Available online at www.sciencedirect.com
ScienceDirect
Annals of Diagnostic Pathology 15 (2011) 19–24
Annals of DIAGNOSTIC PATHOLOGY

Effects of interventions on intra- and interobserver agreement on interpretation of nonalcoholic fatty liver disease histology☆

Samer Gawrieh, MD[a,b,*], Darly M. Knoedler, MD[c], Kia Saeian, MD, MScEpi[a], James R. Wallace, MD, PhD[d], Richard A. Komorowski, MD[c]

[a]Division of Gastroenterology and Hepatology, Medical College of Wisconsin, Milwaukee, WI 53226, USA
[b]Zablocki VA Medical Center, Milwaukee, WI, USA
[c]Department of Pathology, Medical College of Wisconsin, Milwaukee, WI 53226, USA
[d]Department of Surgery, Medical College of Wisconsin, Milwaukee, WI 53226, USA

Abstract Accurate and reproducible interpretation of nonalcoholic fatty liver disease (NAFLD) histology has significant clinical and research-related implications. We evaluated the impact of 2 interventions ([1] review of illustrative histologic images of NAFLD with the study pathologists; [2] use of a scoring sheet with written diagnostic criteria for different NAFLD phenotypes) on intra- and interobserver agreement on interpretation of NAFLD histology. Before and after the interventions, 2 pathologists twice read 65 liver biopsies done for evaluation of suspected NAFLD. The intra- and interobserver agreement was highest on assessment of steatosis and fibrosis. The interventions significantly improved the intraobserver agreement only on assessment of hepatocellular ballooning. The interobserver agreement was only fair on assessment of lobular inflammation, ballooning, and diagnostic classification and did not improve after the interventions. Methods to improve interobserver agreement on assessment of lobular inflammation and ballooning are needed and would likely increase pathologists' agreement on NAFLD diagnostic classification.

Keywords: Intraobserver; Interobserver; Agreement; Steatosis; Steatohepatitis; NAFLD; NASH

1. Introduction

Nonalcoholic fatty liver disease (NAFLD) is the most common liver disease in the United States [1-3]. More than a third of US adults and a tenth of US children have NAFLD [4,5]. The histologic spectrum of the disease ranges from simple steatosis (SS), to steatosis with variable mix of inflammation, hepatocyte injury, and fibrosis (nonalcoholic steatohepatitis [NASH]) [6]. Although patients with SS do not usually develop liver complications, those with NASH are at increased risk for progressive hepatic fibrosis and development of cirrhosis, liver failure, and hepatocellular carcinoma [7-9].

Pathologist's assessment of liver biopsy is currently considered the gold standard for definitive diagnosis of NAFLD, distinction of its histologic phenotypes, and assessment of its severity [6]. Because of their significantly worse hepatic outcome, patients with NASH are generally more closely monitored in clinical practice than those with SS. For the same reason, clinical trials have focused on finding effective therapeutic interventions that halt or reverse the course of hepatic injury in NASH [10,11], with improvement in major NASH histologic features being the most desired outcome measure. Assessment of NAFLD histology, therefore, has significant clinical and research-related implications.

Previous studies described a wide range of agreement among pathologists on diagnostic classification of NAFLD histologic phenotypes (SS, possible NASH, and NASH) and

☆ Disclosures: No conflict of interest to disclose.

* Corresponding author. Division of Gastroenterology and Hepatology, Medical College of Wisconsin, Milwaukee, WI 53226, USA. Tel.: +1 414 955 6850; fax: +1 414 955 6214.
E-mail address: sgawrieh@mcw.edu (S. Gawrieh).

1092-9134/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.anndiagpath.2010.08.001

detection of its cardinal features [12-17]. The aim of this study was to test the hypothesis that intra- and interobserver agreement on interpretation of NAFLD histology can be improved by introduction of an educational intervention and a pathology scoring sheet that includes written diagnostic criteria for different NAFLD phenotypes.

