosis	0.89	0.77	0.79	0.83	0.61

Methodology substudy results are based on nonmissing data.

^aAverage of the pairwise kappas.

^bPairwise kappas.

agreement, and results from this study were generally consistent with those of Kleiner et al (Table 1).^[10]

Interpanel agreement

Kappa scores between the 2 panels for the 4 parameters of interest are shown in Table 2, demonstrating high concordance for fibrosis and steatosis and moderate concordance for lobular inflammation and ballooning. These kappa scores compared favorably with scores from prior methodologies using both the Shrout–Fleiss and Cicchetti–Allison approaches (Table 2). The Cicchetti–Allison kappas between panels A and B were comparable to those calculated by Davison et al.^[7] More importantly, the Shrout–Fleiss kappas between panels A and B are similar to the NASH CRN benchmark for fibrosis, hepatocyte ballooning, inflammation, and steatosis.^[10]

DISCUSSION

The complex analysis of NASH histopathology, with 2 stains and 4 categorically scored parameters grouped in composite endpoints for the 2 major histologic outcomes produces variability and likely underestimates true therapeutic benefit.^[6–9] Repeat biopsies for response assessment have the potential for regression to the mean to be seen as apparent fibrosis regression; however, sampling and observer variability also contribute to an observed placebo response in NASH trials.^[19] Sampling variability can be mitigated to some extent by ensuring a sufficient number of portal tracts and a sufficient length of biopsy specimen (eg, a threshold of 25 mm), but it cannot be completely avoided given the heterogeneity of the disease throughout the liver.^[19,20] Due to its inherent subjectivity, the well-known variability in histopathology interpretation across many disease states can be difficult to control, even with targeted and iterative training. Efforts to address this subjectivity have gained traction in recent

years to aid the clinical development of drugs targeting NASH. Historical use of single readers in studies simplified the process but raised questions about accuracy and reproducibility. While more logistically cumbersome than the single-reader approach, the consensus panel approach may increase reproducibility and decrease variability in histology scoring but does not address the potential underestimation of the true treatment effect associated with the NASH CRN ordinal scoring system.

The original NASH CRN scoring system assessed inter-reader agreement based on biopsies from 32 adult subjects.^[10] Scaling this process to a much larger sample size (N = 446) yielded similar values, notwithstanding the slight decrease in the fibrosis kappa score.^[11] Despite this, the FDA suggested potential solutions for increasing inter-reader and intra-reader concordance, including using a central reading method in which 2 pathologists read all slides independently and then together if they did not match on a parameter, with a third pathologist available as a tie-breaker.^[15] Using this method, the phase 2 study of semaglutide in NASH by Newsome and colleagues (N = 320) yielded much smaller kappas. The 2 pathologists agreed on all variables in only 24% of the assessments and on individual components in 62%–75% of the assessments, necessitating a consensus call to reach an agreement^[21] and further illustrating the challenges in assessing a complex histology endpoint based on 4 parameters. Inter-reader and intra-reader variability was also assessed in paired liver biopsy samples from the EMMINENCE study in which digitized slides were read independently by 3 histopathologists using the NASH CRN scoring system. Inter-reader kappas demonstrated lower reliability than those in NASH CRN.^[7]

To find a suitable alternative consistent with the published recommendations from the FDA,^[6] we looked to biopsy scoring methods that have been effectively employed in other therapeutic areas. In oncology, a successful approach has been implemented that uses 2 initial independent readers and a third tie-breaker reader, with joint adjudication where necessary.^[22] Our

proposed method was adapted from this approach, but it was modified to address the inherent complexity of scoring the histologic features of NASH. The use of 3 independent readers in phase 1 essentially fulfills the role of having 2 readers and a tie-breaker because we defined consensus as the agreement between at least 2 of the 3 readers, with digital assessment allowing for efficient collection of the 3 independent reads, each with 4 scored parameters, in parallel. The fellowship-trained liver pathologists in our study were selected based on reported experience in reading liver biopsies and underwent targeted training, harmonization sessions, and proficiency testing. It is promising in that it generated inter-reader concordance that has similar kappas as the original NASH CRN assessments^[10] while also potentially mitigating bias or inaccuracies from the use of a single central reader or 2 central readers reading separate sets of samples.

