Title

Reliability of histologic assessment for NAFLD and development of an expanded NAFLD activity score

Journal

Hepatology, 76(4)

ISSN

0270-9139

Authors

Pai, Rish K

Jairath, Vipul

Hogan, Malcolm

et al.

Publication Date

2022-10-01

DOI

10.1002/hep.32475

Peer reviewed

HHS Public Access

Author manuscript

Hepatology. Author manuscript; available in PMC 2023 October 01.

Published in final edited form as:

Hepatology. 2022 October ; 76(4): 1150–1163. doi:10.1002/hep.32475.

Reliability of Histologic Assessment for NAFLD and Development of an Expanded NAFLD Activity Score

Rish K. Pai[1], **Vipul Jairath**[2], **Malcolm Hogan**[3], **Guangyong Zou**[4], **Oyedele A. Adeyi**[5], **Quentin M. Anstee**[6], **Bashar A. Aqel**[7], **Cynthia Behling**[8], **Elizabeth J. Carey**[9], **Andrew D. Clouston**[10], **Kathleen Corey**[11], **Brian G. Feagan**[12], **David E. Kleiner**[13], **Christopher Ma**[14], **Stefanie C. McFarlane**[15], **Mazen Noureddin**[16], **Vlad Ratziu**[17], **Mark A. Valasek**[18], **Zobair M. Younossi**[19], **Stephen A. Harrison**[20], **Rohit Loomba**[21]

[1]Department of Laboratory Medicine & Pathology, Mayo Clinic Arizona, Scottsdale, AZ, USA

[2]Department of Medicine, Division of Gastroenterology, University of Western Ontario, London, ON, Canada; Department of Epidemiology and Biostatistics, University of Western Ontario, London, ON, Canada; Alimentiv Inc., London, ON, Canada

[3]Alimentiv Inc., London, ON, Canada

[4]Department of Epidemiology and Biostatistics, University of Western Ontario, London, ON, Canada; Robarts Research Institute, Schulich School of Medicine and Dentistry, University of Western Ontario, London, ON, Canada; Alimentiv Inc., London, ON, Canada

[5]Department of Laboratory Medicine and Pathology, University of Minnesota Medical School, Minneapolis, MN, USA

[6]Translational & Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, UK; NIHR Newcastle Biomedical Research Center, Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK

[7]Division of Gastroenterology and Hepatology, Mayo Clinic Arizona, Phoenix, AZ, USA

[8]Pacific Rim Pathology, San Diego, CA, USA; Department of Pediatrics, University of California San Diego, La Jolla, CA, USA

[9]Division of Gastroenterology and Hepatology, Mayo Clinic Arizona, Phoenix, AZ, USA

CO-CORRESPONDING AUTHORS: Rish K. Pai, MD, PhD, Department of Laboratory Medicine and Pathology, Mayo Clinic Arizona, 13400 E. Shea Blvd., Scottsdale, AZ 85259, Tel: 480-301-4081, Fax: 480-301-9158, pai.rish@mayo.edu, Rohit Loomba, MD, MHSc, Director, NAFLD Research Center, Professor of Medicine, Director of Hepatology and Vice Chief, Division of Gastroenterology, Adjunct Professor, Division of Epidemiology, 1W202 ACTRI Building # MC0887, 9452 Medical Center Drive, University of California at San Diego, La Jolla, CA, 92093-0887, Tel: 858-246-2201, Fax: 858-246-2255, roloomba@health.ucsd.edu.

AUTHOR CONTRIBUTIONS:

All authors have made substantial contributions and reviewed and approved the final manuscript.
Article guarantor: RKP, RL
Study conception and design: RKP, VJ, RL
Investigator: RKP, OAA, EJC, ADC, VR, ZY, RL
Data acquisition: RKP, OAA, ADC, MAV
Data analysis: MH, GZ
Data interpretation: all authors
Study supervision: RKP, VJ, RL
Manuscript drafting: RKP, VJ, SM, RL
Manuscript editing for important intellectual content: all authors

[10]Faculty of Medicine and Biomedical Sciences, University of Queensland, Brisbane, Queensland, Australia

[11]Division of Gastroenterology, Massachusetts General Hospital, Boston, MA, USA; Harvard Medical School, Boston, MA, USA

[12]Department of Medicine, Division of Gastroenterology, University of Western Ontario, London, ON, Canada; Department of Epidemiology and Biostatistics, University of Western Ontario, London, ON, Canada; Alimentiv Inc., London, ON, Canada

[13]Laboratory of Pathology, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA

[14]Division of Gastroenterology & Hepatology, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada; Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada; Alimentiv Inc., London, ON, Canada

[15]Alimentiv Inc., London, ON, Canada

[16]Division of Gastroenterology and Hepatology, Department of Medicine, Cedars-Sinai Medical Center, Los Angeles, CA, USA

[17]Sorbonne Université, Institute of Cardiometabolism and Nutrition, Pitié-Salpêtrière Hospital, Paris, France

[18]Department of Pathology, University of California at San Diego, La Jolla, CA, USA

[19]Betty and Guy Beatty Center for Integrated Research, Inova Health System, Falls Church, VA, USA; Department of Medicine, Center for Liver Diseases, Inova Fairfax Medical Campus, Falls Church, VA, USA

[20]Radcliffe Department of Medicine, University of Oxford, Oxford, UK; Medical Director, Pinnacle Clinical Research, San Antonio, TX, USA

[21]NAFLD Research Center, Division of Gastroenterology and Hepatology, Department of Medicine, University of California at San Diego, La Jolla, CA, USA

Abstract

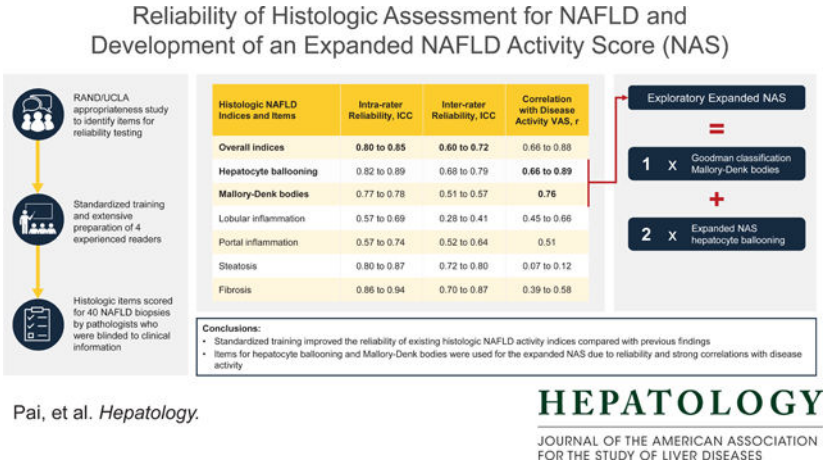
Background & Aims—The Nonalcoholic Steatohepatitis Clinical Research Network (NASH CRN) histologic scoring system, the gold standard NASH histology assessment for clinical trials, has demonstrated intrarater and interrater variability. An expert panel in a previous systematic Research and Development/University of California Los Angeles (RAND/UCLA) study determined that existing histologic scoring systems do not fully capture NASH disease activity and fibrosis and standardized definitions of histologic features are needed. We evaluated the reliability of existing and alternate histologic measures and their correlations with a disease activity visual analog scale (VAS) to propose optimal components for a new expanded NAFLD activity score (NAS).

Approach & Results—Four liver pathologists who were involved in the prior RAND/UCLA study underwent standardized training and multiple discussions with the goal of improving agreement. They were blinded to clinical information and scored histologic measures twice, ≥2

weeks apart, for 40 liver biopsies representing the full spectrum of NAFLD. Index intraclass correlation coefficient (ICC) estimates demonstrated intrarater (0.80–0.85) and interrater reliability (0.60–0.72). Hepatocyte ballooning items had similar interrater ICCs (0.68–0.79), including those extending scores from 0–2 to 0–4. Steatosis measures (interrater ICCs, 0.72–0.80) correlated poorly with disease activity. Correlations with disease activity were largest for hepatocyte ballooning and Mallory-Denk bodies (MDBs), with both used to develop the expanded NAS (intrarater ICC, 0.90; interrater ICC, 0.80). Fibrosis measures had ICCs of 0.70 to 0.87.

Conclusions—After extensive preparation among a group of experienced pathologists, we demonstrated improved reliability of multiple existing histologic NAFLD indices and fibrosis staging systems. Hepatocyte ballooning and MDBs most strongly correlated with disease activity and were used for the new expanded NAS. Further validation including evaluation of responsiveness is required.

