Validation of the Prognostic Value of Histologic Scoring Systems in Primary Sclerosing Cholangitis; An International Cohort Study

Elisabeth M G De Vries[1*], Manon de Krijger[1*], Martti Farkkila[2], Johanna Arola[3], Peter Schirmacher[4], Daniel Gotthardt[5], Benjamin Goeppert[4], Palak J. Trivedi[6], Gideon M Hirschfield[6], Henriette Ytting[7], Ben Vainer[8], Henk R van Buuren[9], Katharina Biermann[10], Maren H Harms[9], Olivier Chazouillers[11], Dominique Wendum[12], Astrid Donald Kemgang[11], Roger W Chapman[13, 14], Lai Mun Wang[13,14,], Kate D Williamson[13,14], Annette Gouw[15], Valerie Paradis[16], Christine Sempoux[17], Ulrich Beuers[1], Stefan Hubscher[6,18],Joanne Verheij[19#], Cyriel Ponsioen[1#]

1 Department of Gastroenterology and Hepatology, Academic Medical Center, Amsterdam, the Netherlands

2 2 Helsinki University and Helsinki University Hospital, Department of Gastroenterology and Helsinki University Hospital, Helsinki, Finland

3 Department of Pathology Helsinki University and Helsinki University Hospital, Helsinki, Finland

4 Institute of Pathology, University Hospital Heidelberg, Heidelberg, Germany

5 Department of Gastroenterology and Hepatology, University Hospital Heidelberg, Heidelberg, Germany

6 National Institute for Health Research (NIHR) Birmingham Liver Biomedical Research Unit (BRU), Institute of Immunology and Immunotherapy, University of Birmingham, United Kingdom

7 Department of Hepatology, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark

8 Department of Pathology, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark

9 Department of Gastroenterology and Hepatology, Erasmus University Medical Center, Rotterdam, the Netherlands

10 Department of Pathology, Erasmus University Medical Center, Rotterdam, the Netherlands

11 Department of Hepatology, Hôpital Saint Antoine, Sorbonne Universités, Paris, France

12 Department of Pathology, Hôpital Saint Antoine, Sorbonne Universités, Paris, France

13 Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom

14 Translational Gastroenterology Unit, John Radcliffe Hospital, Oxford, United Kingdom

15 Department of Pathology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

16 Department of Pathology, Beaujon Hospital, Assistance Publique-Hôpitaux de Paris, Paris, France

17 Institut Universitaire de Pathologie-IUP, Lausanne, Switzerland

18 Department of Cellular Pathology, Queen Elizabeth Hospital, Birmingham, United Kingdom

19 Department of Pathology, Academic Medical Center, Amsterdam, the Netherlands

* these authors contributed equally to the manuscript

shared senior authors

Correspondence

C.Y. Ponsioen, MD PhD

Dept. of Gastroenterology and Hepatology

Academic Medical Center

Meibergdreef 9,1105 AZ,

Amsterdam, The Netherlands

Fax nr: 0031206917033, phone nr: 0031205666012

Email: c.y.ponsioen@amc.uva.nl

Electric word count manuscript: 5644

Number of tables: 4

Number of figures: 2

Authors' contribution

CP designed the study and supervised the project. EdV collected patient data and histologic material, performed the statistical analyses, interpretation of the data and prepared the first draft of the manuscript. MdK coordinated the interobserver study, performed the statistical analyses of the interobserver study, and adjusted the manuscript accordingly. JV contributed to the design of the study and review of the histological material. MF, DG, GH, HY, HvB, OC and RC identified PSC patients that were included in the study. JA, PS, SH, BV, KB, DW and LMW collected and reviewed the histological material. MF, BG, PT, HY, HvB, MH, ADK, and KW collected clinical patient data. JV, SH, JA, DW, LMW, CS, AG and VP scored the

biopsies. All authors reviewed the manuscript for critical content, and approved the final version.

Key words: liver histology; risk stratification;; Nakanuma scoring system; prognosis; interobserver variability, sclerosing cholangitis

Conflicts of Interest and Source of Funding:

PJT is funded by a Wellcome Trust Clinical Research Fellowship. GMH and PJT receive funding through the NIHR BRU. The other authors disclosed no financial relationship relevant to this publication

List of abbreviations:

PSC = primary sclerosing cholangitis

AIH = autoimmune hepatitis

PBC = primary biliary cholangitis

CCA = cholangiocarcinoma

HR = hazard ratio

CI = confidence interval

IBD = inflammatory bowel disease

IQR = inter quartile range

ICC = intra-class correlation coefficient

κ = kappa index

.

Abstract (word limit 275)

Histologic scoring systems specific for primary sclerosing cholangitis (PSC) are not validated. We recently determined the applicability and prognostic value of three histological scoring systems in a single cohort of PSC patients. The aim of this study was to validate their prognostic utility and reproducibility, using a well-characterized multicenter PSC cohort.

Liver biopsies from PSC patients were collected across 7 European centers. Histologic scoring was performed using the Nakanuma, Ishak, and Ludwig scoring systems. Biopsies were independently scored by six liver pathologists for interobserver agreement. The prognostic value of clinical, biochemical, and all three histologic scoring systems on predicting composite endpoint 1: PSC-related death and liver transplantation, 2: liver transplantation, and 3: liver related events, were assessed using uni- and multivariable Cox proportional hazards modelling.

