Reliability of the nonalcoholic steatohepatitis clinical research network and steatosis activity fibrosis histological scoring systems

Howard Ho-Wai Leung,* Pavitratha Puspanathan,† Anthony Wing-Hung Chan,* Nik Raihan Nik Mustapha,†Vincent Wai-Sun Wong‡ and Wah-Kheong Chan§
Departments of *Anatomical and Cellular Pathology, Prince of Wales Hospital, ‡Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong, China; †Department of Pathology, Hospital Sultanah Bahiyah, Alor Setar, §Gastroenterology and Hepatology Unit, Department of Medicine, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia

Introduction

In 2005, Kleiner et al. developed and validated a system of histological valuation for non-alcoholic fatty liver disease (NAFLD) that would allow for assessment of therapeutic response for the Non-alcoholic Steatohepatitis Clinical Research Network (NASH CRN).[1] The NAFLD activity score (NAS), which included only features of active injury that are potentially reversible in the short term, was proposed. The NAS was defined as the unweighted sum of the scores for steatosis, lobular inflammation, and ballooning. Eventually, NASH was to be defined as the presence of at least grade 1 each of steatosis, hepatocyte ballooning, and lobular inflammation.[2]
In 2012, Bedossa et al. developed and validated an algorithm for categorization (subsequently called the fatty liver inhibition of progression or FLIP algorithm) and scoring (called the steatosis, activity and fibrosis [SAF] score) of obesity-associated liver disease.[3]
The difference between the scoring system by Bedossa et al. and Kleiner et al. is only in the activity score (i.e. grading of lobular inflammation and hepatocyte ballooning), whereas the grading of steatosis and staging of fibrosis were identical in the two systems. Furthermore, the grading of hepatocyte ballooning by Bedossa et al. is qualitative in nature, whereas the grading of hepatocyte ballooning by Kleiner et al. is semiquantitative.

	Kleiner et al.	Bedossa et al.
Lobular inflammation	0: No foci 1: <2 foci per 20x 2: 2-4 foci per 20x 3: >4 foci per 20x	0: None 1: ≤2 foci per 20x 2: >2 foci per 20x
Hepatocyte ballooning	0: None 1: Few 2: Many	0: normal hepatocytes with cuboidal shape and pink eosinophilic cytoplasm 1: presence of clusters of hepatocytes with a rounded shape and pale cytoplasm usually reticulated; although shape is different, size is quite similar to that of normal hepatocytes 2: same as grade 1 with some enlarged hepatocytes, at least 2-fold that of normal cells

In the study by Bedossa et al., among the 249 patients with activity (A) score ≥ 2, 230 (92%) had NASH, whereas all patients with A < 2 did not have NASH. Furthermore, there was strong correlation between activity score and serum alanine aminotransferase (ALT) and aspartate aminotransferase (AST) levels. In other words, the activity score provided a more robust histological approach that clearly distinguished most patients with NASH and matched transaminase levels. Moreover, the authors found no significant differences in the ALT and AST levels between patients with normal liver and patients with pure steatosis, supporting the exclusion of steatosis as a marker of activity. The FLIP algorithm and SAF score improved inter-observer variability and have been validated clinically.[4,5] However, data are limited, especially from Asian centres.
Worldwide, many studies have been performed using the system by Kleiner et al. with massive amount of valuable clinical data before the system by Bedossa et al. was introduced. If the NAS can be directly translated into the SAF score, previously collected clinical data can be analyzed using the newer system without needing to look back at the liver biopsies, a task that is enormous and associated with many limitations. The primary objective of this study was to determine whether the grade for lobular inflammation and ballooning in the NAS can be directly translated into the activity score for the SAF score. Secondary objectives included studying the intra-observer and inter-observer agreement for each individual histological component (i.e. steatosis, lobular inflammation, ballooning, and fibrosis) using the two scoring systems, and for the diagnosis of NASH, between pathologists.