Graphical Abstract



Keywords

Nonalcoholic fatty liver disease; NASH; steatohepatitis; intraclass correlation coefficients; hepatocyte ballooning

Nonalcoholic fatty liver disease is characterized by the presence of hepatic steatosis, defined by imaging or histology, in individuals who consume little or no alcohol and do not have secondary causes of hepatic fat accumulation.[1] Encompassing a spectrum of liver diseases, NAFLD ranges from nonalcoholic fatty liver (NAFL), where steatosis is present without hepatocyte injury, to NASH, where steatosis is typically accompanied by hepatocellular injury and lobular inflammation with or without fibrosis.[2,3] Given that NASH is more likely than NAFL to progress to advanced fibrosis, cirrhosis, and liver-related mortality, targeting NASH with significant fibrosis (NASH stage ≥2 fibrosis) has become a focus of drug development.[4] Although many pharmacologic agents are under evaluation in clinical trials for the treatment of NASH,[5–8] none are currently approved by the U.S. Food and Drug Administration (FDA) or the European Medicines Agency.[3] In the United States, NASH-related cirrhosis is the leading cause of liver transplantation in women and projected leading cause in men.[9]

The FDA considers histologic evaluation essential for defining patient eligibility criteria and primary efficacy endpoints for phase 2b and 3 clinical trials in NASH.[10] Of the histologic measures available,[11–16] the NAFLD activity score (NAS) developed by the NASH Clinical Research Network (NASH CRN) is the most commonly used scale.[5] When the NAS was developed, understanding of the histologic features that drive disease progression was limited. Subsequently, two NAS items, histologic steatosis and lobular inflammation, were shown not to be associated with progression to cirrhosis and liver-related outcomes.[14,17–20] Furthermore, lobular inflammation and steatosis have a wider dynamic range (0–3) in the NAS relative to hepatocyte ballooning (0–2), an item strongly linked to fibrosis progression.[17,19–21] Additionally, the NAS excludes two other items of potential value, namely Mallory-Denk bodies (MDBs)[19,20] and portal inflammation,[22–24] which have been associated with NAFLD progression. Although portal inflammation has also been shown to increase after treatment with some otherwise effective therapies, improvements in this feature have been correlated with improvements in fibrosis in most studies.[19,22,25] Finally, the NASH CRN staging system may not capture the full spectrum of the dynamic range of fibrosis in NAFLD. Specifically, it has 3 levels for stage 1 fibrosis (1a, 1b, and 1c) but does not evaluate the considerable morphologic diversity seen in bridging fibrosis. For example, this system regards a single delicate fibrous bridge and complex bridging with numerous fibrous bridges as stage 3 disease.

In a prior modified Research and Development/University of California, Los Angeles (RAND/UCLA) appropriateness process on standardizing interpretation of NASH biopsies for clinical trials, evaluation of alternate and improved methods to measure fibrosis, hepatocyte ballooning, MDBs, and portal inflammation were regarded by the panelists as appropriate.[26] The panelists did not regard the NAS as an optimal index for measuring NAFLD disease severity in clinical trials and rated this statement as uncertain given the emphasis on steatosis and lobular inflammation, limited range of ballooning scores, and exclusion of portal inflammation and MDBs.

Efficient conduct of clinical trials requires standardized use of validated, reliable, and responsive evaluative instruments. In the absence of a comprehensive evaluation of the reliability of existing NAFLD histologic indices, we evaluated the intrarater and interrater reliability of these indices, their component items, and additional histologic items identified through the aforementioned RAND/UCLA appropriateness process.[26] In addition, we aimed to determine if the intrarater and interrater reliability of existing histologic indices can be improved with standardized training. Given that the NAS was not rated by panelists as an optimal index, we also aimed to develop an exploratory expanded NAS index using systematic correlations between reliable histologic items and a disease activity visual analog scale (VAS).

MATERIALS AND METHODS

Study Design

This study included 4 experienced liver pathologists (ADC, MAV, OAA, and RKP) with expertise in NAFLD and more than 10 years as liver pathology subspecialists, who initially participated in the RAND/UCLA process. They were subsequently invited to serve as central

readers and to retrospectively review 40 liver biopsy slides from patients with NAFLD. In addition to extensive discussions during the RAND/UCLA process, the pathologists completed standardized training on scoring all histologic measures and met as a group to review numerous illustrations and discuss potential issues, with the goal of improving agreement in this study. For each slide, the central readers scored its quality (Supplementary Table 1) and completed 2 assessments of all items, with ≥2 weeks between each assessment to facilitate memory extinction, for a total of 320 observations. Pathologists were blinded to clinical information.

Study Population

We used 40 liver biopsies that had been obtained from patients with histologically confirmed NAFLD as per their routine care at the Mayo Clinic (Phoenix, AZ). The slides were selected to represent the full histologic spectrum of disease, with optimal biopsy quality, ranging from NAFL to NASH and included stages 0, 1, and 2, bridging fibrosis, and cirrhosis. The Masson's trichrome– and hematoxylin and eosin–stained slides were scanned at 40× magnification using an Aperio® AT2 (Leica Biosystems; Buffalo Grove, Illinois, USA) and compressed using WebMicroscope® Compressor (Fimmic, Helsinki, Finland). Central readers viewed the slides on the Alimentiv WebMicroscope database hosted on a secure remote server.

Index and Item Selection and RAND Process

All NAFLD indices and items were selected for this study based on previous ratings in a modified RAND/UCLA appropriateness process.[26] In this process, 14 liver pathologists and 3 hepatologists completed 2 rounds of voting using an online survey of 130 statements identified through literature review and expert opinion. Survey responses from each round and any disagreement among the panelists were discussed, and statements classified as *appropriate* or *uncertain* based on the median panel rating and degree of panel disagreement were included for reliability testing. The reliability of existing NAFLD histologic indices (Supplementary Table 2), including the Brunt criteria,[11] modified Brunt criteria,[12] NAS,[13] Goodman classification,[14] steatosis, activity fibrosis (SAF) scoring system,[15] and Ishak staging of fibrosis (modification of the Knodell staging system),[16] and the component items of these indices was evaluated.[26] Additionally, histologic items identified during the previous modified RAND/UCLA appropriateness process (Supplementary Table 3)[26] and a 100-mm VAS (0, *completely normal*; 100, *worst disease ever seen*) for overall NAFLD histologic disease activity and for each of 4 individual features (hepatocyte ballooning, lobular inflammation, steatosis, and fibrosis) were included for reliability assessment. Each VAS for a given feature was scored as the readers' overall impression of severity immediately prior to the remaining items for that feature.

Histologic Features

Reliability was assessed for the indices overall, the individual items for hepatocyte ballooning and MDBs, lobular inflammation, portal inflammation, steatosis, and fibrosis, and for each VAS. Of all indices, only the Goodman classification has no calculation of an overall score. For each index, fibrosis was assessed separately from the overall histologic activity grade or score.

Hepatocyte Ballooning and Mallory-Denk Bodies

Given the RAND/UCLA-identified importance of improving the measurement of hepatocyte ballooning in NASH, 7 hepatocyte ballooning items were assessed for reliability, including items from the NAS (0–2), SAF scoring system (0–2), and Goodman classification (0–3), a previously published expanded NAS ballooning item (0–4),[26,27] a ballooning item based on percent involvement of 20× fields (0–3), an alternate expanded ballooning item based on clusters of ballooned cells (0–4), and a ballooning VAS. The presence of MDBs is captured in the Goodman classification (0–3) and in a RAND/UCLA-identified alternate item (0–2).

Lobular Inflammation

Lobular inflammation assessments included multiple definitions from the NAS, SAF scoring system, modified Brunt criteria, Goodman classification, RAND/UCLA-identified alternate items, and a lobular inflammation VAS. The item from the Goodman classification uses the descriptors *none*, *mild*, *moderate*, and *severe*, whereas the other items assess the number of foci with lobular inflammation per 20× field. *Focus* is defined in the NAS lobular inflammation item as a cluster of inflammatory cells the size of a hepatocyte and in the SAF scoring system as ≥2 inflammatory cells. The reliability of defining a *focus* as a cluster of ≥5 inflammatory cells was also explored using 2 alternate items—one that assessed the *number* of foci per 20× field (alternate lobular inflammation #1) and the other that assessed *percentage* of 20× fields with a focus (alternate lobular inflammation #2). The reliability of intra-acinar neutrophils and acidophil bodies were assessed separately from other lobular inflammation measures.

Portal Inflammation

Portal tracts and fibrous septa can become inflamed in NASH and are not captured in the NAS or SAF scoring system. The portal inflammation item from the Goodman classification (*none*, *mild*, *moderate*, *severe*) was assessed based on the average inflammation in all portal tracts and the most involved portal tract.

Steatosis

The reliability of macrovesicular steatosis items from the NAS (0–3), Goodman classification (0–4), and SAF scoring system (0–3) and a steatosis VAS was assessed. Reliability was also determined for additional RAND/UCLA-identified measures of steatosis, including an alternate item that estimates percentage of nonfibrotic parenchyma replaced by steatosis evaluated at low to medium power (4×–10×), the size of steatotic droplets (small/mixed/large), and microvesicular steatosis.

Fibrosis

Reliability was assessed for fibrosis items from the NAS, Goodman classification, Brunt and modified Brunt criteria, SAF staging system, Ishak fibrosis staging system (modified slightly to account for fibrosis in the central areas for stages 1 and 2), 2 RAND/UCLA-proposed alternative fibrosis staging systems, and a fibrosis VAS. The first alternate fibrosis staging system uses the NAS fibrosis item but expands bridging fibrosis into 3 levels (stage 3a: one definite fibrous bridge connecting 2 structures; stage 3b: few fibrous bridges without nodule

formation; stage 3c: complex bridging with or without rare nodule formation). The second alternate fibrosis staging system separately measures zone 3/sinusoidal fibrosis (0–4) and portal fibrosis (A-D).