119 PSC patients were identified, The median follow-up was 142 months. During follow-up, 31 patients died (20 PSC-related deaths), 31 underwent liver transplantation, and 35 experienced one or more liver-related events. All three staging systems were independent predictors of endpoints 2 and 3; Nakanuma HR 3.16 (95%CI 1.49-6.68), HR 2.05 (95%CI 1.17-3.57); Ishak: HR 1.55 (95%CI 1.10-2.18), HR 1.43 (95%CI 1.10-1.85); Ludwig: HR 2.62 (95%CI 1.19-5.80), HR 2.06 (95%CI 1.09-3.89), respectively. Only the Nakanuma staging system was independently associated with endpoint 1: HR 2.14 (95%CI 1.22-3.77). The inter-observer agreement was moderate for Nakanuma stage (κ=0.56), and substantial for Nakanuma component fibrosis (κ=0.67), Ishak stage (κ=0.64) and Ludwig stage (κ=0.62). ***Conclusion:***

We confirm the independent prognostic value and reproducibility of predicting disease progression in PSC by the Nakanuma, Ishak and Ludwig staging systems. The Nakanuma staging system – that incorporates features of chronic biliary disease – showed to hold the strongest predictive value..

Word count: 275

Primary sclerosing cholangitis (PSC) is an idiopathic cholestatic liver disease, affecting the intra- and/or extrahepatic bile ducts(1). Chronic inflammation induces formation of fibrotic strictures, interspersed with dilatations, resulting in a typical ‘beaded’ appearance when visualizing the PSC bile ducts on a cholangiogram.

Liver biopsy is no longer routinely performed to establish a diagnosis of PSC (2, 3). However, there are still some circumstances in which the assessment of

liver histology is important diagnostically, such as the suspicion of small duct PSC or PSC/auto-immune hepatitis (AIH) overlap syndrome. Besides a diagnostic purpose, histological assessment of disease severity may also be useful to stratify clinical outcomes in autoimmune and cholestatic liver diseases; particularly to determine inclusion criteria for clinical trials and to evaluate the efficacy of therapeutic interventions(4).

Histologically, the characteristic fibrosing duct lesions of PSC mainly involve medium-sized (septal) ducts and are thus infrequently seen in needle biopsy specimens(5). Varying degrees of portal and periportal inflammation are commonly present, particularly during the early stages of PSC, and may be associated with bile duct inflammation. As the disease progresses, there is loss of small and medium-sized ducts (ductopenia) and the development of secondary changes related to chronic cholestasis including ductular reaction and the accumulation of copper binding protein in periportal hepatocytes(6). These changes are associated with the development of progressive periportal fibrosis, ultimately leading to the development of biliary cirrhosis(6).

In the absence of a validated histologic scoring system specifically designed to assess disease severity in PSC, we recently assessed the applicability and prognostic value of three histologic scoring systems, designed primarily to assess disease severity in chronic hepatitis (Ishak system) and primary biliary cholangitis (PBC) (Ludwig and Nakanuma systems), across a group of patients with PSC(7). Staging by each of these three scoring systems was shown to be strongly associated with disease progression in PSC(7). However, this was a single center study with a relatively small sample size, and only included biopsies taken at time of diagnosis. Furthermore, although all three histologic scoring systems are being applied in

clinical practice, their reproducibility for PSC is unknown. With the present study we aimed to validate the prognostic value and reproducibility of liver histology, scored by the three previously evaluated scoring systems, using several well-characterized external cohorts of PSC patients.

PATIENTS AND METHODS

Patients and tissue preparation

For this multicenter cohort study, PSC patients from 7 centers in 6 European countries were included: Helsinki University Hospital, Finland; University Hospital Heidelberg, Germany; Queen Elizabeth Hospital, United Kingdom; Rigshospitalet, Denmark; Erasmus University Medical Center, the Netherlands; Hôpital Saint

Antoine, France; John Radcliffe Hospital, United Kingdom. PSC diagnosis was established according to the European Association for the Study of the Liver guidelines(2). A diagnosis of PSC/AIH-overlap syndrome was made in keeping with the expertise of the contributing center.

To ensure adequate follow-up time, PSC patients with liver biopsies taken at least 10 years ago were included and biopsy material requested from the archives by liver pathologists in the participating centers. Biopsies taken between 1980-2006 were studied. The original haematoxylin and eosin, connective tissue, and orcein stained sections were collected and the quality of the staining was checked. In those cases where staining had faded over time, additional stainings were carried out. In the event that the pathology laboratory of the participating center did not perform orcein histochemistry, additional orcein staining was performed by the laboratory of SH in Birmingham. Clinical data were retrospectively collected from patient charts considering patient demographics, medication use, liver biochemistry at time of diagnosis (+/- 1 month), liver-related events (variceal bleeding, ascites, splenomegaly and hepatic encephalopathy), and the occurrence of cholangiocarcinoma, liver transplantation, or death.

Medical Research Involving Human Subjects Act (WMO) did not apply to this study. Participation by individual centers was approved at a local level. Patient data were treated anonymously.