Methods

Four pathologists (H. H. L., A. W. C., P. P., and N. R. N. M.) from two Asian centres reviewed a standardized document on grading and staging of the NASH CRN scoring system, following which they independently reviewed 20 digitalized liver biopsy slides and reported them according to the scoring system using a standardized form. Additionally, the pathologist gave an overall diagnosis of whether a case was NASH or not. The reports were finalized and set aside after completion. The process was repeated 2 weeks after the first examination to minimize bias during the second examination.
A further 2 weeks, the pathologists reviewed another standardized document on grading and staging of NAFLD according to the SAF scoring system, following which they repeated the same abovementioned process according to the scoring system.
The digitalized liver biopsy slides were assembled from a library of NAFLD cases from a previous study.[6] The slides were scanned using a digital slide scanner (Pannoramic® MIDI, 3DHISTECH, Budapest, Hungary) and were viewed by all pathologists using a standardized software (CaseViewer 2.1, 3DHISTECH, Budapest, Hungary). The slides were selected to cover the range of grades or stages of the different NAFLD histological components. The pathologists were not involved in the selection of cases and were blinded to the clinical data and the grades or stages of the different histological components of the cases. Each of the cases had hematoxylin and eosin (HE) and Masson trichrome stains and were provided to the pathologists in an external hard disk.
Two of the pathologists (A.W. C. and N. R. N. M.) were senior pathologists, each with 15 years of experience in liver pathology, while each of the two other pathologists (H. H. L. and P. P.) had 3.5 years of experience in liver pathology.

Statistical analysis

Sample size calculation was based on the minimum number of subjects required to study the agreement between pathologists at 0.05 significance level and 80% power, with an assumed disagreement rate between pathologists of 5%. The sample size needed was 18. Data were analyzed using IBM PSS Statistics 27 (IBM, New York, USA).
Intra- and inter-observer agreement was analyzed using Fleiss' kappa, weighted kappa or Cohen kappa, where appropriate.

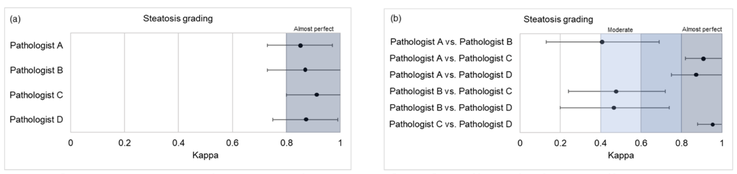
Kappa	Strength of agreement
0-0.19	Slight
0.20-0.39	Fair
0.40-0.59	Moderate
0.60-0.79	Substantial
0.80-1.00	Almost perfect

Result

Patient characteristics

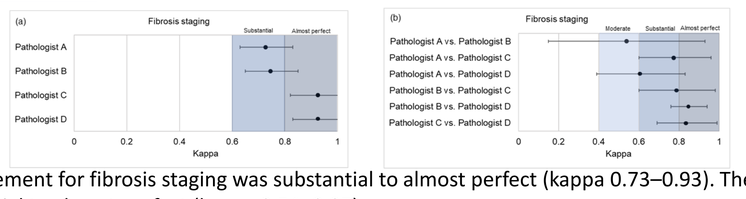
Median age of the study population was 58 (52–63) years old, consisting of 7 men and 13 women. The length of the liver biopsy specimens was 14 (12–15) mm with 6 (4–7) portal tracts. Fourteen patients had NASH while six patients did not have NASH.

Steatosis grading



The intra-observer agreement for steatosis grading was almost perfect for all pathologists (kappa 0.85–0.91). The inter-observer agreement was only moderate between Pathologist B and the other three pathologists (kappa 0.41–0.47). However, the inter-observer agreement was almost perfect among the other three pathologists (kappa 0.88–0.96).

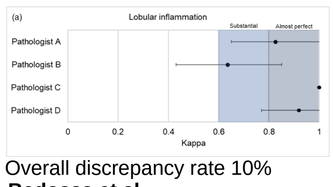
Fibrosis



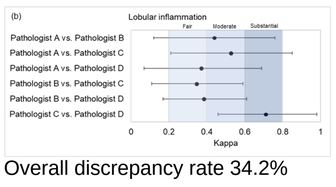
The intra-observer agreement for fibrosis staging was substantial to almost perfect (kappa 0.73–0.93). The inter-observer agreement was substantial to almost perfect (kappa 0.54–0.85).