We used 2 analytical methodologies to appropriately compare our results with those of Kleiner et al and Davison et al.^[7,10,11] The values reported by Kleiner et al using the Shrout–Fleiss method have generally been considered the gold standard for concordance among raters when using the NASH CRN scoring system.^[10,11] We used the Cicchetti–Allison method to compare our results to those of Davison et al, as they were also evaluating the reliability of NASH CRN scoring and the effect of hepatopathologist interpretations on NASH clinical trial endpoints.^[7] While this method yielded smaller kappas, the agreement between panels A and B using the consensus panel method is comparable to or higher than the Davison kappas and almost approaches those of the NASH CRN expert panel reported by Kleiner et al.^[10,11]

Our high inter-reader and intra-reader agreement for fibrosis and steatosis are consistent with the literature, suggesting that it is easiest to obtain concordance on these 2 parameters. Also consistent with the literature are the lower inter-reader and intra-reader kappas we observed for hepatocyte ballooning and inflammation, as the grading of hepatocyte ballooning and lobular inflammation are not well defined.^[23] With all 4 components contributing to the assessment of a liver biopsy for NASH diagnosis and as the primary endpoint in NASH clinical trials, the variability of hepatocyte ballooning assessments, in particular, remains a significant issue to be addressed. The use of precisely defined and accepted criteria is recommended for scoring of each histologic feature when designing clinical trials.^[23]

Despite the concordance achieved with our panel-based approach, the method is not as efficient as a single-reader process, and the risk of bias from a dominant voice still exists for the small fraction of slides that were reviewed in phase 2, although not to the extent that is observed with only 2 pathologist readers. In addition, using a joint panel over a tie-breaker

pathologist in the adjudication step for discordant reads potentially sacrifices logistical efficiency and should be studied to confirm the former's advantages in NASH. While several publications have reported the use of WSI in liver biopsies and the results support noninferiority to the use of glass slides, digital pathology methods are likely to be affected by slide quality.^[24,25] As such, the evaluation of slide quality should be a routine part of clinical trial reporting. In this analysis, ~5%–10% of trichrome slides were regarded as nonevaluable, which could impact the robustness of the results. This was attributed, in part, to stain degradation over time and should be addressable through earlier digitization for future reads.

Although there is a range in the kappas between the pairs within panels, the agreement rate between any 2 of 3 pathologists is high (~90%; [Figure 2](#)) for all 4 parameters during phase 1. Furthermore, based on prior central pathologist reading data, it was observed that even the best-trained and most experienced pathologists might not perform equally well for all 4 parameters. For the proposed reading method, using a panel of 3 pathologists allows for the specific skill set (ie, high skill in reading a specific parameter) that is common/strongest between any 2 of 3 pathologists to drive the overall scoring of that parameter. This may further help to explain why this proposed method yields such high agreement. It should also be noted that the current study involved a total of 6 pathologists, each reading 400 slides; the high concordance rates achieved in this study might not be reproducible when WSIs are read over a longer time by a larger number of pathologists in phase 3 clinical trials of NASH.

This substudy of REGENERATE also provided additional qualitative insights that may further enhance the performance of the consensus panel method in the clinical trial setting (and other panel-based reader protocols). Because pathology training focuses on diagnoses rather than specific skills for quantitation, as the NASH CRN methodology requires, providing more specific definitions and targeted training of pathologists in areas with high variability to support precision in scoring may be useful. Specifically, clear instructions on how inflammatory foci should be separated, how many cells should define a focus, distinctly defining terms such as *few* and *many* for ballooning scores, and clarifying the inclusion of droplets/microvesicles for steatosis scores would be beneficial.

Regardless of the biopsy reading approach (ie, single reader or consensus panel), the NASH CRN ordinal rating scale may underestimate histologic improvements and thus clinical benefits in NASH clinical trials due to the broad range of disease included in each of the stages. Incremental improvements in the width of the septa may indicate significant regression of fibrosis^[26,27] that is not adequately captured by this scoring system. An

alternative for fibrosis staging is the Ishak scoring system, which categorizes fibrosis into 6 stages rather than 4. While this system may provide more descriptive and comprehensive information on fibrosis and increase sensitivity, it also has limitations.^[28] In a study investigating interobserver agreement for categoric and quantitative scores of liver fibrosis, the kappas between any 2 pathologists ranged from 0.57 to 0.67 for Ishak staging and from 0.47 to 0.57 for NASH CRN staging. The authors also showed that categoric scores such as Ishak or NASH CRN perform less well than quantitative scores.^[29] Noninvasive tests, such as transient elastography, may also show improvements even in subjects with no histologic change in fibrosis as measured by an ordinal scale^[30]; however, it should be noted that the degree of change sufficient to judge a true response using noninvasive tests has not yet been defined.