Histologic evaluation

Two expert liver pathologists (JV, SH) scored the liver biopsies in tandem using a multihead microscope, with the intention to reach consensus. This was considered

the reference diagnosis for evaluation of prognostic value. Liver biopsy sections containing less than 6 evaluable portal tracts were excluded from the study.

The criteria for scoring according to the Nakanuma, Ishak and Ludwig scoring systems have been described previously(8-10). (Supplementary Table 1,2) To summarize, both Nakanuma and Ishak scoring systems evaluate disease severity in terms of the histologic grade. The Nakanuma grading system includes degree of cholangitis activity (score 0-3), and hepatitis activity (score 0-3)(9), while the Ishak grading system evaluates degree of interface hepatitis (score 0-4), confluent necrosis (score 0-6), lobular inflammation (score 0-4) and portal inflammation (score 0-4) (8) (Supplementary Table 1).

All three histologic scoring systems evaluate disease progression in terms of stage, by scoring the degree of fibrosis; Nakanuma score 0-3, Ishak score 0-6; Ludwig score 0-4 (8-10). The Nakanuma scoring system in addition scores the degree of bile duct loss (0-3) and deposition of orcein positive granules (representing copper binding proteins) (0-3). The final Nakanuma stage is obtained from the total score of these three features: stage I (no or minimal progression) corresponds to a score of 0, stage II (mild progression) corresponds to a score of 1-3, stage III (moderate progression) corresponds to a score of 4-6 and stage IV (advanced progression) corresponds to a score of 7-9(9). (Supplementary Table 2.)

Interobserver analysis

A panel of 6 international expert liver pathologists (JA, DW, LMW, AG, CS and VP) attended a tutorial of Nakanuma scoring contents and its application for PSC. During a consensus reading, 12 separate cases (7) were discussed in the panel using a multihead microscope, in order to clarify uncertainties. Interobserver variability was

determined based on the scoring of 76 biopsies on original glass slides (H&E, connective tissue and orcein) from the present cohort. The selection of these biopsies was based on the availability and quality of the biopsies and stainings. All pathologists individually scored all biopsies and no clinical or laboratory data were available to the observers.

Endpoints

Several endpoints were assessed for the time-to-event analyses of association with endpoints. Endpoint 1 was transplant-free survival, defined as a composite of PSC-related death (death from end-stage liver disease, death from liver surgery, death from cholangiosepsis and death from cholangiocarcinoma (CCA)), and liver transplantation. Although the risk of developing CCA in PSC is up to 20% increased(11), the pathophysiological mechanism leading to CCA in PSC is yet to be clarified, and may partly depend on the duration and intensity of the chronic inflammation, and to genetic predisposition of the host. It is uncertain if death from CCA lies within a biological pathway similar to disease progression in PSC, and hence death from CCA may not be heralded by liver histology at time of diagnosis. To accommodate for this uncertainty, endpoint 2: liver transplantation, was introduced, which focuses on disease progression only. Endpoint 3, the occurrence of liver related events at follow-up, was defined as the occurrence of gastro-esophageal variceal bleeding, development of ascites, splenomegaly(12) or hepatic encephalopathy. Ascites and splenomegaly were assessed during yearly surveillance imaging - either abdominal ultrasonography or magnetic resonance cholangiopancreatography. Splenomegaly was defined as a spleen length of more than 120 mm.

Statistical Analysis

Patient characteristics and laboratory values were expressed as mean ± standard deviation, and median and interquartile range when having a symmetrical and skewed distribution, respectively. Dichotomous variables were expressed as percentage (%) of the cohort. Reference values of biochemical variables may vary between hospitals and over time, therefore biochemical variables were expressed as ratio of upper limit of normal, or lower limit of normal.

Associations of grading and staging with the different clinical endpoints were estimated using Kaplan Meier curve and log-rank test. The potential prognostic value on predicting outcome of clinical, biochemical, and histological variables was assessed using univariable Cox proportional hazards model. Potential multicollinearity between the different staging systems, as well as Mayo risk score and its laboratory components aspartate aminotransferase, albumin, and bilirubin were checked by Spearman's rank correlation coefficient. Using stepwise multivariable Cox proportional hazards regression, the independent predictive value of the histologic scoring systems – entered as continuous variables – was evaluated, from which the hazard ratio (HR), and corresponding 95% confidence interval (CI) could be calculated. The criterion for retaining predictors was a p-value <0.05.

Interobserver agreement was measured using the multireader linear weighted Light's kappa index (κ), which adjusts for the chance-expected agreement. For continuous variables, additionally an intra class correlation coefficient (ICC) was calculated. Concordance was qualified according to the arbitrary scale of Landis and Koch, considering a value <0.01 as poor, values between 0.01-0.20 as slight, values

between 0.21-0.40 as fair, values between 0.41-0.60 as moderate, values between 0.61-0.80 as substantial and values >0.81 as almost perfect agreement(13).

Statistical analyses were performed using SPSS version 22.0 software (SPSS, Chicago, IL). A p-value of <0.05 was considered statistically significant.