Lobular inflammation

Kleiner et al.

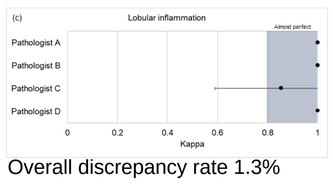


Overall discrepancy rate 10%

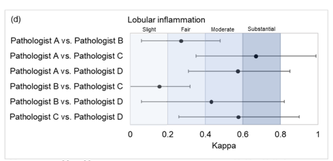


Overall discrepancy rate 34.2%

Bedossa et al.

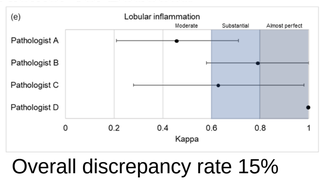


Overall discrepancy rate 1.3%



Overall discrepancy rate 38.3%

Kleiner vs. Bedossa

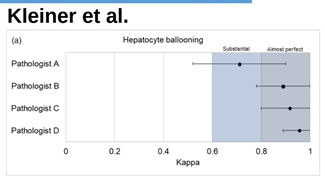


Overall discrepancy rate 15%

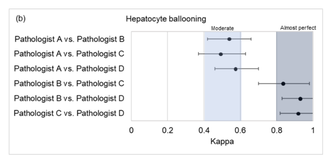
NASH CRN scoring system: The intra-observer agreement for lobular inflammation (Fig. a) was substantial for one pathologist (kappa 0.64), almost perfect for two pathologists (kappa 0.83–0.92) and perfect for one pathologist (kappa 1.00). However, inter-observer agreement (Fig. b) was only fair to moderate (kappa 0.35–0.53), except between Pathologist C and Pathologist D, where the inter-observer agreement was substantial (kappa 0.72).
SAF scoring system: The intra-observer agreement (Fig. c) was almost perfect for one pathologist (kappa 0.86) and perfect for three pathologists (kappa 1.00). However, inter-observer agreement (Fig. d) was only fair to moderate (kappa 0.27–0.58), except between Pathologist B and Pathologist C, where the inter-observer agreement was only slight (kappa 0.13), and between Pathologist A and Pathologist C, where the inter-observer agreement was substantial (kappa 0.67).
The intra-observer agreement for lobular inflammation during examinations using the NASH scoring system and the SAF scoring system (Fig. e) was moderate for one pathologist (kappa 0.46), substantial for two pathologists (kappa 0.63–0.79) and perfect for one pathologist (kappa 1.00).

Hepatocyte Ballooning

Kleiner et al.

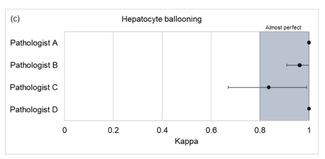


Overall discrepancy rate 12.5%

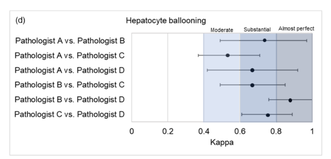


Overall discrepancy rate 35%

Bedossa et al.

