Inter-observer variability in histological evaluation of liver fibrosis using categorical and quantitative scores

Michael Pavlides[1,2], Jacqueline Birks[3], Eve Fryer[4], David Delaney[4], Nikita Sarania[5], Rajarshi Banerjee[6], Stefan Neubauer[2], Eleanor Barnes[1,7], Kenneth. A. Fleming[4,5], Lai Mun Wang[4]

[1]Translational Gastroenterology Unit, University of Oxford, Oxford, UK
[2]Radcliffe Department of Medicine, University of Oxford, Oxford, UK
[3]Centre for Statistics in Medicine, University of Oxford, Oxford, UK
[4]Department of Histopathology, Oxford University Hospitals, Oxford, UK
[5]Division of Medical Sciences, University of Oxford, Oxford, UK
[6]Perspectum Diagnostics, Oxford, UK
[7]Peter Medawar Building, University of Oxford, Oxford, UK

Short title: Liver fibrosis scores: inter-observer variability

Conflicts of interest: MP, RB, SN and EB are shareholders in Perspectum Diagnostics (PD), a university spin out company. RB and SN are on the board of directors of PD. RB is employed by PD. JB, EF, DD, KF and LMW declare no conflict of interest.

Word count: 2454

Corresponding author
Dr Lai Mun Wang
Department of Histopathology, John Radcliffe Hospital
Headley Way, Oxford, OX3 9DU
United Kingdom

Email : laimun.wang@ouh.nhs.uk
Telephone: +44 1865 220510
Fax : +44 1865 220519

Abstract

Aims: The aim of the study was to investigate the inter-observer agreement for categorical and quantitative scores of liver fibrosis.

Methods and Results: Sixty five consecutive biopsies from patients with mixed liver disease aetiologies were assessed by three pathologists using the Ishak and NASH CRN scoring systems and the fibrosis area (collagen proportionate area; CPA) was estimated by visual inspection (visual-CPA). A subset of 20 biopsies were analysed using DIA for the measurement of CPA (DIA-CPA). The bivariate weighted kappa between any two pathologists ranged from 0.57 to 0.67 for Ishak staging and from 0.47 to 0.57 for the NASH CRN staging. Bland Altman analysis showed poor agreement between all possible pathologist pairings for visual-CPA, but good agreement between all pathologist pairings for DIA-CPA. There was good agreement between the two pathologists that assessed biopsies by visual-CPA and DIA-CPA. The intra-class correlation coefficient, (equivalent to the kappa statistic for continuous variables), was 0.78 for visual-CPA and 0.97 for DIA-CPA.

Conclusion: These results suggest that DIA-CPA is the most robust method for assessing liver fibrosis followed by visual-CPA. Categorical scores perform less well than both the quantitative CPA scores assessed here.

Keywords: collagen proportionate area, digital imaging analysis, Bland- Altman plot, kappa statistic.

Introduction

Fibrosis represents the final common pathway of injury in chronic liver pathologies including chronic viral hepatitis and non-alcoholic fatty liver disease (NAFLD). The assessment of liver fibrosis is therefore central to the evaluation of patients with liver disease, and liver biopsy is frequently needed to achieve this. Differences in the pathophysiology of each liver disease can result in distinct patterns of fibrosis distribution, and this has led to the development of disease specific semi-quantitative categorical scores like the Ishak score[1] for viral hepatitis or the scoring system for fibrosis developed by the non-alcoholic steatohepatitis Clinical Research Network (NASH CRN) for NAFLD[2].

Quantitative digital imaging analysis (DIA) techniques have also been used for fibrosis evaluation. The quantification of collagen as a proportion of the total liver biopsy area (collagen proportionate area; CPA) was found to correlate with categorical scores of fibrosis and with portal pressure[3], and to predict outcomes in patients with chronic hepatitis C[4] and cirrhosis[5]. Despite this, the use of CPA remains restricted to research studies and has not seen widespread clinical application yet.