RESULTS

Patient and biopsy characteristics

We received a total of 154 liver biopsies, of which 17 biopsies were excluded due to insufficient amount of portal tracts, poor quality of the staining or an accidental case of sending the wrong tissue biopsy. In 12 PSC cases, more than one biopsy was received, of which the first taken biopsy was included in the final cohort.

Table 1 summarizes the patient characteristics. A total of 119 patients (corresponding to 119 liver biopsies) were included in the final cohort, of which 81

(68%) were male, 38 (32%) female. The majority of 101 (85%) patients were diagnosed with large duct PSC, 2 (2%) small duct PSC, and 16 (13%) patients had PSC/AIH-overlap syndrome. Concomitant inflammatory bowel disease (IBD) was present in 80 (67%) patients, the majority (68 (57%)) suffered from ulcerative colitis. There was a median follow-up from PSC diagnosis until reaching an endpoint or date of last follow up, of 142 months (inter quartile range (IQR) 98-188).

Biopsies had a median of 17 (IQR 11-24) portal tracts, and a median biopsy length of 15 mm (IQR 11-20). The median disease duration at time of biopsy was 0 (range 0-38) months. Sixteen (13%) biopsies showed concentric periductal fibrosis, characteristic for PSC diagnosis.

Patient outcome

During follow-up, 31 (26%) patients died, 20/31 (65%) of deaths were PSC related. In that subgroup, the median time from PSC diagnosis until PSC related death was 152 months (IQR 92-223). 31 (26%) patients underwent liver transplantation, which they underwent within a median time of 79 months (IQR 17-112) from PSC diagnosis. A total of 11 (9%) patients were diagnosed with CCA. For those patients the median time from diagnosis until developing CCA was 146 months (IQR 62-224); 8 (73%) patients died.

A total of 35 (29%) patients experienced one of more liver related events within a median of 85 months (IQR 14-132) after PSC diagnosis. 7 (6%) patients had a variceal bleed. Ascites occurred in 13 (11%) patients, 18 (15%) patients had an enlarged spleen, and 5 (4%) patients developed hepatic encephalopathy.

Distribution of grade and stage

Evaluation of grade and stage according to the three histologic scoring systems was reached in consensus in 100% of cases by the two in tandem scoring pathologists.

Nakanuma grading components cholangitis activity and hepatitis activity were absent in the vast majority of cases, resulting in a median grade of 0 (range 0-3). The Ishak grading component interface hepatitis was distributed with a median of 1 (range 0-3), while confluent necrosis was only present in 4 cases, resulting in a median score of 0 (range 0-6). Degree of lobular inflammation and portal inflammation as scored by Ishak showed a median of 1 (range 0-4), and 1 (range 0-3), respectively. Figures 1 A and B illustrate the total Nakanuma and Ishak grades. The distribution of Nakanuma and Ishak grading components is illustrated in Supplementary Figure 1.

All three histologic components (fibrosis, bile duct loss, and orcein positive granules) staged by Nakanuma were distributed with a median of 1 (range 0-3), resulting in a median total Nakanuma stage of 2 (range 1-4). (Figure 1C; Supplementary Figure. 1) The distribution of degree of fibrosis, as scored according Ishak and Ludwig are shown in Figure 1D, E: median 2 (range 0-6) and 2 (range 0-4), respectively.

Association between histologic grade and stage and clinical outcome.

We found a significant association between the Nakanuma stage, Ishak stage and Ludwig stage and endpoints 1 and 2,; the Ishak and Ludwig staging systems were also significantly associated with endpoint 3 (Figure 2A-C; supplementary Figure A-F). Table 2 shows the results of the univariable Cox proportional hazard analyses, which were used to analyze potential associations between clinical, biochemical and

histological variables and outcome in PSC - as defined by the three different endpoints.

The Ishak and Ludwig staging systems were highly correlated, with a correlation coefficient of 0.93, p<0.001. This multicolinearity precluded the analyses of all three staging system in one multivariable cox model. Therefore, using stepwise regression, the independence of each individual staging system was assessed by entering each staging system separately. In addition, there was a significant correlation between Mayo risk score and aspartate aminotransferase (correlation coefficient 0.665, p <0.001), Mayo risk score and bilirubin (correlation coefficient 0.685, p <0.001), and Mayo risk score and albumin (correlation coefficient -0.647, p <0.001). Therefore, the Mayo risk score was not included in multivariable regression analyses.

Endpoint 1. Composite PSC-related death and liver transplantation

Nakanuma grade, Nakanuma stage, Ishak stage, Ludwig stage, center of inclusion, total bilirubin, albumin, platelet count, and Mayo risk score were predictors of endpoint 1 in univariable analysis. (Table 2). Separate multivariable analyses for the three different staging systems showed that the Nakanuma grading system was inversely associated with endpoint 1 (corrected for Ishak stage: HR 0.55 (95%CI 0.31-0.99), p=0.046; corrected for Ludwig stage: HR 0.54 (95%CI 0.30-0.98), p=0.041). Each incremental grade corresponds to a 55% and 54% decreased risk, respectively, of reaching an endpoint. In contrast, the Nakanuma staging system was positively associated with endpoint 1 (HR 2.14 (95%CI 1.22-3.77), p=0.008). (Table 3A)