2. Methods

The study protocol was reviewed and approved by the Medical College of Wisconsin's Internal Review Board. Our liver biopsy database was searched for patients who underwent liver biopsy for suspected NAFLD (unexplained elevated liver tests, steatosis on imaging, or cryptogenic cirrhosis). Medical records were reviewed and patients were excluded if they had documented alcohol consumption of more than 20 g/d, evidence of viral hepatitis (positive hepatitis C antibody or hepatitis B surface antigen), or evidence of other chronic liver disease based on positive serologic tests (anti–smooth muscle antibody, antinuclear antibody, antimitochondrial antibody, iron panel, or stains of the biopsy) and supportive histologic features.

A senior (RAK) and a junior (DMK) pathologist semiquantitatively scored the individual histologic features, including steatosis grade (SG), lobular inflammation (LI), hepatocellular ballooning (HB), Mallory's hyaline, and fibrosis stage (FS) according to the scoring system suggested by the NIH NASH Clinical Research Network working group [13]. Portal inflammation (PI), however, was graded according to the Brunt classification to allow broader assessment of its severity [18]. The NAFLD activity score (NAS) was calculated as the sum of SG, LI, and HB scores. The definition of NASH was based on modified criteria used by Dixon et al [19]. Each liver biopsy specimen was classified as SS, possible NASH, NASH, or normal. The diagnosis of SS required presence of macrovesicular steatosis in 5% or more of the hepatocytes. In addition to steatosis, the diagnosis of NASH required the presence of 2 of the following centrilobular criteria: (1) lobular inflammation; (2) hepatocyte ballooning; (3) pericellular fibrosis. Patients with steatosis and only one of the above criteria were diagnosed to have possible NASH. Pathologists were unaware of the prior clinical diagnosis. Each pathologist blindly read and scored the set twice with at least a 2-week interval between each reading. A scoring sheet that included scoring 14 features of NAFLD [13] and only criteria for diagnosis of NASH was used in this phase of the study.

3. Interventions

In the second phase of this study, results of the first phase, illustrative histologic images of NAFLD histologic phenotypes that show the different individual features in different severity and different published diagnostic criteria for NAFLD histologic phenotypes (SS, possible NASH, and NASH) were reviewed with each pathologist using a PowerPoint presentation. Each image was discussed in detail to highlight the characteristics that define each feature and its severity followed by discussion on how to apply our diagnostic criteria to reach a final histologic phenotype. The scoring sheet was modified to include scoring only 8 NAFLD features (Table 1). In addition, the above criteria used for the diagnosis of the 3 different NAFLD phenotypes were included on the new sheet. Each pathologist blindly re-read the same set of biopsies 2 additional times after these 2 interventions with at least a 2-week interval between each reading.

4. Statistical analysis

For comparison of NAFLD feature frequency pre- and postinterventions, a 2-tailed t test was applied.

Simple and weighted κ statistics were calculated using SAS statistical software, version 9.1 (The SAS Institute, Cary, NC). Simple κ was used for dichotomous or nonordered measures and weighted κ was used for those measures with a natural ordering. To ensure square tables for the calculation of κ, observations with very small weight (0.0001) were added to the diagonal cells in the table.

An inverse variance weighted mean of repeated κ values was calculated for both intra- and interobserver κ using methods introduced by Fleiss [20]. This method includes a χ^2 test for the null hypothesis of equal κ's. Comparison of pre- and postintervention κ values was performed using a Z test based on asymptotic normality of the estimated κ statistics for intra- and interrater reliability. These calculations were performed in an Excel spreadsheet.

Because these measures of κ were actually paired measures on the same samples and not independent groups, we investigated the correlation of the pre- and post-κ estimates using bootstrap methods [21]. The variance of the difference in pre- and postintervention κ was estimated by resampling among the 65 observations, using one randomly selected replicate for interobserver κ and both replicates for intraobserver κ. The statistical significance was determined by calculating a normal test statistic using the observed difference and the estimated variance (a bias corrected bootstrap test). We also calculated the Wald statistic using the bootstrap estimate of correlations.

5. Results

5.1. Distribution of NAFLD features and diagnostic classifications

Sixty-five liver biopsies were included in this study. The average biopsy length was 2.17 ± 0.75 cm with only 2 biopsies being less than 1 cm (0.6 and 0.9 cm) long. The median number of biopsy fragments was 2 (range, 1-6). The