Because fibrosis is the strongest predictor of clinical outcomes, including death and liver-related morbidity, it is the primary surrogate outcome accepted by the FDA as a marker for efficacy in NASH clinical trials.^[15] Reliable measures of fibrosis improvement are thus critical for developing NASH treatments. Consistent with previous findings, fibrosis staging had the highest concordance rates among the pathologists in our study, providing additional confidence in the antifibrotic effect size estimates of our REGENERATE clinical trial of obeticholic acid. Even with the improvements in consistent scoring through a consensus panel method, the NASH CRN nominal scoring system potentially underestimates the benefit of an antifibrotic therapeutic.

CONCLUSIONS

Our consensus panel method for reading NASH biopsy slides from an ongoing clinical study has an interobserver agreement that is similar to that reported by the NASH CRN pathologists considered the gold standard in the field. This approach potentially mitigates variability and provides an efficient and reliable method for reading digitized slides that can increase confidence in treatment effect sizes in clinical studies.

DATA AVAILABILITY

All data supporting the findings of this analysis are available within the article. The REGENERATE study is ongoing at the time of publication, and, owing to its proprietary nature, data from the study will not be made publicly available.

AUTHOR CONTRIBUTIONS

All authors contributed to the conception or design of the work; acquisition, analysis, or interpretation of data; drafting the manuscript or revising it critically for important intellectual content; and final approval of the version to be published.

ACKNOWLEDGMENTS

Medical writing assistance was provided by Peloton Advantage, LLC, an OPEN Health company, and MedLogix Communications, LLC, Itasca, Illinois, and was funded by Intercept Pharmaceuticals, Inc. The authors thank the PathAI pathologists LiJuan Wang, Sandy Liu, Sanjay Kakar, Dhanpat Jain, Evgeny Yakirevich, Robert Najarian, Xiuli Liu, Zhiyong Ren, Hannah Chen, Kevin Anderson, Andres Matoso, Alexander Christakis, Carl Jacobs, Michael Idowu, and Naziheh Assarzagdegan for their important contribution to this study.

FUNDING INFORMATION

This analysis was sponsored by Intercept Pharmaceuticals, Inc. Medical writing assistance was provided by MedLogix Communications, LLC, and was funded by Intercept Pharmaceuticals, Inc.

CONFLICTS OF INTEREST

Arun J. Sanyal has stock options for Genfit, Exhalenz, Tiziana, Indalo, NorthSea, Durect, HemoShear, and Rivus. He has been a consultant for Intercept Pharmaceuticals, Inc., AstraZeneca, Amgen, Salix, Janssen, Gilead, Mallinckrodt, Terns, Merck, Siemens, 89bio, NGM Bio, Poxel, Boehringer Ingelheim, Eli Lilly, HemoShear, Bristol Myers Squibb, Novartis, Novo Nordisk, Pfizer, Albireo, Regeneron, Genentech, Anylam, Roche, Madrigal, Inventiva, Covance, ProSciento, HistolIndex, PathAI, and Genfit. He has received research grant support to Virginia Commonwealth University from Intercept Pharmaceuticals, Inc., Gilead, Bristol Myers Squibb, Eli Lilly, Merck, Boehringer Ingelheim, Novo Nordisk, Fractyl, Mallinckrodt, Madrigal, Inventiva, Novartis, and Pfizer. He receives royalties from Elsevier and UpToDate. Rohit Loomba has served as a consultant for Aardvark, Altimmune, Anylam/Regeneron, Amgen, Arrowhead, AstraZeneca, Bristol Myers Squibb, CohBar, Eli Lilly, Galmed, Gilead, Glympse, HighTide, Inipharm, Intercept Pharmaceuticals, Inc., Inventiva, Ionis, Janssen, Madrigal, Metacrine, NGM Bio, Novartis, Novo Nordisk, Merck, Pfizer, Sagimet, Theratechnologies, 89bio, Terns, and Viking; has received research grants from Arrowhead, AstraZeneca, Boehringer Ingelheim, Bristol Myers Squibb, Eli Lilly, Galectin, Galmed, Gilead, Hanmi, Intercept Pharmaceuticals, Inc., Inventiva, Ionis, Janssen, Madrigal, Merck, NGM Bio, Novo Nordisk, Pfizer, Sonic Incytes, and Terns; and is a cofounder of LipoNexus. Quentin M. Anstee is coordinator of the EU IMI-2 LITMUS consortium, which is funded by the EU Horizon 2020 programme and EFPIA. This multistakeholder consortium includes industry partners. He has also received research grant funding from AstraZeneca, Boehringer Ingelheim, and Intercept Pharmaceuticals, Inc.; has served as a consultant on behalf of Newcastle University for Alimentiv, Akero, AstraZeneca, Axcella, 89bio, Boehringer Ingelheim, Bristol Myers Squibb, Galmed, Genfit, Genentech, Gilead, GSK,