Exploratory Analysis

An exploratory objective was to determine if highly reliable items could be used to derive an optimized index. Five candidate hepatocyte ballooning, MDB, lobular inflammation, portal inflammation, and steatosis items were selected for preliminary development of this index, with the disease activity VAS used as an anchor. Items retained after a stepwise regression procedure were included in the final exploratory optimized index.

Statistical Analysis

Descriptive statistics were used to summarize demographics and clinical data. The intrarater and interrater reliability of each index and item was determined using intraclass correlation coefficients (ICCs), estimated using a 2-way random effects analysis of variance model with interaction.[28] Point estimates of ICCs are equivalent to Kappa.[29] The associated 2-sided 95% CIs were obtained using the nonparametric cluster bootstrap method with 2000 replicates to address the potential for data nonnormality. Correlation coefficients between each measure and VAS (disease activity, fibrosis) were estimated using a mixed effect model approach with cluster bootstrap methods used to obtain 2-sided 95% CIs[28,30–32] and were interpreted according to the benchmarks established by Cohen, where 0.1, 0.3, and 0.5 indicate small, medium, and large effects, respectively.[33]

For the exploratory index, 5 candidate items with interrater ICCs of 0.41 or higher were selected according to bivariate relationships with the disease activity VAS. To avoid violating the assumption of independent observations, a single observation was randomly chosen for each subject. Spearman correlation coefficients and associated 95% CIs were determined between the candidate items and several clinical characteristics. A full preliminary model incorporating all 5 items was obtained, and items were then removed in a step-down model-building approach where $P = 0.05$ for the individual predictors was used as the criterion for item retention. Remaining items were used in the final exploratory model, with coefficients standardized using division by the smallest coefficient, and were examined for obvious violations of key assumptions for linear regression analysis, including homogeneity of residual variance, normality of residual distribution, and the lack of outliers. The ICCs of the final model were used to evaluate its intrarater and interrater reliability in an exploratory analysis.

All statistical calculations were performed using SAS version 9.4 (SAS Institute Inc.; Cary, North Carolina, USA).

Sample Size

The study sample size was based upon estimating the ICC using a one-way random effects model.[34] Assuming a true ICC of 0.7, evaluation of 40 histologic slides by 4 central readers would yield an approximate 80% chance of obtaining a 2-sided 95% lower bound for the ICC that is greater than 0.5, which is higher than the midpoint of the Landis and

Koch benchmark for moderate reliability (0.41–0.60).[35] The sample size was conservative to account for duplicate reading of the 40 slides by 4 central readers. This study was primarily powered to evaluate the reliability of existing NAFLD histologic indices and items, additional RAND/UCLA-identified items, and the disease activity VAS but not the optimized histologic index comprising selected items. Therefore, the development of this optimized index and related reliability assessments should be considered exploratory.

Ethical Considerations

This study was conducted in accordance with the protocol, which was written and codified before proceeding, and the ethical guidelines of the 1975 Declaration of Helsinki. The Mayo Clinic institutional review board approved the protocol, including the use of the deidentified, digitized histology slides of liver biopsies, and granted a waiver of the requirement to obtain written informed consent.

RESULTS

Patient Characteristics

The clinical characteristics associated with the histologic slides in this study are consistent with those of fatty liver disease (Table 1). Of the 40 patients with NAFLD (mean age, 55.3 years; mean BMI, 35.7 kg/m^2), 15 (37.5%) were men and 17 (42.5%) had diabetes. Mean liver enzyme values were 99.8 U/L for alanine aminotransferase (ALT), 76.1 U/L for aspartate aminotransferase (AST), and 91.2 U/L for alkaline phosphatase.

Reliability Study

Intrarater and interrater ICCs for existing NAFLD indices and items are shown in Table 2, and ICCs for the RAND/UCLA-identified items are shown in Table 3.

Intrarater Reliability—Existing disease activity indices had intrarater ICCs that ranged from 0.80 to 0.85. The hepatocyte ballooning items had similar intrarater ICCs, regardless of the scale used (0 to 2 [ICC, 0.85–0.86]; 0 to 3 [ICC, 0.87–0.88]; 0 to 4 [ICC, 0.89]). The MDB items had intrarater ICCs of 0.77 to 0.78. Intrarater ICCs for lobular inflammation ranged from 0.57 for the alternate lobular inflammation #1 item to 0.69 for the alternate lobular inflammation #2 item. Intrarater ICCs were 0.54 for intra-acinar neutrophils and 0.47 for apoptotic hepatocytes. Goodman classification portal inflammation items had an intrarater ICC of 0.74 for the alternate item that is based on the most involved portal tract and 0.57 for the item that is based on the average inflammation in all portal tracts. Intrarater ICCs for steatosis ranged from 0.80 to 0.87. Intrarater reliability was lower for items measuring size of steatotic droplets (ICC, 0.44) and microvesicular steatosis (ICC, 0.53–0.63). The intrarater ICC of the disease activity VAS was 0.86.

Fibrosis measures had intrarater ICCs of 0.86 for the pericellular/perisinusoidal fibrosis item from the Goodman classification to 0.94 for the fibrosis VAS.

Interrater reliability—Interrater ICCs for existing disease activity indices ranged from 0.60 to 0.72. Hepatocyte ballooning items had interrater ICCs of 0.68 to 0.79, and expanding

the range of scores from 0 to 2 (ICC, 0.68–0.73) to 0 to 4 (ICC, 0.77–0.79) did not decrease reliability. The MDB items had interrater ICCs of 0.51 to 0.57. Items for measuring lobular inflammation had ICCs from 0.28 for the SAF lobular inflammation item to 0.41 for the alternate lobular inflammation #2 item. Intra-acinar neutrophils had an interrater ICC of 0.48, and apoptotic hepatocytes had an ICC of 0.42. Interrater reliability of Goodman classification portal inflammation items was better when based on the most involved portal tract (ICC, 0.64) than when based on the average inflammation in all portal tracts (ICC, 0.52). Most steatosis measures demonstrated interrater ICCs of 0.72 to 0.80. Size of steatotic droplets and the 2 microvesicular steatosis items demonstrated lower interrater agreement (ICC, 0.05–0.07), consistent with the challenges of scoring these measures. The disease activity VAS had an interrater ICC of 0.69.

Measures of fibrosis had ICCs of 0.70 to 0.87, with higher interrater reliability observed for the Ishak fibrosis staging system (ICC, 0.87) and the alternate NAS fibrosis staging system that expands bridging fibrosis (ICC, 0.83). The interrater ICC of the fibrosis VAS was 0.86.

Correlations Between Histologic Indices/Items and the Disease Activity/Fibrosis VAS

Correlations with the disease activity VAS and fibrosis VAS are shown in Table 2 for existing NAFLD indices and items and in Table 3 for the RAND/UCLA-identified items.

For the overall disease activity indices, correlations with the disease activity VAS ranged from r = 0.66 (95% CI: 0.52–0.71) for the SAF activity grade to r = 0.88 (95% CI: 0.85–0.91) for the Brunt criteria necro-inflammatory grade of steatohepatitis. The largest correlations with the disease activity VAS were observed for the hepatocyte ballooning items (r = 0.88; 95% CI: 0.84–0.90 for the expanded NAS ballooning score) followed by MDBs (r = 0.76; 95% CI: 0.59–0.82 for the Goodman classification MDB item), lobular inflammation (r = 0.63; 95% CI: 0.49–0.67 for the alternate lobular inflammation #2 item), and portal inflammation (r = 0.51; 95% CI: 0.23–0.55 for the item that is based on the most severely inflamed portal tract). Correlations with the disease activity VAS were small for all steatosis items (r = −0.02 to 0.12).

Correlations between fibrosis items and the fibrosis VAS ranged from r = 0.77 (95% CI: 0.70–0.88) for the cirrhosis item from the Goodman classification to r = 0.96 (95% CI: 0.95–0.97) for the Ishak fibrosis staging system. The fibrosis correlations were largest for the Ishak fibrosis staging system and the alternate NAS fibrosis staging system that expands bridging fibrosis (r = 0.95; 95% CI: 0.94–0.96), indicating that these 2 staging systems are accurate for measuring the degree of fibrosis within the liver.

Exploratory Development of an Optimized Histologic NAFLD Index: Expanded NAS Index

Of the 5 candidate items with interrater ICCs ≥0.41 selected for a preliminary NAFLD index, the NAS steatosis item and the Goodman classification MDB item were from existing NAFLD indices. The remaining 3 items had been previously identified through the modified RAND/UCLA process: the expanded NAS hepatocyte ballooning item, the alternate lobular inflammation #2 item, and the Goodman classification portal inflammation item that is based on the most involved portal tract. Clinical correlations were strongest between the lobular inflammation item and both AST (r_S = 0.49; 95% CI: 0.21–0.69) and ALT (r_S = 0.43;

95% CI: 0.14–0.65) and between the hepatocyte ballooning item and AST (r_s = 0.41; 95% CI: 0.11–0.64) (Supplementary Table 4). Relationships between each candidate item and the disease activity VAS are shown in Fig. 1A, and linear regression models for the candidate items are shown in Supplementary Table 5.