Endpoint 2. Liver transplantation

The same predictors of endpoint 1 were shown to be associated with endpoint 2 in univariable Cox regression analysis (Table 2A). When correcting for potential interacting variables in multivariable analyses, the Nakanuma, Ishak, and Ludwig staging systems were all independently associated with liver transplantation, HR 3.16 (95% CI 1.49-6.68), p=0.003; HR 1.55 (95% CI 1.10-2.18), p=0.013; HR 2.62 (95% CI 1.19-5.80), p=0.017, respectively. The Nakanuma grading system was inversely associated with liver transplantation in the multivariable analysis correcting for the Ishak and Ludwig staging systems, HR 0.25 (95% CI 0.07-0.85), p= 0.027; HR 0.25 (95% CI 0.07-0.84), p=0.025, respectively. (Table 3B)

Endpoint 3. Liver related events

The variables Nakanuma stage, Ishak stage and Ludwig stage, age at PSC diagnosis, co-existing IBD, center of inclusion, total bilirubin, and Mayo risk score were significant predictors for the occurrence of liver related events in univariable analysis (Table 2.). The Nakanuma, Ishak and Ludwig staging systems were all shown to be independently associated with the occurrence of liver related events in multivariable analyses: HR 2.05 (95% CI 1.17-3.57), p 0.012; HR 1.43 (95% CI 1.10-1.85) p=0.007; HR 2.06 (95% CI 1.09-3.89), p = 0.027. (Table 3C)

Multivariable analysis Nakanuma staging components

When analyzing the prognostic value of the individual Nakanuma staging components to predict endpoints 1, 2 and 3, orcein deposition was shown to be the strongest independent predictor: HR 1.59 (95% CI 1.16-2.18), p=0.004; HR 1.76 (95% CI 1.19-2.61), p=0.005; HR 1.40 (95% CI 0.95-2.06), p= 0.088, respectively. The Nakanuma component liver fibrosis was an independent predictor of endpoint 2

only (HR 2.25 (95% CI 1.08-4.68), p=0.030, while the Nakanuma component bile duct loss was never associated with outcome. (Table 4).

Interobserver agreement

A total of 76 biopsies were scored by each of the 6 participating pathologists. One case was excluded due to a difference in cutting depth between stainings within the same case, which caused a variation in staging and grading among different slides. In two cases, no orcein staining was available. They were therefore excluded from the analysis for orcein positive granules and Nakanuma total stage.

The inter-observer agreement for Nakanuma grade and stage, Ishak grade and stage and Ludwig stage is shown in table 5. Interobserver agreement was moderate for Nakanuma stage (κ=0.56), and substantial for both Ishak and Ludwig stage (κ=0.64 and κ=0.62 respectively). Interobserver agreement was highest for the fibrosis component of Nakanuma staging, compared with both bile duct loss and deposition of orcein positive granules (κ=0.67, κ=0.47 and κ=0.58 respectively). For Nakanuma grade cholangitis activity, agreement was fair (κ=0.35), hepatitis activity showed moderate agreement (κ=0.55). Ishak grade using 4 different categories showed moderate agreement (κ=0.52), with an intraclass coefficient of 0.745 (95%CI 0.67-0.81).

DISCUSSION

This study confirms the prognostic value and reproducibility of staging disease progression using the Nakanuma, Ishak, and Ludwig staging systems.

We demonstrated that all three staging systems were predictive of liver transplantation and development of liver-related events. In contrast, the Ishak and Ludwig scoring systems were not independently predictive of endpoint 1 (composite PSC related death and liver transplantation). The inclusion of death from CCA in endpoint 1– which may not be predictable by histology – could have contributed to this difference. Additional univariable analysis showed that, indeed, neither grading nor staging liver histology were of prognostic importance for the development of CCA. (Supplementary Table 3.)

In accordance with the previous study on the prognostic value of histology in PSC(ref)., we found that the Nakanuma staging system was the strongest histological predictor of long-term clinical outcomes(7). This is in concordance with study results in PBC, where the Nakanuma staging system was repeatedly shown to be superior to the classical Ludwig and Scheuer staging systems in terms of predicting clinical outcome(14, 15). These observations may be explained by the inclusion of three staging components that reflect histopathological progression of liver injury in vanishing bile duct diseases(15). Of the individual staging components, orcein deposition was the most valuable histologic predictor. In addition, the degree of orcein deposition was strongly correlated with alkaline phosphatase – an important prognostic biochemical marker in PSC(16). (Supplementary Table 4) These results are in concordance with the recent findings in PBC of both Kakuda et al. and Chan et al.. They demonstrated that the copper binding protein score was the most powerful histologic prognostic Nakanuma staging component(14, 15). The deposition of orcein positive granules is a sensitive marker of chronic cholestasis in the liver (17), most likely related to the inability to excrete copper associated protein in bile due to bile duct loss. The accumulation of copper-associated protein in periportal hepatocytes is frequently associated with other morphological changes related to the accumulation of toxic bile acids, such as hepatocyte ballooning and Mallory-Denk body formation (“cholate stasis”), and may thus be an important marker of disease progression in PSC(5). The inter-observer variability assessment of orcein positive granules showed moderate agreement (κ=0.58), which is comparable with the agreement observed in the study by Nakanuma et al. (κ=0.41) (9). Although bile duct loss might theoretically be a more sensitive marker of disease progression in ductopenic diseases such as PSC, loss of bile ducts is often patchy in distribution and may thus