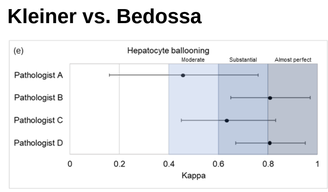


Overall discrepancy rate 5%



Overall discrepancy rate 37.5%

Kleiner vs. Bedossa

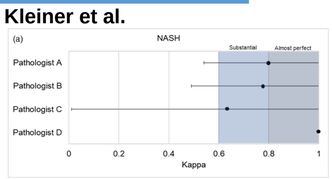


Overall discrepancy rate 33.8%

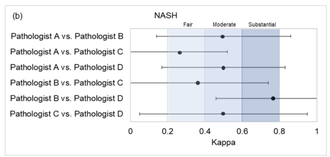
NASH CRN scoring system: The intra-observer agreement for hepatocyte ballooning (Fig. a) was substantial for one pathologist (kappa 0.71) and almost perfect for three pathologists (kappa 0.89–0.96). The inter-observer agreement (Fig. b) was only moderate between Pathologist A and the other three pathologists (kappa 0.50–0.58), but almost perfect between the other three pathologists (kappa 0.84–0.93).
SAF scoring system: The intra-observer agreement (Fig. c) was almost perfect for two pathologists (kappa 0.83–0.97) and perfect for two pathologists (kappa 1.00). The inter-observer agreement (Fig. d) was substantial to almost perfect (kappa 0.67–0.88), except between Pathologist A and Pathologist C, where the inter-observer agreement was moderate (kappa 0.54).
NASH CRN scoring system vs SAF scoring system: The intra-observer agreement (Fig. e) was moderate for one pathologist (kappa 0.46), substantial for one pathologist (kappa 0.64) and almost perfect for two pathologists (kappa 0.81).

NASH

Kleiner et al.

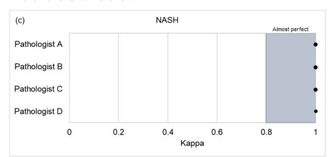


Overall discrepancy rate 6.3%

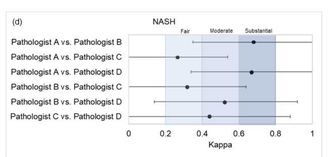


Overall discrepancy rate 35%

Bedossa et al.

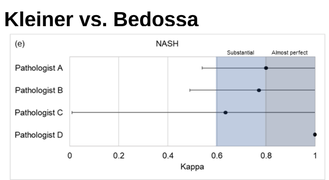


Overall discrepancy rate 0%



Overall discrepancy rate 33.8%

Kleiner vs. Bedossa



Overall discrepancy rate 6.3%

NASH CRN scoring system: The intra-observer agreement for the diagnosis of NASH (Fig. a) was substantial for two pathologists (kappa 0.64–0.78), almost perfect for one pathologist (kappa 0.80) and perfect for one pathologist (kappa 1.00). However, the inter-observer agreement (Fig. b) was only fair to moderate (kappa 0.20–0.50), except between Pathologist B and Pathologist D, where the inter-observer agreement was substantial (kappa 0.77).
SAF scoring system: The intra-observer agreement (Fig. c) was perfect for all pathologists (kappa 1.00). However, the inter-observer agreement (Fig. d) was only slight to fair between Pathologist C and the other three pathologists (kappa 0.15–0.27), and moderate to substantial between the other three pathologists (0.53–0.68).
The intra-observer agreement for the diagnosis of NASH during examinations using the NASH CRN scoring system and the SAF scoring system (Fig. e) was substantial for two pathologists (kappa 0.64–0.78), almost perfect for one pathologist (kappa 0.80) and perfect for one pathologist (kappa 1.00).