After a step-down procedure was performed, 3 items were removed including the NAS steatosis item, the alternate lobular inflammation #2 item, and the Goodman classification portal inflammation item based upon the most involved portal tract. The remaining 2 items —the Goodman classification MDB item and the expanded NAS hepatocyte ballooning item—comprised the final exploratory model for predicting the disease activity VAS. Using standardized coefficients (Table 4), an exploratory expanded NAS index was generated that can be scored from 0 to 11, with higher scores indicating more severe disease activity, and calculated as:

Expanded NAS = 1 × Goodman classification MDBs (4 levels) + 2 × expanded NAS hepatocyte ballooning (5 levels).

The R^2 was similar between the full preliminary model comprising the 5 candidate items (R^2 = 0.87) and the final 2-item model (R^2 = 0.86). Diagnostic plots for the final model did not indicate severe violation of model assumptions (eg, homogeneity of residual variance, normality of residual distribution, lack of outliers) (Supplementary Fig. 1).

For the exploratory expanded NAS, the mean ± SD score among the 40 slides in this study was 6.00 ± 3.20. This index demonstrated an intrarater ICC of 0.90 (95% CI: 0.84–0.93). The interrater ICC of the expanded NAS was 0.80 (95% CI: 0.70–0.86) and numerically higher than that of all other indices (Table 2).

Exploratory Analysis of Fibrosis Staging Systems

The Ishak fibrosis staging system and the expanded NASH CRN fibrosis staging system that expands bridging fibrosis were selected based on their relationships with the fibrosis VAS (Fig. 1B). There is a large gap in the fibrosis VAS observed between stages 2 and 3 in the original NASH CRN staging system. In contrast, the increase in fibrosis VAS between each stage of fibrosis in both the Ishak fibrosis staging system and the expanded NASH CRN staging system is more uniform.

DISCUSSION

Evaluation of pharmacologic therapies for NASH is increasing in clinical trials,[5–8] with histologic assessment being considered essential by the FDA for defining entry criteria and evaluating efficacy.[10] However, histologic assessments are often inconsistent across trials for several reasons including variations in histologic definitions and scoring systems. Minimizing heterogeneity in the patient population and standardizing histologic outcomes are critical considerations for the conduct of efficient drug trials in NASH. To achieve standardization in clinical trials, fully validated and objective histologic measures of disease activity should be used, requiring formal evaluations of reliability and responsiveness. It is notable that these requirements have not been completed for currently available

histologic measures.[11–16] Operationalizing histopathology into clinical trials is challenging. Methodologies for integrating histopathology slide reading in clinical trials include use of a single central reader, simultaneous review by multiple readers, individual review by multiple readers with consensus or adjudication paradigms, duplicate reviews of screening and end-of-study biopsies, use of digital images/glass slides, and central/local slide processing.[26] Each approach may impact scoring reliability. The FDA encourages the development of an imaging charter for use in a clinical trial to standardize specific imaging processes including image acquisition, quality, and interpretation, reader training, and controlling for bias and variability.[36] To this end, all readers involved in this study participated in a training session where scoring issues were discussed and a detailed scoring document with numerous illustrations was provided. The critical importance of central reader training and attainment of high interrater reliability has been underscored by Davison *et al.* who used a large clinical trial dataset of paired liver biopsies to demonstrate suboptimal reliability of biopsy evaluations that would have impacted histologic eligibility and outcomes during the trial.[37] Approximately half of the patients with NASH who were included in the EMMINENCE study based on qualifying readings by 1 hepatopathologist were subsequently not considered to fulfill the histologic inclusion criteria according to ≥1 of the 3 hepatopathologists.[37] This lack of interrater reliability might also be expected to diminish statistical power to detect potential treatment effects. In this instance, simulations have shown that low interrater reliability could reduce study power from >90% to as low as 40%,[37] and many studies do not currently incorporate known variability into the power calculation.

In this study, we evaluated the intrarater and interrater reliability of several measures of histologic NAFLD activity. Our key findings demonstrate intrarater and interrater reliability for existing histologic NAFLD activity indices that was improved with standardized training compared with previous findings.[37] Among the index component items and additional RAND/UCLA-identified items, hepatocyte ballooning and fibrosis items were most reliable. The disease activity VAS allowed us to determine the individual features that correlate most strongly with a pathologist's global impression of disease activity, a well-accepted methodology used in multiple index development studies.[38,39] Among all histologic features, measures of hepatocyte ballooning and MDBs had the strongest correlations with disease activity. Moderate correlations with disease activity were observed for portal inflammation and lobular inflammation measures. However, negligible correlations with disease activity were seen for all steatosis items, an important finding given that steatosis is among the items assessed in most existing histologic indices of NAFLD activity including the NAS.[13] Steatosis is a key feature of the disease but does not always correlate with inflammation and hepatocellular injury. Our findings are consistent with prior studies demonstrating the critical importance of hepatocyte ballooning in disease activity, as demonstrated by the strong association between this feature and the development of adverse liver-related outcomes and fibrosis progression.[14,17,20,21,40] In contrast, measures of steatosis have shown poor correlation with these outcomes.[17,18,20,41]

We conducted an exploratory analysis to determine whether histologic scoring of NAFLD could be optimized or simplified using reliable items that were most strongly correlated with the disease activity VAS. Our final exploratory model, comprising the expanded NAS hepatocyte ballooning item (0–4 × 2) and the Goodman classification MDB item

(0–3 × 1), demonstrated intrarater and interrater ICCs of 0.90 and 0.80, respectively. This index could have particularly good interobserver agreement for patients with comorbid diabetes, which has a known association with more prominent ballooning and MDBs. The simplicity and reliability of this 2-item expanded NAS suggest that its ability to predict liver-related outcomes and fibrosis progression in clinical trials should be formally validated. In particular, the reliability of this index should ideally be tested by other groups of pathologists, and the relative responsiveness to change following effective therapy should be quantified.

Correlations between fibrosis measures and a fibrosis VAS were also evaluated to identify items that best capture the spectrum of fibrosis seen in NASH. While all fibrosis measures had strong correlations with the fibrosis VAS, the Ishak fibrosis staging system and the expanded NASH CRN fibrosis staging system that includes substages for stage 3 had the strongest correlations. Furthermore, visual inspection of the fibrosis VAS stratified by staging system demonstrated a marked difference in mean VAS scores between stages 2 and 3 in the original NASH CRN staging system, underscoring the large diversity of fibrosis seen in stage 3. The Ishak and expanded NASH CRN staging systems have a more uniform change in VAS for each increase in stage. Given that expanding the fibrosis staging did not decrease interrater reliability, inclusion of these staging systems in future clinical trials should be considered and may be especially helpful in studies aiming to identify fibrosis regression. However, additional study is needed to determine the clinical significance of a change in fibrosis stage for these systems.

This study has some methodological strengths. To identify the most complete list of appropriate histologic measures of disease activity for this reliability study, we used the rigorous RAND/UCLA methodology that captures expert insights from a panel of internationally recognized liver pathologists and hepatologists and aims to minimize bias.[26] The central readers were expert liver pathologists with expertise in NAFLD who were involved in the prior RAND/UCLA study and participated in multiple additional discussions and standardized training on scoring all histologic indices and items in this study. This extensive preparation may have helped to improve the reliability of existing histologic indices in this study in comparison with that observed previously.[37] Standardized definitions of features may improve agreement on scoring, as demonstrated in the SAF literature.[15] Additionally, most liver biopsy slides were considered good quality, another important factor in improving interrater agreement. Furthermore, this reliability study is an important step towards the goal of standardizing the use of fully validated histologic measures of disease activity for more efficient clinical trials of NASH, for which no treatments are currently approved.[3]

The limitations of this study should be acknowledged. First, slides for the study were acquired from a single center rather than from a randomized controlled trial that would have allowed for assessment of both baseline and posttreatment biopsies to further ensure representation of a broad range of disease activity. Reliability testing and correlations with disease activity should be evaluated in a larger data set involving multiple centers. Second, the time interval of ≥2 weeks between the 2 readings for each measure is relatively short, and a longer interval could have affected the performance of the measures. Third, the

study was powered to evaluate the reliability of existing NAFLD histologic indices and items, additional RAND/UCLA-identified items, and a disease activity VAS. However, the use of our reliability findings to investigate whether scoring tools could be optimized using a single histologic index comprising selected items should be regarded as exploratory only. The 2 items of this index are highly correlated, with MDBs often present in classic ballooned hepatocytes, and may reflect bias toward the existing literature.[14,17,20,21,40] However, these 2 distinct features were evaluated separately in prior studies and had independent associations with fibrosis progression and regression.[13,14,19] In addition, while the expanded hepatocyte ballooning item from the new index has not been formally validated, it has been used by the NASH CRN for several years.[27] Finally, formal validation of the expanded NAS requires further analysis, including validation in a larger clinical trial setting of medically diverse cohorts (eg, nondiabetic obese patients).

In conclusion, we showed that intrarater and interrater reliability of existing histologic NAFLD indices and fibrosis staging systems can be improved (vs previous findings[37]) with extensive training and preparation among a group of 4 experienced pathologists. Among component items of existing indices and the additional RAND/UCLA-identified items, histologic measures of hepatocyte ballooning and MDBs had the largest correlations with disease activity and were incorporated into a novel exploratory index, the expanded NAS. Further validation including analysis of the responsiveness to change after a therapeutic intervention is required for the histologic indices and items evaluated in this study.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

FINANCIAL SUPPORT

This study was supported in part by the Intramural Research Program of the National Cancer Institute, NIH

CONFLICTS OF INTEREST

RKP has received consulting fees from AbbVie, Eli Lilly, Allergan, Genentech, and Alimentiv Inc.