be subject to problems with sampling variability in liver biopsy specimens. The absence of an association between degree of bile duct loss and clinical outcomes was thereforeunsurprising, and similar results were found in a recent study by the Vries et al(7). Of all three Nakanuma staging components, bile duct loss showed the lowest agreement (κ=0.47). In a previous study by Wendum et al., interobserver agreement of bile duct loss in liver biopsies of PBC patients increased from moderate to substantial by adding a keratin 7 immunostaining(18). We chose to use orcein staining for the evaluation of bile duct loss in the present study, in order to obtain the best replicate of the original scoring system by Nakanuma et al.,. Adding a keratin 7 staining may, however, improve performance. Unfortunately, our study was insufficiently powered to assess the independent prognostic value of the individual Nakanuma staging components in a multivariable model, which also included clinical and biochemical parameters.

Similar to what has been demonstrated in PBC, we found no association between Nakanuma grade and liver related events(14, 15). There was, however, an unexpected negative association between Nakanuma grade and the composite endpoint 1 and endpoint 2 - a higher Nakanuma grade being associated with a decreased chance on reaching these endpoints. This difference could not be explained by the inclusion of PSC/AIH-patients, who would be expected to have relatively high scores for the degree of hepatitis activity, since the prognostic value of Nakanuma grade was driven by the cholangitis activity component. (Table 2) We checked if Nakanuma grade could reflect an early (inflammatory) disease stage and a corresponding better prognosis, with less fibrosis by assessing the association between cholangitis activity and degree of fibrosis, but no such association was found. (supplementary Table 5.) Overall, cholangitis activity and hepatitis activity are

not prominent features in PSC, and were present in only 27% versus 14% of cases, respectively. The resulting limited sample size per grade is important to take into consideration, and the negative association between Nakanuma grade and outcome should therefore be interpreted with caution. Compared with the slight agreement for cholangitis activity and hepatitis activity observed in PBC . (κ=0.110, κ=0.197 respectively), agreement in the present study - yielding fair and moderate agreement, respectively (Table 5) - was much higher. The low agreement by Nakanuma et al. was thought to be due to differences in assessment processes in different institutions(9). The consensus reading that took place before scoring may have added to the improved agreement within the present study. Furthermore, the majority of readers in the panel in Nakanuma's study consisted of non-expert liver pathologists.

The proportion of PSC/AIH-overlap patients (13%) included in the present study was relatively high when compared with the 4-7% reported in large epidemiologic studies(11, 19). PSC/AIH-overlap may be regarded as a distinct disease entity, with an altered prognosis(20, 21). A sensitivity analysis in which PSC/AIH- overlap patients were excluded was performed to assess the potential impact of PSC/AIH-overlap on the prognostic value of the histologic scoring systems. No differences were found (data not shown). Moreover, the independent association of all three histologic staging systems with outcome confirms the applicability of liver histology as a risk stratifier in PSC. Carbone at al. recently reported a similar independent association between degree of fibrosis and long-term outcome in PBC(22). Through risk stratification (into low/high risk for poor outcome) patient subgroups that potentially have the greatest benefit from inclusion and treatment in clinical trials may be identified using their initial liver biopsy(4).

The patchy distribution of the PSC liver carries the risk of sampling variability in PSC biopsies(23, 24). In addition, the use of liver biopsy is hampered by its invasive nature, and interpretative inconsistencies. The correlation of histology with clinical outcomes in the present study suggests that sampling variability when it comes to staging is not a significant issue in the context of using liver biopsy findings as surrogate endpoints in clinical trials. Furthermore, interobserver reproducibility regarding staging was shown to be moderate; disagreement of more than one stage was not observed. Repeated biopsies were present in a small subset of PSC cases; n= 12, of which orcein staining was missing in 1 case. A sub analysis of the remaining 11 cases showed progression of Nakanuma stage in 5 cases, steady status in 5 cases, and regression in 1 case (data not shown). This result supports the ability of the Nakanuma staging system to study histologic change over time – despite the risk of sampling variability.

The development of liver fibrosis and the end stage of liver cirrhosis were long thought to be unidirectional and irreversible. However, increased understanding of the pathophysiology of hepatic fibrogenesis – especially in the field of chronic viral hepatitis and auto-immune hepatitis – has revealed that this is a dynamic process, which can be prevented or reversed by either eliminating the etiologic agent, or intervening with the pathogenic mechanism(25-27). Further research is necessary to assess if treatment effects could be measured by a change in histologic stage. If this would be the case, this would warrant further study as a potential surrogate endpoint in clinical trials.