Discussion

Several notable observations were made in this study on intra-observer and inter-observer agreement when using the NASH CRN scoring system and the SAF scoring system among four pathologists from two Asian centres active in NAFLD research. Firstly, the intra-observer agreement for lobular inflammation grading was almost perfect (except in one pathologist, who demonstrated substantial intra-observer agreement) regardless of the NASH CRN scoring system or the SAF scoring system. However, there were slightly more discrepancies from grading of either 2 or 3 by the different pathologists due to the additional grade (i.e. grade 3) in the NASH CRN scoring system. On the other hand, the inter-observer agreement for lobular inflammation grading was only fair to moderate regardless of the NASH CRN scoring system or the SAF scoring system. The lower inter-observer agreement may be due to the different methods used by the different pathologists for estimating the number of foci of lobular inflammation, for example, eyeballing, focusing on hot spot, averaging for the entire examination. When comparing the NASH CRN scoring system and the SAF scoring system, although the intra-observer agreement was substantial to almost perfect (except in one pathologist, who demonstrated only moderate intra-observer agreement), the intra-observer discrepancy rate (15%) was higher than the intra-observer discrepancy rate of the individual systems (10% and 1.8% for the NASH CRN scoring system and the SAF scoring system, respectively). These findings suggested that the grades in the NASH CRN scoring system and the SAF scoring system were not interchangeable.
Secondly, the intra-observer agreement for hepatocyte ballooning grading was almost perfect (except in one pathologist, who demonstrated substantial intra-observer agreement), regardless of the NASH CRN scoring system or the SAF scoring system. At one look, the inter-observer agreement for hepatocyte ballooning grading appeared better when using the SAF scoring system than the NASH CRN scoring system. However, the three comparisons showing moderate inter-observer agreement when using the NASH CRN scoring system were likely related to Pathologist A, who also demonstrated lower intra-observer agreement when using the NASH CRN scoring system. We observed that 96.3% (26/27) of the discrepancies between the NASH CRN scoring system and the SAF scoring system were due to an increased score when the SAF scoring system was used. These findings may be explained by the additional morphological description for hepatocyte ballooning grading in the SAF scoring system compared with the semi-quantitative nature for hepatocyte ballooning grading in the NAS system (e.g. a few large and well-formed ballooned hepatocytes may be graded as 1 in the NASH CRN scoring system but as 2 in the SAF scoring system). While this may provide more clarity for the individual pathologist, it still depended on the subjective interpretation by the different pathologists. When comparing the NASH CRN scoring system and the SAF scoring system, although the intra-observer agreement was substantial to almost perfect (except in one pathologist, who demonstrated only moderate intra-observer agreement), the intra-observer discrepancy rate (33.8%) was higher than the intra-observer discrepancy rate of the individual systems (12.5% and 5% for the NASH CRN scoring system and the SAF scoring system, respectively). These findings suggested that the grades for hepatocyte ballooning in the NAS system and the SAF system were not interchangeable.
Thirdly, the intra-observer agreement for NASH diagnosis was better when using the SAF scoring system than when using the NASH CRN scoring system, although the inter-observer agreement varied widely between fair to substantial. When comparing the NASH CRN scoring system and the SAF scoring system, the intra-observer agreement for NASH diagnosis was substantial to almost perfect with an intra-observer discrepancy rate of 6.3% (5/80), which was within the intra-observer discrepancy rate of the individual systems (6.3% and 0% for the NAS system and the SAF system, respectively). Overall, the findings from our study did not support the direct inter-translation between the lobular inflammation grade and hepatocyte ballooning grade of the two systems. Nevertheless, NASH diagnosis using the NAS system appeared comparable to NASH diagnosis using the SAF system.
Fourth, our study provided important multi-center Asian data on intra-observer and inter-observer variability for histopathological examination of liver biopsy specimen, which is lacking in the existing literature. The intra-observer agreement was almost perfect in almost all instances for grading of steatosis, lobular inflammation and hepatocyte ballooning, and substantial to almost perfect for fibrosis staging and for overall diagnosis of NASH. On the other hand, inter-observer agreement varied widely between fair and substantial for lobular inflammation grading and NASH diagnosis, and appeared better for steatosis grading, hepatocyte ballooning grading and fibrosis staging. While the SAF scoring system demonstrated slightly lower discrepancy rates for grading of lobular inflammation and hepatocyte ballooning and for the diagnosis of NASH, the level of agreement as documented by kappa scores did not clearly favor one system over the other. In the end, the individual pathologist's familiarity with and self-consistency using a particular system should probably play the largest role in deciding which system to use. The experience of pathologist did not appear to consistently impact on inter-observer agreement. For example, inter-observer agreement for steatosis grading and lobular inflammation grading was negatively impacted by Pathologist B, who had less experience. However, inter-observer agreement for hepatocyte ballooning grading and fibrosis staging was negatively impacted by Pathologist A, while inter-observer agreement for NASH diagnosis was negatively impacted by Pathologist C, both of whom were more experienced. The use of artificial intelligence and automated quantification may help to improve the inter-observer variability in the reporting of histological components of NAFLD and the overall diagnosis of NASH. [7]
Despite our best effort, this study had several limitations. First, the pathologists were from centres active in NAFLD research, so the findings may not be generalizable to centres with low volume of NAFLD cases. Second, despite the interval of at least 2 weeks between each of rounds of histopathological examination, there is a possibility that the pathologists may be able to recall the scoring of a particular case from earlier rounds of histopathological examination. Third, it was not possible to cover the full range of possible combinations of the different grades or stages of the different histological components. However, the cases were selected to cover all the grades or stages of the different histological components of NAFLD. Fourth, although two pathologists had longer experience, all four pathologists were experienced in liver pathology. The interpretation of the impact (or the lack of impact) of pathologist's experience on intra-observer and inter-observer agreement in this study may be limited by the experience threshold, which may have been exceeded by all four pathologists who have been active in NAFLD research. Finally, the sample size calculation for this study was based on an assumed disagreement rate of 5%. The disagreement rates between pathologists for the different histological components and for the diagnosis of NASH in this study was found to range from 5% to 65%. The minimum number of subjects required would range from 18 if the assumed disagreement rate were 5% to 96 if the assumed disagreement rate were 50%. If the assumed disagreement rate were 65%, the minimum number of subjects required would be 87.