VJ has received consulting/advisory board fees from AbbVie, Alimentiv Inc, Arena pharmaceuticals, Asieris, Bristol Myers Squibb, Celltrion, Eli Lilly, Ferring, Fresenius Kabi, Galapagos, GlaxoSmithKline, Genentech, Gilead, Janssen, Merck, Mylan, Pandion, Pendopharm, Pfizer, Reistone Biopharma, Roche, Sandoz, Takeda, Topivert; speaker's fees from, Abbvie, Ferring, Janssen Pfizer Shire, Takeda

MH is an employee of Alimentiv Inc.

GZ and his institution have received consulting fees from Alimentiv Inc.

OAA has nothing to disclose.

QMA is coordinator of the EU IMI-2 LITMUS consortium, which is funded by the EU Horizon 2020 program and EFPIA. In addition, his institution has received grant support from Allergan/Tobira, AstraZeneca, GlaxoSmithKline, Glympse Bio, Novartis Pharma AG, Pfizer Ltd.; he serves as a consultant or advisory board member for 89Bio, Allergan/Tobira, Altimmune, AstraZeneca, Axcella, Blade, BMS, BNN Cardio, Cirius, CymaBay, EcoR1, E3Bio, Eli Lilly & Company Ltd., Galmed, Genentech, Genfit SA, Gilead, Grunthal, HistoIndex, Indalo, Intercept Pharma Europe Ltd., Inventiva, IQVIA, Janssen, Madrigal, MedImmune, Medpace, Metacrine, NGMBio, North Sea Therapeutics, Novartis, Novo Nordisk A/S, PathAI, Pfizer Ltd., Poxel, ProSciento, Raptor Pharma, Roche,

Servier, Terns, The Medicines Company, Viking Therapeutics; royalties/speakers fees from Abbott Laboratories, Allergan/Tobira, BMS, Clinical Care Options, Falk, Fishawack, Genfit SA, Gilead, Integritas Communications, Kenes, Medscape, Elsevier Ltd.

BAA has nothing to disclose.

CB's institution currently receives support for biopsy-related work through laboratory services and/or consulting agreements with: Akero Therapeutics Inc, Allergan, Celgene, Covance, Eli Lilly, E study site, Genesis Imaging Service, ICON, Medical Research Group, Inc., Pinnacle Clinical Research, Southern California Research Center, Imaging Health Care Specialists, Imperial Valley Family Medical Care Group, Leica, Mayo Laboratories, Roche-Ventana, Sharp Community Medical Group, Sharp Reese Stealy Medical Group, and Sharp Healthcare.

EJC has nothing to disclose.

ADC has nothing to disclose.

KC serves on the scientific advisory board for Novo Nordisk and BMS and has received grant funding from Boehringer-Ingelheim, BMS, and Novartis.

BGF is a scientific advisory board member for AbbVie, Allergan, Amgen, AstraZeneca, Avaxia Biologics Inc., Boehringer-Ingelheim, Bristol-Myers Squibb, Celgene, Elan, Biogen, Ferring, Genentech–Roche, Janssen–Johnson & Johnson, Merck, Millennium, Nestlé, Novo Nordisk, Novartis, Pfizer, Prometheus, Protagonist, Receptos, Salix, Sigmoid Pharma, Takeda, Teva, TiGenix, Tillotts Pharma, and UCB Pharma; consulting fees from AbbVie, Actogenix, Akros, Albireo Pharma, Allergan, Amgen, AstraZeneca, Avaxia Biologics, Avir Pharma, Axcan, Baxter Healthcare, Biogen Idec, Boehringer Ingelheim, Bristol-Myers Squibb, Calypso Biotech, Celgene, Elan–Biogen, EnGene, Ferring, Genentech–Roche, GiCare Pharma, Gilead Sciences, Given Imaging, GlaxoSmithKline, Ironwood, Janssen Biotech–Centocor, Janssen–Johnson & Johnson, Kyowa Hakko Kirin, Eli Lilly, Merck, Mesoblast Pharma, Millennium, Nestlé, Novo Nordisk, Novartis, Pfizer, Prometheus, Protagonist, Receptos Salix, Sanofi, Shire, Sigmoid Pharma, Synergy Pharma, Takeda, Teva, TiGenix, Tillotts Pharma, UCB Pharma, Vertex, VHsquared, Wyeth, Zealand, and Zygenia; lecture fees from AbbVie, Janssen–Johnson & Johnson, Takeda, and UCB Pharma, and grant support from AbbVie, Amgen, AstraZeneca, Bristol-Myers Squibb, Janssen Biotech–Centocor, Janssen–Johnson & Johnson, Pfizer, Receptos, Sanofi, and Takeda, and is the Senior Scientific Officer of Alimentiv Inc.

DEK has nothing to disclose.

CM has received consulting fees from AbbVie, Alimentiv, Amgen, AVIR Pharma Inc, Bristol Myers Squibb, Ferring, Fresenius Kabi, Janssen, McKesson, Mylan, Takeda, Pfizer, and Roche; speaker's fees from AbbVie, Amgen, AVIR Pharma Inc, Alimentiv, Ferring, Janssen, Takeda, and Pfizer; and research support from Pfizer.

SCM is an employee of Alimentiv Inc.

MN has received advisory fees from Allergan, Gilead, and Novartis; consulting fees from Allergan, Gilead, Novartis, 89Bio, Intercept, Pfizer, Novo Nordisk, Blade, Echosens, Fractyl, Terns, OWL, Siemens, Roche Diagnostic, and Abbott; grants from Viking, Bristol Myers Squibb, Galmed, Galectin, Genfit, Conatus, Enanta, Madrigal, Shire, and Zydus; and owns stock in Anaetos and Viking.

VR has received consulting fees for Boehringer-Ingelheim, Novo-Nordisk, Galmed, Terns, Theratechnologies, Bristol-Myers-Squibb, Genfit, Madrigal, and NGM Bio.

MAV has received consulting fees from Alimentiv and AcelaBio.

ZMY has received research funding and/or consulting fees from Gilead Sciences, Intercept, BMS, NovoNordisk, Viking, Terns, Siemens, Quest, Abbvie, Madridgal, Merck, Abbott, and Novartis.

SAH has received grant and research support from Akero, Axcella, Cirius, Civi Biopharma, Cymabay, Enyo, Galectin, Galmed, Genfit, Gilead, Hepion, Hightide, Intercept, Madrigal, Metacrine, NGM, Northsea, Novartis, Novo Nordisk, Poxel, Sagimet and Viking and is a consulting adviser for Akero, Alentis, Alimentiv, Altimmune, Arrowhead, Axcella, Boston Pharmaceuticals, B Riley FBR, Canfite, Chronwell, Cirius, Civi Biopharma, Corcept, Cymabay, Echosens, Enyo, Fibronostics, Foresite Labs, Fortress, Galectin, Galmend, Genfit, Gilead, GNS Healthcare, Hepion, Hightide, Histoindex, Inipharm, Intercept, Ionis, Madrigal, Medpace, Metacrine, Microba, NGM, Northsea, Novartis, Novo Nordisk, Nutrasource, PathAI, Piper Sandler & Co, Poxel, Prometic, Sagimet, Sonic Incytes Medical Corp, Terns, Theratechnologies and Viking and holds shares in Akero, Chronwell, Cirius, Galectin, Genfit, Hepion, Histoindex, Metacrine, NGM, Northsea and Sonic Incytes Medical Corp.

RL serves as a consultant or advisory board member for Arrowhead Pharmaceuticals, AstraZeneca, Bird Rock Bio, Boehringer Ingelheim, Bristol-Myer Squibb, Celgene, Cirius, CohBar, Conatus, Eli Lilly, Galmed, Gemphire, Gilead, Glympse bio, GNI, GRI Bio, Intercept, Ionis, Janssen Inc., Merck, Metacrine, Inc., NGM Biopharmaceuticals, Novartis, Novo Nordisk, Pfizer, Prometheus, Sanofi, Siemens, and Viking Therapeutics. In addition, his institution has received grant support from Allergan, Boehringer-Ingelheim, Bristol-Myers Squibb, Cirius, Eli Lilly and Company, Galectin Therapeutics, Galmed Pharmaceuticals, GE, Genfit, Gilead, Intercept, Janssen, Madrigal Pharmaceuticals, Merck, NGM Biopharmaceuticals, NuSirt, Pfizer, Prometheus, and Siemens. He is also co-founder of Liponexus, Inc.

Alimentiv, Inc. is an academic gastrointestinal contract research organization (CRO), operating under the Alimentiv Health Trust. Alimentiv, Inc. provides centralized imaging management solutions in clinical trials, including endoscopy, histopathology, and magnetic resonance imaging. as well as precision medicine services. The beneficiaries of this trust are the employees of the enterprises it holds. None of the academic authors are beneficiaries of the Alimentiv Health Trust. RKP, VJ, GZ, QMA, CM, BGF, SAH, and RL are consultants to Alimentiv and have neither equity positions nor shares in the corporation.