Although the Ludwig staging system is the most commonly used staging system in PSC, to our knowledge, no inter-observer studies have been performed to assess inter-rater variability in scoring of PSC biopsies. For other liver diseases,

interrater reliability using Ludwig's and Ishak' staging systems showed fair and substantial agreement (κ=0.32 in PBC and κ=0.61 in HCV respectively)(18, 28), whereas both staging systems showed substantial agreement in the present study. Since PSC is a rare disease, often requiring tertiary care, only expert liver pathologists were invited to participate in our interobserver study. which may explain these differences. Only two of the grading criteria (Nakanuma's cholangitis activity and Ishak's confluent necrosis) showed fair agreement, and more consensus in these scoring components would be desirable. All other scoring components showed moderate or substantial agreement, implicating that scoring liver biopsies by the Nakanuma, Ishak and Ludwig scoring systems is reproducible in PSC - especially with regards to staging.

Promising non-invasive modalities such as the enhanced liver fibrosis score and transient elastography may be a future substitute for the measure of fibrosis(29, 30). However, the value of liver histology in PSC is not restricted to staging liver fibrosis, and the important additional prognostic value of staging the degree of orcein deposition currently cannot be replaced by non-invasive tools. An additional advantage of liver histology is that it measures disease progression within the affected organ itself and has the potential advantage of construct validity. Furthermore, it allows for mechanistic studies to investigate the effect of an investigational drug(4, 31).

This study has several limitations. Firstly, since not all PSC patients routinely undergo liver biopsy, and since this was a retrospective accrual of cases, selection bias of patients cannot be excluded. Secondly, the development of liver-related events has not been systematically assessed, but relied on clinician reported incidence that may yield an underestimation. Lastly, since the recently presented

novel prognostic model for PSC (de Vries et al. unpublished data) – including biochemistry values at time of diagnosis – showed adequate performance in predicting prognosis in PSC, we decided to use available biochemistry values at time of diagnosis for the multivariable regression analyses. However, biochemistry results at time of diagnosis may differ from the results at time of biopsy since not all biopsies were taken at the time of diagnosis. A potential limitation of the liver histologic staging systems overall is that they may have no prognostic value for development of biliary dysplasia or CCA, which is a significant cause of morbidity and death of PSC patients.

In conclusion, although the inter-observer agreement on the histologic staging systems according to Nakanuma, Ishak, and Ludwig were only moderate and low for cholangitis activity, the staging systems were confirmed to be independent predictors of long-term outcome in PSC, in this multicenter cohort. The Nakanuma staging system –incorporating features of chronic biliary disease – was shown to hold the strongest predictive value and reproducibility was comparable for all three staging systems. Therefore, the Nakanuma staging systemmay be considered the preferred system for scoring PSC liver biopsies. Histologic scoring may be a useful tool for risk stratification and could be explored as surrogate endpoint in clinical trials in PSC..

ACKNOWLEDGEMENTS

REFERENCES

1. Hirschfield GM, Karlsen TH, Lindor KD, Adams DH. Primary sclerosing cholangitis. Lancet 2013;382:1587-1599.

2. European Association for the Study of the L. EASL Clinical Practice Guidelines: management of cholestatic liver diseases. J Hepatol 2009;51:237-267.

3. Chapman R, Fevery J, Kalloo A, Nagorney DM, Boberg KM, Shneider B, Gores GJ, et al. Diagnosis and management of primary sclerosing cholangitis. Hepatology 2010;51:660-678.

4. Trivedi PJ, Corpechot C, Pares A, Hirschfield GM. Risk stratification in autoimmune cholestatic liver diseases: Opportunities for clinicians and trialists. Hepatology 2016;63:644-659.

5. Nakanuma Y, Zen Y, Portmann B. Diseases of the bile ducts. In. Burt AD, Portmann BC, Ferrell LD, eds. MacSween's pathology of the liver. 6th ed. Edinburgh: Churchill Livingstone.; 2012.

6. Portmann B, Zen Y. Inflammatory disease of the bile ducts-cholangiopathies: liver biopsy challenge and clinicopathological correlation. Histopathology 2012;60:236-248.

7. de Vries EM, Verheij J, Hubscher SG, Leeflang MM, Boonstra K, Beuers U, Ponsioen CY. Applicability and prognostic value of histologic scoring systems in primary sclerosing cholangitis. J Hepatol 2015;63:1212-1219.

8. Ishak K, Baptista A, Bianchi L, Callea F, De Groote J, Gudat F, Denk H, et al. Histological grading and staging of chronic hepatitis. J Hepatol 1995;22:696-699.

9. Nakanuma Y, Zen Y, Harada K, Sasaki M, Nonomura A, Uehara T, Sano K, et al. Application of a new histological staging and grading system for primary biliary

cirrhosis to liver biopsy specimens: Interobserver agreement. Pathol Int 2010;60:167-174.

10. Ludwig J, Dickson ER, McDonald GS. Staging of chronic nonsuppurative destructive cholangitis (syndrome of primary biliary cirrhosis). Virchows Arch A Pathol Anat Histol 1978;379:103-112.

11. Boonstra K, Weersma RK, van Erpecum KJ, Rauws EA, Spanier BW, Poen AC, van Nieuwkerk KM, et al. Population-based epidemiology, malignancy risk, and outcome of primary sclerosing cholangitis. Hepatology 2013;58:2045-2055.