Hanmi, HistolIndex, Intercept Pharmaceuticals, Inc., Inventiva, Ionis, IQVIA, Janssen, Madrigal, Medpace, Merck, NGM Bio, Novartis, Novo Nordisk, PathAI, Pfizer, Poxel, Resolution Therapeutics, Roche, Ridgeline Therapeutics, RTI, Shionogi, and Terns; has served as a speaker for Fishawack, Integritas Communications, Kenes, Novo Nordisk, Madrigal, Medscape, and Springer Healthcare; and receives royalties from Elsevier Ltd. Vlad Ratziu has served as a consultant for Intercept Pharmaceuticals, Inc., Boehringer Ingelheim, Enyo, Madrigal, NGM Bio, Novo Nordisk, Poxel, and Terns. Kris V. Kowdley has received grant/research/clinical trial support from Corcept, CymaBay, Genfit, Gilead, GSK, Hanmi, Intercept Pharmaceuticals, Inc., Madrigal, Mirum, Novo Nordisk, NGM Bio, Pfizer, Pliant, Terns, Viking, and 89bio; has served as a consultant/advisory board member for CymaBay, Enanta, Genfit, Gilead, HighTide, Inipharm, Intercept Pharmaceuticals, Inc., Madrigal, Mirum, NGM Bio, Pfizer, and 89bio; has served on speakers' bureaus for AbbVie, Gilead, and Intercept Pharmaceuticals, Inc.; and has stock options in Inipharm. Stephen A. Harrison has received research grants from Akeru, Axcella, Cirius, CiVi, CymaBay, Enyo, Galectin, Galmed, Genfit, Gilead, Hepion, HighTide, Intercept, Madrigal, Metacrine, NGM Bio, NorthSea, Novartis, Novo Nordisk, Poxel, Sagimet, and Viking; has served as a consultant for AgomAb, Akeru, Alentis, Alimentiv, Altimune, Axcella, Boston Pharmaceuticals, B. Riley FBR, BVF Partners, CohBar, Can-Fite, Corcept, CymaBay, Echosens, Enyo, Fibronostics, Foresite, Fortress, Galectin, Genfit, GNS, Hepion, HighTide, HistolIndex, Inipharm, Intercept Pharmaceuticals, Inc., Ionis, Kowa Research Institute, Madrigal, Medpace, Metacrine, Microba, NGM Bio, NorthSea, Novo Nordisk, Nutrasource, Perspectum, Piper Sandler, Poxel, Prometic, Ridgeline, Sagimet, Sonic Incytes, Terns, and Viking; has served on advisory boards/panels for 89bio, Akeru, Altimune, Arrowhead, Axcella, ChronWell, CiVi, CymaBay, Echosens, Foresite, Galectin, Galmed, Genfit, Gilead, Hepion, HighTide, HistolIndex, Indalo, Intercept Pharmaceuticals, Inc., Madrigal, Medpace, Metacrine, NGM Bio, NorthSea, Novartis, Novo Nordisk, PathAI, Poxel, Prometic, Ridgeline, Sagimet, Sonic Incytes, Terns, and Theratechnologies; and has stock in Akeru, ChronWell, Cirius, Galectin, Genfit, Hepion, HistolIndex, Metacrine, NGM Bio, NorthSea, and Sonic Incytes. Murray B. Resnick is a consultant for PathAI. Thomas Capozza and Sangeeta Sawhney are employees of Intercept Pharmaceuticals, Inc. Nirav Shelat is a former employee of Intercept Pharmaceuticals, Inc. Zobair M. Younossi has received research funding and/or consultant fees from Gilead, Intercept Pharmaceuticals, Inc., Bristol Myers Squibb, Novo Nordisk, AstraZeneca, Siemens, Quest, Madrigal, Merck, and Abbott. The remaining author has no conflicts to report.