Conclusion

In conclusion, the findings from this study did not support the direct inter-translation between the NASH CRN scoring system and the SAF scoring system. However, the diagnosis of NASH during examinations using the NASH CRN scoring system may be comparable with diagnosis of NASH using the SAF scoring system, vice versa. Furthermore, we confirmed previous findings of inter-observer variability in histopathological examination of NAFLD cases, especially for lobular inflammation. Further research is needed to improve the consistency in grading or staging of histological components of NAFLD.

Reference
1. Kleiner, D. E., Brunt, E. M., Van Natta, M., Behling, C., Contos, M. J., Cummings, O. W., ... & Nonalcoholic Steatohepatitis Clinical Research Network. (2005). Design and validation of a histological scoring system for nonalcoholic fatty liver disease. Hepatology, 41(6), 1313-1321.
2. Sanyal, A. J., Brunt, E. M., Kleiner, D. E., Kowdley, K. V., Chalasani, N., Lavine, J. E., ... & McCullough, A. (2011). Endpoints and clinical trial design for nonalcoholic steatohepatitis.
3. Bedossa, P., Poitou, C., Veyrie, N., Bouillot, J. L., Basdevant, A., Paradis, V., ... & Clement, K. (2012). Histopathological algorithm and scoring system for evaluation of liver lesions in morbidly obese patients. Hepatology, 56(5), 1751-1759.
4. Bedossa, P., & FLIP Pathology Consortium. (2014). Utility and appropriateness of the fatty liver inhibition of progression (FLIP) algorithm and steatosis, activity, and fibrosis (SAF) score in the evaluation of biopsies of nonalcoholic fatty liver disease. Hepatology, 60(2), 565-575.
5. Nascimbeni, F., Bedossa, P., Fedchuk, L., Pais, R., Charlotte, F., Lebray, P., ... & Ratziu, V. (2020). Clinical validation of the FLIP algorithm and the SAF score in patients with non-alcoholic fatty liver disease. Journal of Hepatology, 72(5), 828-838.
6. Harry, S., Lai, L. L., Mustapha, N. R. N., Aziz, Y. F. A., Vijayananthan, A., Rahmat, K., ... & Chan, W. K. (2020). Volumetric liver fat fraction determines grade of steatosis more accurately than controlled attenuation parameter in patients with nonalcoholic fatty liver disease. Clinical Gastroenterology and Hepatology, 18(4), 945-953.
7. Forlano, R., Mullish, B. H., Giannakeas, N., Maurice, J. B., Angkathunyakul, N., Lloyd, J., ... & Manousou, P. (2020). High-throughput, machine learning–based quantification of steatosis, inflammation, ballooning, and fibrosis in biopsies from patients with nonalcoholic fatty liver disease. Clinical Gastroenterology and Hepatology, 18(9), 2081-2090.