12. Ehlken H, Wroblewski R, Corpechot C, Arrive L, Lezius S, Hartl J, Denzer UW, et al. Spleen size for the prediction of clinical outcome in patients with primary sclerosing cholangitis. Gut 2016;65:1230-1232.

13. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-174.

14. Kakuda Y, Harada K, Sawada-Kitamura S, Ikeda H, Sato Y, Sasaki M, Okafuji H, et al. Evaluation of a new histologic staging and grading system for primary biliary cirrhosis in comparison with classical systems. Hum Pathol 2013;44:1107-1117.

15. Chan AW, Chan RC, Wong GL, Wong VW, Choi PC, Chan HL, To KF. Evaluation of histological staging systems for primary biliary cirrhosis: correlation with clinical and biochemical factors and significance of pathological parameters in prognostication. Histopathology 2014;65:174-186.

16. Williamson KD, Chapman RW. Editorial: further evidence for the role of serum alkaline phosphatase as a useful surrogate marker of prognosis in PSC. Aliment Pharmacol Ther 2015;41:149-151.

17. Nakanuma Y, Karino T, Ohta G. Orcein positive granules in the hepatocytes in chronic intrahepatic cholestasis. Morphological, histochemical and electron X-ray microanalytical examination. Virchows Arch A Pathol Anat Histol 1979;382:21-30.
18. Wendum D, Boelle PY, Bedossa P, Zafrani ES, Charlotte F, Saint-Paul MC, Michalak S, et al. Primary biliary cirrhosis: proposal for a new simple histological scoring system. Liver Int 2015;35:652-659.
19. Tischendorf JJ, Hecker H, Kruger M, Manns MP, Meier PN. Characterization, outcome, and prognosis in 273 patients with primary sclerosing cholangitis: A single center study. Am J Gastroenterol 2007;102:107-114.
20. Floreani A, Rizzotto ER, Ferrara F, Carderi I, Caroli D, Blasone L, Baldo V. Clinical course and outcome of autoimmune hepatitis/primary sclerosing cholangitis overlap syndrome. Am J Gastroenterol 2005;100:1516-1522.
21. Luth S, Kanzler S, Frenzel C, Kasper HU, Dienes HP, Schramm C, Galle PR, et al. Characteristics and long-term prognosis of the autoimmune hepatitis/primary sclerosing cholangitis overlap syndrome. J Clin Gastroenterol 2009;43:75-80.
22. Carbone M, Sharp S, Heneghan M, Neuberger J, Hirschfield G, K B. P1198 : Histological stage is relevant for risk-stratification in primary biliary cirrhosis. . In: J. Hepatol. ; 2015.
23. Scheuer PJ. Ludwig Symposium on biliary disorders--part II. Pathologic features and evolution of primary biliary cirrhosis and primary sclerosing cholangitis. Mayo Clin Proc 1998;73:179-183.
24. Olsson R, Hagerstrand I, Broome U, Danielsson A, Jarnerot G, Loof L, Prytz H, et al. Sampling variability of percutaneous liver biopsy in primary sclerosing cholangitis. J Clin Pathol 1995;48:933-935.

25. Czaja AJ. Hepatic inflammation and progressive liver fibrosis in chronic liver disease. World J Gastroenterol 2014;20:2515-2532.
26. Hammel P, Couvelard A, O'Toole D, Ratouis A, Sauvanet A, Flejou JF, Degott C, et al. Regression of liver fibrosis after biliary drainage in patients with chronic pancreatitis and stenosis of the common bile duct. N Engl J Med 2001;344:418-423.
27. Vesterhus M, Hov JR, Holm A, Schrumpf E, Nygard S, Godang K, Andersen IM, et al. Enhanced liver fibrosis score predicts transplant-free survival in primary sclerosing cholangitis. Hepatology 2015;62:188-197.
28. Westin J, Lagging LM, Wejstal R, Norkrans G, Dhillon AP. Interobserver study of liver histopathology using the Ishak score in patients with chronic hepatitis C virus infection. Liver 1999;19:183-187.
29. Corpechot C, Gaouar F, El Naggar A, Kemgang A, Wendum D, Poupon R, Carrat F, et al. Baseline values and changes in liver stiffness measured by transient elastography are associated with severity of fibrosis and outcomes of patients with primary sclerosing cholangitis. Gastroenterology 2014;146:970-979; quiz e915-976.
30. Ponsioen CY, Chapman RW, Chazouilleres O, Hirschfield GM, Karlsen TH, Lohse AW, Pinzani M, et al. Surrogate endpoints for clinical trials in primary sclerosing cholangitis: Review and results from an International PSC Study Group consensus process. Hepatology 2016;63:1357-1367.
31. Sohrabpour AA, Mohamadnejad M, Malekzadeh R. Review article: the reversibility of cirrhosis. Aliment Pharmacol Ther 2012;36:824-832.

Figure legends

Fig.1. Distribution of grading and staging systems.

(A) Nakanuma grading system. (B) Ishak grading system.

(C) Nakanuma staging system. (D) Ishak staging system. (E) Ludwig staging system.

Fig. 2. Kaplan Meier survival curves of the Nakanuma staging system.

(A) Endpoint 1; Transplant-free survival (B) Endpoint 2; Liver transplantation, (C) Endpoint 3; Liver related events.