ORCID

Arun J. Sanyal <https://orcid.org/0000-0001-8682-5748>

Rohit Loomba <https://orcid.org/0000-0002-4845-9991>

Quentin M. Anstee <https://orcid.org/0000-0002-9518-0088>

Vlad Ratziu <https://orcid.org/0000-0002-3051-0111>

Kris V. Kowdley <https://orcid.org/0000-0002-8553-3652>

Stephen A. Harrison <https://orcid.org/0000-0001-6399-5744>

Murray B. Resnick <https://orcid.org/0000-0002-9392-3415>

Zobair M. Younossi <https://orcid.org/0000-0001-9313-577X>

REFERENCES

1. European Medicinal Agency Committee for Medicinal Products for Human Use (CHMP). Reflection paper on regulatory requirements for the development of medicinal products for chronic non-infectious liver diseases (PBC, PSC, NASH). 2021. Accessed September 13, 2021. https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-regulatory-requirements-development-medicinal-products-chronic-non-infectious-liver_en.pdf.
2. US Food and Drug Administration. Noncirrhotic nonalcoholic steatohepatitis with liver fibrosis: Developing drugs for treatment. Guidance for industry. 2018. Accessed September 13, 2021. <https://www.fda.gov/media/119044/download>
3. Loomba R, Ratziu V, Harrison SA, NASH Clinical Trial Design International Working Group. Expert panel review to compare FDA and EMA guidance on drug development and endpoints in nonalcoholic steatohepatitis. In: NASH Clinical Trial Design International Working Group, eds. *Gastroenterology*. 2022;162:680–8.
4. Pai RK, Jairath V, Hogan M, Zou G, Adeyi OA, Anstee QM, et al. Reliability of histologic assessment for NAFLD and development of an expanded NAFLD activity score. *Hepatology*. 2022;76:1150–63.
5. Brunt EM, Clouston AD, Goodman Z, Guy C, Kleiner DE, Lackner C, et al. Complexity of ballooned hepatocyte feature recognition: Defining a training atlas for artificial intelligence-based imaging in NAFLD. *J Hepatol*. 2022;76:1030–41.
6. Anania FA, Dimick-Santos L, Mehta R, Toerner J, Beitz J. Nonalcoholic steatohepatitis: Current tinkering from the Division of Hepatology and Nutrition at the Food and Drug Administration. *Hepatology*. 2021;73:2023–7.
7. Davison BA, Harrison SA, Cotter G, Alkhouri N, Sanyal A, Edwards C, et al. Suboptimal reliability of liver biopsy evaluation has implications for randomized clinical trials. *J Hepatol*. 2020;73:1322–32.
8. Ratziu V, Charlotte F, Heurtier A, Gombert S, Giral P, Bruckert E, et al. Sampling variability of liver biopsy in nonalcoholic fatty liver disease. *Gastroenterology*. 2005;128:1898–906.
9. Vuppalanchi R, Unalp A, Van Natta ML, Cummings OW, Sandrasegaran KE, Hameed T, et al. Effects of liver biopsy sample length and number of readings on sampling variability in nonalcoholic fatty liver disease. *Clin Gastroenterol Hepatol*. 2009;7:481–6.
10. Kleiner DE, Brunt EM, Van Natta M, Behling C, Contos MJ, Cummings OW, et al. Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology*. 2005;41:1313–21.
11. Kleiner DE, Brunt EM, Wilson LA, Behling C, Guy C, Contos M, et al. Association of histologic disease activity with progression of nonalcoholic fatty liver disease. *JAMA Netw Open*. 2019;2:e1912565.