LIST OF ABBREVIATIONS:

ALT	alanine aminotransferase
AST	aspartate aminotransferase
FDA	U.S. Food and Drug Administration
ICC	intraclass correlation coefficients
MDB	Mallory-Denk body
NAFL	nonalcoholic fatty liver
NAS	nonalcoholic fatty liver disease activity score
NASH CRN	Nonalcoholic Steatohepatitis Clinical Research Network
RAND/UCLA	Research and Development/University of California Los Angeles
SAF	steatosis, activity, fibrosis
VAS	visual analog scale

REFERENCES

1. Chalasani N, Younossi Z, Lavine JE, Charlton M, Cusi K, Rinella M, et al. The diagnosis and management of nonalcoholic fatty liver disease: practice guidance from the American Association for the Study of Liver Diseases. Hepatology. 2018;67:328–357. [PubMed: 28714183]
2. Kleiner DE, Makhlouf HR. Histology of nonalcoholic fatty liver disease and nonalcoholic steatohepatitis in adults and children. Clin Liver Dis. 2016;20:293–312. [PubMed: 27063270]
3. Siddiqui MS, Harrison SA, Abdelmalek MF, Anstee QM, Bedossa P, Castera L, et al. Case definitions for inclusion and analysis of endpoints in clinical trials for nonalcoholic steatohepatitis through the lens of regulatory science. Hepatology. 2018;67:2001–2012. [PubMed: 29059456]
4. Singh S, Allen AM, Wang Z, Prokop LJ, Murad MH, Loomba R. Fibrosis progression in nonalcoholic fatty liver vs nonalcoholic steatohepatitis: a systematic review and meta-analysis of paired-biopsy studies. Clin Gastroenterol Hepatol. 2015;13:643–654. [PubMed: 24768810]
5. Guirguis E, Grace Y, Bolson A, DellaVecchia MJ, Ruble M. Emerging therapies for the treatment of nonalcoholic steatohepatitis: a systematic review. Pharmacotherapy. 2021;41:315–328. [PubMed: 33278029]

6. Harrison SA, Bashir MR, Guy CD, Zhou R, Moylan CA, Frias JP, et al. Resmetirom (MGL-3196) for the treatment of non-alcoholic steatohepatitis: a multicentre, randomised, double-blind, placebo-controlled, phase 2 trial. Lancet. 2019;394:2012–2024. [PubMed: 31727409]
7. Neuschwander-Tetri BA, Loomba R, Sanyal AJ, Lavine JE, Van Natta ML, Abdelmalek MF, et al. Farnesoid X nuclear receptor ligand obeticholic acid for non-cirrhotic, non-alcoholic steatohepatitis (FLINT): a multicentre, randomised, placebo-controlled trial. Lancet. 2015;385:956–965. [PubMed: 25468160]
8. Ratziu V, Harrison SA, Francque S, Bedossa P, Lehert P, Serfaty L, et al. Elafibranor, an agonist of the peroxisome proliferator-activated receptor-alpha and -delta, induces resolution of nonalcoholic steatohepatitis without fibrosis worsening. Gastroenterology. 2016;150:1147–1159. [PubMed: 26874076]
9. Noureddin M, Vipani A, Bresee C, Todo T, Kim IK, Alkhouri N, et al. NASH leading cause of liver transplant in women: updated analysis of indications for liver transplant and ethnic and gender variances. Am J Gastroenterol. 2018;113:1649–1659. [PubMed: 29880964]
10. U.S. Food and Drug Administration. Noncirrhotic nonalcoholic steatohepatitis with liver fibrosis: developing drugs for treatment; draft guidance for industry. https://www.federalregister.gov/documents/2018/12/04/2018-26333/noncirrhotic-nonalcoholic-steatohepatitis-with-liver-fibrosis-developing-drugs-for-treatment-draft. Published December 4, 2018. Accessed February 10, 2021.
11. Brunt EM, Janney CG, Di Bisceglie AM, Neuschwander-Tetri BA, Bacon BR. Nonalcoholic steatohepatitis: a proposal for grading and staging the histological lesions. Am J Gastroenterol. 1999;94:2467–2474. [PubMed: 10484010]
12. Harrison SA, Torgerson S, Hayashi P, Ward J, Schenker S. Vitamin E and vitamin C treatment improves fibrosis in patients with nonalcoholic steatohepatitis. Am J Gastroenterol. 2003;98:2485–2490. [PubMed: 14638353]
13. 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;41:1313–1321. [PubMed: 15915461]
14. Younossi ZM, Stepanova M, Rafiq N, Makhlouf H, Younoszai Z, Agrawal R, et al. Pathologic criteria for nonalcoholic steatohepatitis: interprotocol agreement and ability to predict liver-related mortality. Hepatology. 2011;53:1874–1882. [PubMed: 21360720]
15. Bedossa P, Poitou C, Veyrie N, Bouillot JL, Basdevant A, Paradis V, et al. Histopathological algorithm and scoring system for evaluation of liver lesions in morbidly obese patients. Hepatology. 2012;56:1751–1759. [PubMed: 22707395]
16. Ishak K, Baptista A, Bianchi L, Callea F, De Groote J, Gudat F, et al. Histological grading and staging of chronic hepatitis. J Hepatol. 1995;22:696–699. [PubMed: 7560864]
17. Angulo P, Kleiner DE, Dam-Larsen S, Adams LA, Bjornsson ES, Charatcharoenwitthaya P, et al. Liver fibrosis, but no other histologic features, Is associated with long-term outcomes of patients with nonalcoholic fatty liver disease. Gastroenterology. 2015;149:389–397. [PubMed: 25935633]
18. Hagstrom H, Nasr P, Ekstedt M, Hammar U, Stal P, Hultcrantz R, et al. Fibrosis stage but not NASH predicts mortality and time to development of severe liver disease in biopsy-proven NAFLD. J Hepatol. 2017;67:1265–1273. [PubMed: 28803953]
19. 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.
20. Matteoni CA, Younossi ZM, Gramlich T, Boparai N, Liu YC, McCullough AJ. Nonalcoholic fatty liver disease: a spectrum of clinical and pathological severity. Gastroenterology. 1999;116:1413–1419. [PubMed: 10348825]
21. Gramlich T, Kleiner DE, McCullough AJ, Matteoni CA, Boparai N, Younossi ZM. Pathologic features associated with fibrosis in nonalcoholic fatty liver disease. Hum Pathol. 2004;35:196–199. [PubMed: 14991537]
22. Brunt EM, Kleiner DE, Wilson LA, Sanyal AJ, Neuschwander-Tetri BA, Nonalcoholic Steatohepatitis Clinical Research Network. Improvements in histologic features and diagnosis associated with improvement in fibrosis in nonalcoholic steatohepatitis: results from

the Nonalcoholic Steatohepatitis Clinical Research Network treatment trials. Hepatology. 2019;70:522–531. [PubMed: 30549292]

23. Brunt EM, Kleiner DE, Wilson LA, Unalp A, Behling CE, Lavine JE, et al. Portal chronic inflammation in nonalcoholic fatty liver disease (NAFLD): a histologic marker of advanced NAFLD-Clinicopathologic correlations from the Nonalcoholic Steatohepatitis Clinical Research Network. Hepatology. 2009;49:809–820. [PubMed: 19142989]
24. Rakha EA, Adamson L, Bell E, Neal K, Ryder SD, Kaye PV, et al. Portal inflammation is associated with advanced histological changes in alcoholic and non-alcoholic fatty liver disease. J Clin Pathol. 2010;63:790–795. [PubMed: 20819880]
25. Brunt EM, Neuschwander-Tetri BA, Oliver D, Wehmeier KR, Bacon BR. Nonalcoholic steatohepatitis: histologic features and clinical correlations with 30 blinded biopsy specimens. Hum Pathol. 2004;35:1070–1082. [PubMed: 15343508]
26. Pai RK, Kleiner DE, Hart J, Adeyi OA, Clouston AD, Behling CA, et al. Standardising the interpretation of liver biopsies in non-alcoholic fatty liver disease clinical trials. Aliment Pharmacol Ther. 2019;50:1100–1111. [PubMed: 31583739]
27. Kleiner DE, Brunt EM, Belt PH, Behling CA, Gill RM, Guy CD, et al. Extending the ballooning score beyond 2: a proposal for a new balloon score [Abstract]. Hepatology. 2015;62(suppl 1):288A–289A.
28. Eliasziw M, Young SL, Woodbury MG, Fryday-Field K. Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. Phys Ther. 1994;74:777–788. [PubMed: 8047565]
29. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas. 1973;33:613–619.
30. Hamlett A, Ryan L, Serrano-Trespalacios P, Wolfinger R. Mixed models for assessing correlation in the presence of replication. J Air Waste Manag Assoc. 2003;53:442–450. [PubMed: 12708508]
31. Kowalik D, Choi Y-H, Zou GY. Confidence interval estimation for a difference between two dependent intraclass correlation coefficients with variable class sizes. Journal of Statistical Theory and Practice. 2011;5:613–625.
32. Zou GY. Toward using confidence intervals to compare correlations. Psychol Methods. 2007;12:399–413. [PubMed: 18179351]
33. Cohen J. A power primer. Psychol Bull. 1992;112:155–159. [PubMed: 19565683]
34. Zou GY. Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. Stat Med. 2012;31:3972–3981. [PubMed: 22764084]
35. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–174. [PubMed: 843571]
36. U.S. Food and Drug Administration (FDA). Clinical trial imaging endpoint process standards. Guidance for industry. https://www.fda.gov/files/drugs/published/Clinical-Trial-Imaging-Endpoint-Process-Standards-Guidance-for-Industry.pdf. Published April 2018. Accessed February 10, 2021.
37. 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;73:1322–1332. [PubMed: 32610115]
38. Jairath V, Jeyarajah J, Zou G, Parker CE, Olson A, Khanna R, et al. A composite disease activity index for early drug development in ulcerative colitis: development and validation of the UC-100 score. Lancet Gastroenterol Hepatol. 2019;4:63–70. [PubMed: 30343116]
39. Mosli MH, Feagan BG, Zou G, Sandborn WJ, D'Haens G, Khanna R, et al. Development and validation of a histological index for UC. Gut. 2017;66:50–58. [PubMed: 26475633]
40. Richardson MM, Jonsson JR, Powell EE, Brunt EM, Neuschwander-Tetri BA, Bhathal PS, et al. Progressive fibrosis in nonalcoholic steatohepatitis: association with altered regeneration and a ductular reaction. Gastroenterology. 2007;133:80–90. [PubMed: 17631134]
41. Younossi ZM, Gramlich T, Liu YC, Matteoni C, Petrelli M, Goldblum J, et al. Nonalcoholic fatty liver disease: assessment of variability in pathologic interpretations. Mod Pathol. 1998;11:560–565. [PubMed: 9647594]