In clinical practice liver fibrosis assessment is needed to prioritise patients for treatment, and to monitor disease progression over time. Up until now, the majority of development and validation for histological scores has been focused on hepatitis C assessment. However, recent development of highly effective directly acting

antiviral drugs (DAA) is likely to change this. There is already evidence suggesting that the most cost effective way to manage hepatitis C in the era of DAAs would be to offer treatment to all patients irrespective of disease severity[6], therefore negating the need for fibrosis assessment. The focus of liver fibrosis evaluation is therefore shifting towards other aetiologies and NAFLD in particular. The re-examination of histological scores in this context is now required.

The reproducibility of categorical scores for viral hepatitis has been studied extensively[7-9]. DIA-CPA techniques, which in theory could be used universally irrespective of the underlying aetiology, have also been studied primarily in chronic hepatitis C[3,10]. The inter-observer variability of categorical scores and visual- and DIA-CPA remains largely untested in routine clinical populations and mixed cohorts.

The aim of this study was to assess inter-observer variability in the interpretation of biopsies from a patient population with mixed liver disease aetiologies using both categorical and continuous histological scores of liver fibrosis.

Materials and Methods

Study design and patient population

Sixty five consecutive liver biopsy slides from patients participating in a study of a multi-parametric magnetic resonance imaging technique for liver fibrosis evaluation[11] were assessed. Biopsies were performed under radiological guidance using 18G cutting needles, according to the local clinical practice. The study was approved by the UK National Research Ethics Service (Ref 11/H0504/2), and all patients gave written informed consent.

Routine clinical reporting of biopsies

All biopsies were clinically indicated and were included irrespective of biopsy length. Biopsies were processed according to the local clinical routine, which involves review by two pathologists and a discussion in a clinico-pathological meeting before a final consensus report is issued. For the purposes of this study, this consensus report was used as the reference standard.

In our local practice, fibrosis is staged using the Ishak score[1] in all biopsies. The NASH CRN score[2] is used in addition for cases of NAFLD.

Blinded reporting using standard microscopy

At least six months after the routine clinical reporting, three liver pathologists (LMW, DWD, KAF) independently and blindly re-assessed the biopsies. They evaluated fibrosis using the Ishak score[1] and the NASH CRN fibrosis score[2], in all biopsies irrespective of the underlying aetiology. The amount of excess fibrosis (collagen) as a percentage of the total biopsy area (Collagen Proportionate Area; CPA) was also

estimated by visual inspection of Sirius red stained slides, using a standard light microscope (visual-CPA).

Digital Imaging Analysis for Collagen Proportionate Area

At least six months later, three pathologists (LMW, KAF and EF), used digital imaging analysis (DIA) to calculate the CPA (DIA-CPA) in a subset of 20 slides, that were representative of the disease aetiologies and staging distribution of the whole cohort. Slides stained with Sirius red were scanned using a Hamamatsu Nanozoomer 2.0 HT Digital Pathology System (Hamamatsu, Hamamatsu City, Japan) to produce high quality digital images. Image processing was done using ImageJ (U. S. National Institute of Health, Bethesda, Maryland, USA, http://imagej.nih.gov/ij/, v1.47).

Low magnification images were used so that the whole biopsy sample was visible in one frame. The digital images were manually cropped to remove native collagen from large portal tracts or liver capsule, artefacts in the background and any non-liver tissue (e.g. skeletal muscle). The images were then split into the RGB channels and the green channel was used for further processing. The “threshold” function of the software was used to estimate the total area of liver tissue and the total area of collagen. The area of collagen was expressed as a percentage of the total biopsy area (CPA). Figure 1 illustrates the digital image processing steps.

The DIA technique was developed by MP (hepatology clinical researcher) and LMW (liver histopathologist) based on the published literature[3]. The other pathologists (EF and KAF) received training by MP and LMW in the use of the DIA technique. The

training involved a 45 minute session where the use of the software and the DIA technique was demonstrated. A leaflet with step wise instructions of how to use the software was supplied. The pathologists were provided with a training set of 5 slides and were asked to use these to practice the DIA technique until they were confident in the use of the software. The 5 training slides were different from the 20 used in the final analysis.