12. Sanyal AJ, Harrison SA, Ratziu V, Abdelmalek MF, Diehl AM, Caldwell S, et al. The natural history of advanced fibrosis due to nonalcoholic steatohepatitis: Data from the simtuzumab trials. *Hepatology*. 2019;70:1913–27.
13. Angulo P, Kleiner DE, Dam-Larsen S, Adams LA, Bjornsson ES, Charatcharoenwithaya P, et al. Liver fibrosis, but no other histologic features, is associated with long-term outcomes of patients with nonalcoholic fatty liver disease. *Gastroenterology*. 2015;149:389–97 e310.
14. Younossi ZM, Stepanova M, Rafiq N, Henry L, Loomba R, Makhlof H, et al. Nonalcoholic steatofibrosis independently predicts mortality in nonalcoholic fatty liver disease. *Hepatol Commun*. 2017;1:421–8.
15. Matsubayashi T. Drug development for nonalcoholic steatohepatitis (NASH) with fibrosis: A regulatory perspective [presentation]. Regulatory perspectives for development of drugs for treatment of NASH [webinar]. Updated January 29, 2021. Accessed November 8, 2022. <https://www.fda.gov/drugs/news-events-human-drugs/regulatory-perspectives-development-drugs-treatment-nash-01292021-01292021>.
16. Younossi ZM, Ratziu V, Loomba R, Rinella M, Anstee QM, Goodman Z, et al. Obeticholic acid for the treatment of non-alcoholic steatohepatitis: interim analysis from a multicentre, randomised, placebo-controlled phase 3 trial. *Lancet*. 2019;394:2184–96.
17. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull*. 1979;86:420–8.
18. Cicchetti DV, Allison T. A new procedure for assessing reliability of EEG sleep recordings. *Am J EEG Technol*. 1971;11:101–10.
19. Rowe IA, Parker R. The placebo response in randomized trials in nonalcoholic steatohepatitis simply explained. *Clin Gastroenterol Hepatol*. 2022;20:e564–72.
20. Bedossa P, Dargere D, Paradis V. Sampling variability of liver fibrosis in chronic hepatitis C. *Hepatology*. 2003;38:1449–57.
21. Newsome PN, Buchholtz K, Cusi K, Linder M, Okanou T, Ratziu V, et al. A placebo-controlled trial of subcutaneous semaglutide in nonalcoholic steatohepatitis. *N Engl J Med*. 2021;384:1113–24.
22. Speight PM, Abram TJ, Floriano PN, James R, Vick J, Thornhill MH, et al. Interobserver agreement in dysplasia grading: Toward an enhanced gold standard for clinical pathology trials. *Oral Surg Oral Med Oral Pathol Oral Radiol*. 2015;120:474–482.e472.
23. Longerich T, Schirmacher P. Determining the reliability of liver biopsies in NASH clinical studies. *Nat Rev Gastroenterol Hepatol*. 2020;17:653–4.
24. Saco A, Ramirez J, Rakislova N, Mira A, Ordi J. Validation of whole-slide imaging for histopathological diagnosis: Current state. *Pathobiology*. 2016;83:89–98.
25. Saco A, Diaz A, Hernandez M, Martinez D, Montironi C, Castillo P, et al. Validation of whole-slide imaging in the primary diagnosis of liver biopsies in a University Hospital. *Dig Liver Dis*. 2017;49:1240–6.
26. Naoumov NV, Brees D, Loeffler J, Chng E, Ren Y, Lopez P, et al. Digital pathology with artificial intelligence analyses provides greater insights into treatment-induced fibrosis regression in NASH. *J Hepatol*. 2022;77:1399–409.
27. Sun Y, Zhou J, Wang L, Wu X, Chen Y, Piao H, et al. New classification of liver biopsy assessment for fibrosis in chronic hepatitis B patients before and after treatment. *Hepatology*. 2017;65:1438–50.
28. Chowdhury AB, Mehta KJ. Liver biopsy for assessment of chronic liver diseases: A synopsis. *Clin Exp Med*. 2022;23:273–85.
29. Pavlides M, Birks J, Fryer E, Delaney D, Sarania N, Banerjee R, et al. Interobserver variability in histologic evaluation of liver fibrosis using categorical and quantitative scores. *Am J Clin Pathol*. 2017;147:364–9.
30. Rinella ME, Dufour JF, Anstee QM, Goodman Z, Younossi Z, Harrison SA, et al. Non-invasive evaluation of response to obeticholic acid in patients with NASH: Results from the REGENERATE study. *J Hepatol*. 2022;76:536–48.

How to cite this article: Sanyal AJ, Loomba R, Anstee QM, Ratziu V, Kowdley KV, Rinella ME, et al. Utility of pathologist panels for achieving consensus in NASH histologic scoring in clinical trials: Data from a phase 3 study. *Hepatol Commun*. 2024;8:e0325. <https://doi.org/10.1097/HC9.0000000000000325>