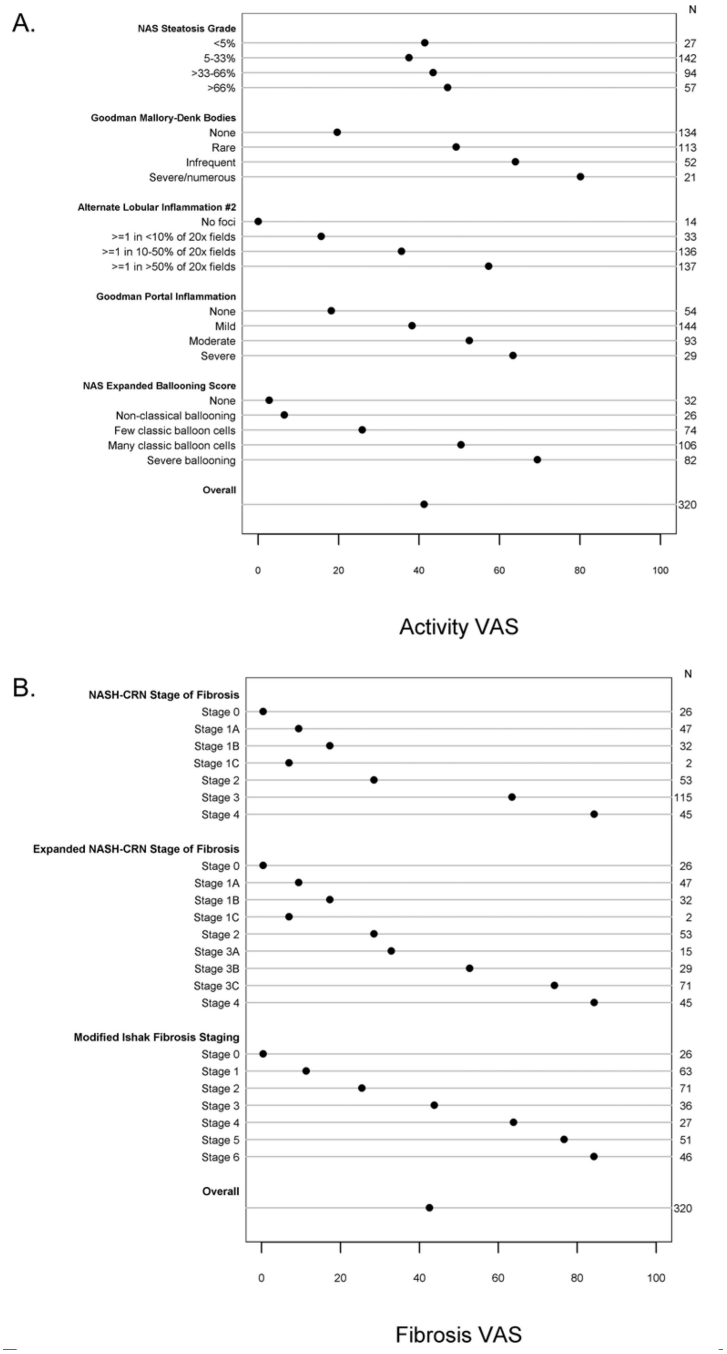


Fig. 1.

Univariable summaries of mean disease activity VAS scores stratified by levels of candidate items (A) and mean fibrosis VAS scores stratified by levels of selected staging systems (B). Abbreviation: VAS, visual analog scale.

Table 1.

Demographics and patient characteristics of patients represented in the histology slides

Characteristic	Patients (n = 40)
Age (y), mean ± SD	55.30 ± 11.90
Male, No. (%)	15 (37.50)
Diabetes, No. (%)	17 (42.50)
Hypertension, No. (%)	23 (57.50)
Statin, No. (%)	15 (37.50)
Non-Hispanic Race, No. (%)	37 (92.50)
BMI (kg/m^2), mean ± SD	35.70 ± 8.20
AST (U/L), mean ± SD	76.1 ± 46.6
ALT (U/L), mean ± SD	99.8 ± 84.9
Alkaline phosphatase (U/L), mean ± SD	91.2 ± 54.8
Total bilirubin (mg/dL), mean ± SD	0.65 ± 0.50
LDL-C (mg/dL), mean ± SD	105.1 ± 32.2
HDL-C (mg/dL), mean ± SD	46.6 ± 12.8
Triglycerides (mg/dL), mean ± SD	165.3 ± 110.7

Abbreviations: ALT, alanine aminotransferase; AST, aspartate aminotransferase; BMI, body mass index.

Table 2.

Intrarater and interrater reliability of existing NAFLD histologic indices and their components and correlations with the disease activity and fibrosis VAS

Indices of NAFLD Disease Activity/Severity and Component Items		Reliability ICC (95% CI)		Correlation, r (95% CI) with	
Feature	**Index/Item/Measure**	**Intrarater**	**Interrater**	**Disease Activity VAS**	**Fibrosis VAS**
Overall Disease Activity	Brunt necro-inflammatory grade of steatohepatitis	0.82 (0.71, 0.88)	0.70 (0.55, 0.79)	0.88 (0.85, 0.91)	0.59 (0.38, 0.67)
	Modified Brunt necro-inflammatory grade *	0.80 (0.70, 0.86)	0.60 (0.45, 0.70)	0.84 (0.76, 0.86)	0.56 (0.32, 0.60)
	NAS activity score **	0.85 (0.75, 0.90)	0.72 (0.58, 0.80)	0.69 (0.51, 0.76)	0.37 (–0.06, 0.48)
	SAF activity grade †	0.82 (0.68, 0.89)	0.63 (0.40, 0.75)	0.66 (0.52, 0.71)	0.47 (0.12, 0.55)
	Disease activity VAS	0.86 (0.79, 0.91)	0.69 (0.57, 0.77)	N/A	0.61 (0.37, 0.68)
Hepatocyte Ballooning and MDBs	Modified Brunt hepatocyte degeneration/necrosis	0.82 (0.72, 0.88)	0.69 (0.56, 0.79)	0.86 (0.83, 0.89)	0.56 (0.36, 0.65)
	NAS hepatocyte ballooning	0.86 (0.76, 0.92)	0.73 (0.57, 0.82)	0.85 (0.81, 0.87)	0.57 (0.34, 0.67)
	Goodman hepatocellular ballooning	0.87 (0.78, 0.92)	0.73 (0.60, 0.81)	0.88 (0.84, 0.90)	0.55 (0.30, 0.65)
	Goodman MDBs	0.78 (0.68, 0.85)	0.57 (0.46, 0.66)	0.76 (0.59, 0.82)	0.49 (0.26, 0.56)
	SAF hepatocyte ballooning	0.85 (0.76, 0.91)	0.68 (0.45, 0.82)	0.66 (0.49, 0.73)	0.50 (0.24, 0.58)
	Ballooning VAS	0.88 (0.81, 0.92)	0.74 (0.62, 0.82)	0.96 (0.93, 0.97)	0.60 (0.38, 0.68)
Inflammation	Modified Brunt inflammation	0.61 (0.48, 0.72)	0.36 (0.23, 0.47)	0.66 (0.49, 0.69)	0.44 (0.20, 0.48)
	NAS lobular inflammation	0.59 (0.41, 0.72)	0.39 (0.21, 0.52)	0.59 (0.46, 0.65)	0.35 (0.14, 0.40)
	Goodman lobular inflammation	0.60 (0.44, 0.72)	0.37 (0.22, 0.49)	0.64 (0.54, 0.69)	0.39 (0.10, 0.44)
	Goodman portal inflammation	0.57 (0.45, 0.69)	0.52 (0.38, 0.62)	0.51 (0.23, 0.57)	0.56 (0.53, 0.60)
	SAF lobular inflammation	0.62 (0.48, 0.74)	0.28 (0.13, 0.41)	0.45 (0.27, 0.53)	0.28 (–0.05, 0.36)
	Lobular inflammation VAS	0.71 (0.58, 0.80)	0.40 (0.22, 0.54)	0.63 (0.44, 0.70)	0.40 (0.12, 0.44)
Steatosis	NAS steatosis	0.87 (0.80, 0.91)	0.80 (0.70, 0.87)	0.12 (−0.13, 0.28)	−0.06 (−0.44, 0.15)
	Goodman steatosis	0.80 (0.66, 0.88)	0.72 (0.58, 0.82)	0.07 (−0.14, 0.25)	−0.10 (−0.41, 0.12)
	SAF steatosis	0.87 (0.80, 0.91)	0.80 (0.70, 0.87)	0.12 (−0.13, 0.28)	−0.06 (−0.44, 0.15)
	Steatosis VAS	0.92 (0.88, 0.95)	0.85 (0.75, 0.91)	0.06 (−0.12, 0.23)	−0.11 (−0.45, 0.07)
Fibrosis	Brunt stage of fibrosis	0.90 (0.84, 0.94)	0.79 (0.68, 0.87)	0.56 (0.32, 0.63)	0.92 (0.90, 0.94)
	Modified Brunt stage of fibrosis	0.90 (0.84, 0.94)	0.79 (0.68, 0.87)	0.56 (0.32, 0.63)	0.92 (0.90, 0.94)
	NAS stage of fibrosis	0.92 (0.86, 0.96)	0.77 (0.64, 0.86)	0.55 (0.29, 0.64)	0.88 (0.85, 0.91)
	Goodman pericellular/perisinusoidal fibrosis	0.86 (0.76, 0.92)	0.70 (0.54, 0.79)	0.57 (0.33, 0.67)	0.81 (0.77, 0.84)

Indices of NAFLD Disease Activity/Severity and Component Items		Reliability ICC (95% CI)		Correlation, r (95% CI) with	
Feature	**Index/Item/Measure**	**Intrarater**	**Interrater**	**Disease Activity VAS**	**Fibrosis VAS**
	Goodman portal fibrosis	0.91 (0.86, 0.94)	0.82 (0.75, 0.87)	0.51 (0.31, 0.57)	0.91 (0.89, 0.95)
	Goodman bridging fibrosis	0.93 (0.88, 0.96)	0.84 (0.74, 0.92)	0.49 (0.30, 0.55)	0.93 (0.91, 0.95)
	Goodman cirrhosis	0.87 (0.77, 0.93)	0.75 (0.56, 0.87)	0.39 (0.30, 0.44)	0.77 (0.70, 0.88)
	SAF stage of fibrosis	0.92 (0.86, 0.96)	0.77 (0.64, 0.86)	0.55 (0.29, 0.64)	0.88 (0.85, 0.91)
	Ishak fibrosis staging[≠]	0.94 (0.89, 0.96)	0.87 (0.79, 0.92)	0.55 (0.32, 0.62)	0.96 (0.95, 0.97)
	Fibrosis VAS	0.94 (0.89, 0.97)	0.86 (0.79, 0.91)	0.61 (0.37, 0.68)	N/A

Abbreviations: ICC, intraclass correlation coefficient; NAS, NAFLD activity score; SAF, steatosis, activity, fibrosis; VAS, visual analog scale.

[*]The necro-inflammatory grade is calculated by summing the scores for inflammation and hepatocyte degeneration/necrosis (range, 0–6).

[**]The total NAS is the unweighted sum of the scores for the 3 histologic items (range, 0–8), with higher scores representing more severe NAFLD activity.

[†]The SAF activity grade is calculated by summing the scores for lobular inflammation and hepatocyte ballooning.

[≠]Modified slightly to account for fibrosis in the central areas for stages 1 and 2.

Table 3.

Intrarater and interrater reliability of additional items identified through the RAND/UCLA appropriateness process and correlations with the disease activity and fibrosis VAS

Item		Reliability ICC (95% CI)		Correlation, r (95% CI) with	
Feature	**Description**	**Intrarater**	**Interrater**	**Disease Activity VAS**	**Fibrosis VAS**
Hepatocyte Ballooning and MDBs	NAS expanded ballooning score	0.89 (0.82, 0.93)	0.79 (0.68, 0.85)	0.88 (0.84, 0.90)	0.59 (0.31, 0.69)
	Alternate expanded ballooning score	0.89 (0.82, 0.93)	0.77 (0.64, 0.86)	0.89 (0.85, 0.90)	0.55 (0.30, 0.66)
	Alternate 4-point ballooning score	0.88 (0.80, 0.93)	0.71 (0.56, 0.81)	0.87 (0.83, 0.89)	0.55 (0.34, 0.65)
	MDBs	0.77 (0.66, 0.86)	0.51 (0.38, 0.62)	0.76 (0.53, 0.80)	0.51 (0.24, 0.59)
Inflammation	Alternate lobular inflammation #1*	0.57 (0.40, 0.70)	0.36 (0.20, 0.49)	0.62 (0.49, 0.68)	0.38 (0.13, 0.43)
	Alternate lobular inflammation #2**	0.69 (0.54, 0.79)	0.41 (0.22, 0.54)	0.63 (0.49, 0.67)	0.41 (0.17, 0.44)
	Intra-acinar neutrophils	0.54 (0.41, 0.66)	0.48 (0.33, 0.59)	0.60 (0.37, 0.66)	0.42 (0.15, 0.48)
	Apoptotic hepatocytes (acidophil bodies)	0.47 (0.35, 0.61)	0.42 (0.27, 0.54)	0.52 (0.42, 0.55)	0.35 (0.16, 0.39)
	Goodman inflammation within portal tracts and fibrous areas (based on most severe)	0.74 (0.63, 0.82)	0.64 (0.52, 0.73)	0.51 (0.23, 0.55)	0.57 (0.51, 0.66)
Steatosis	Size of steatotic droplets	0.44 (0.22, 0.62)	0.07 (0, 0.16)	−0.02 (−0.21, 0.09)	−0.03 (−0.17, 0.05)
	Steatosis as measured by % area of non-fibrotic parenchyma replaced by steatosis, evaluated at low to medium power (4× to 10×)	0.86 (0.78, 0.91)	0.80 (0.69, 0.87)	0.08 (−0.15, 0.25)	−0.09 (−0.45, 0.12)
	Microvesicular steatosis	0.53 (0.41, 0.65)	0.05 (0, 0.11)	0.06 (−0.02, 0.15)	0.04 (−0.07, 0.08)
	3-tier microvesicular steatosis	0.63 (0.50, 0.74)	0.07 (0, 0.13)	0.03 (−0.05, 0.13)	0.02 (−0.10, 0.04)
Fibrosis	Alternate fibrosis staging system - stage of zone 3/sinusoidal fibrosis	0.87 (0.81, 0.92)	0.76 (0.65, 0.83)	0.58 (0.30, 0.66)	0.86 (0.82, 0.89)
	Alternate fibrosis staging system - substages for fibrosis from the portal area	0.93 (0.89, 0.96)	0.78 (0.67, 0.85)	0.49 (0.29, 0.56)	0.93 (0.91, 0.95)
	Expanded NASH CRN fibrosis†	0.94 (0.90, 0.96)	0.83 (0.74, 0.90)	0.57 (0.30, 0.64)	0.95 (0.94, 0.96)

*Abbreviations: ICC, intraclass correlation coefficient; MDB, Mallory-Denk body; NAS, NAFLD activity score; NASH CRN, Nonalcoholic Steatohepatitis Clinical Research Network; VAS, visual analog scale. Defines a *focus* as 5 or more inflammatory cells, score is based on average of all 20× fields.

**Defines a *focus* as 5 or more inflammatory cells within the lobule.

†This item (range, 0–8) expands stage 3 of the NAS fibrosis item into 3 levels: stage 3a, one definite fibrous bridge; stage 3b, few fibrous bridges without nodule formation; stage 3c, complex bridging (many bridges) with or without rare nodule formation.

Table 4.

Items and weights for the candidate index

Item	Full model coefficient (SE)	Final model coefficient (SE)	Standardized coefficient
(Intercept)	−3.4 (5.3)	—	—
NAS steatosis grade	−1.5 (1.9)	—	—
0 <5%			
1 5%−33%			
2 >33%−66%			
3 >66%			
Goodman MDBs	8.6 (2.4)	8.7	1
0 None			
1 Rare			
2 Infrequent			
3 Severe/numerous			
Alternate lobular inflammation #2*	3.8 (2.8)	—	—
0 No foci			
1 ≥1 in <10% of 20× fields			
2 ≥1 in 10%−50% of 20× fields			
3 ≥1 in >50% of 20× fields			
Goodman portal inflammation based on most severe portal tract	0.55 (2.2)	—	—
0 None			
1 Mild			
2 Moderate			
3 Severe			
NAS expanded ballooning score	11.9 (2.4)	13.7	2
0 None			
1 Nonclassical ballooning			
2 Few classic balloon cells			
3 Many classic balloon cells, but not severe			
4 Severe ballooning			
R^2	0.869	0.859	—

Abbreviations: MDB, Mallory-Denk body; NAS, NAFLD activity score; SE, standard error.

*Defines a *focus* as 5 or more inflammatory cells within the lobule.