Statistics

The statistical analysis plan was designed and executed by JB, an experienced medical statistician. The weighted kappa statistic was used to assess inter-observer agreement in the ordinal fibrosis scores (Ishak and NASH CRN fibrosis). Higher kappa values indicate better agreement between observers, and agreement is generally considered almost perfect if kappa is 0.81 – 0.99, substantial if kappa is 0.61 – 0.80, moderate if kappa is 0.41-0.60, fair if kappa is 0.21-0.40 and only slight if kappa is 0.01 – 0.20.

Bland-Altman plots were used to calculate the mean of the difference between pairs of assessments for visual-CPA and DIA-CPA by different pathologists or where the same pathologist assessed both visual-CPA and DIA-CPA in the same slide. A mean difference of 0 would be expected if the two sets of measurements are in complete agreement. A significant deviation from 0 would suggest poor agreement. The student's t-test was used to evaluate whether the mean difference of the two measurements was significantly different from 0.

Mixed model methodology taking into account the correlation between measurements made on the same slide was used to calculate the “between slide” and the “within slide” variation for the two methods of CPA evaluation. The “between slide” variation is a measure of the variation of CPA in the whole cohort of slides, and largely depends on the range of “true CPA” values in the cohort. The “within slide” variation is a measure of the CPA variation for the value of each individual slide (i.e. the smaller the “within slide variation” the closer the agreement between assessors for each individual slide).

Two pathologists assessed a set of 20 biopsies by all 3 methods (categorical scores, visual-CPA and DIA-CPA). The intra-class correlation coefficient (ICCC) was calculated for the inter-observer agreement in the visual-CPA and DIA-CPA analysis. The ICCC is equivalent to the kappa statistic for the inter-observer assessment of continuous variables.

Results

In the overall biopsy cohort (n=65), the majority of patients had mild fibrosis [F1 in 15 (23%), F2 in 17 (26%)] and cirrhosis (F6) was present in 11 (17%) cases. The two main primary pathologies were non-alcoholic fatty liver disease (NAFLD; n=21; 32%) and chronic viral hepatitis (n=28; 43%). The biopsies had a median (IQR) length of 20mm (16-29), and contained a median of 11 (IQR: 8-15) portal tracts and 6 (IQR: 4-9) central veins. The distribution of fibrosis stage, primary diagnosis and quality measures for biopsies assessed by DIA (n=20) was similar to the overall cohort (Table 1).

Inter-observer agreement– categorical scores of fibrosis (n=65)

For the Ishak score, the bivariate weighted kappa statistics between any two pathologists ranged from 0.57 to 0.67 and for the NASH CRN fibrosis score from 0.47 to 0.57 (Table 2).

Inter-observer agreement – visual-CPA (n=65)

The mean of the difference between the visual CPA assessments of pathologists 1 and 2 was 2.8; between pathologists 1 and 3 was -3.68 and between pathologists 2 and 3 was -6.55. In all the 3 possible pairings the mean of the difference was significantly different from 0, indicating poor agreement (Table 3). In the 20 slides that were subsequently assessed by DIA-CPA, the “between slide” variation was 73, and the “within slide” variation was 21.

Inter-observer agreement – DIA-CPA (n=20)

The mean of the difference between the DIA-CPA assessments of pathologists 1 and 3 was -0.60, between pathologists 1 and 4 it was 0.04 and between pathologists 3 and 4 it was 0.64. None of the 3 pairings showed any significant difference from 0, indicating good agreement (Table 4). The “between slide” variation was 82 and the “within slide” variation was 4.6.

Agreement between methods for CPA assessment (n=20)

Pathologists 1 and 3 assessed 20 slides using both methods (visual-CPA and DIA CPA). The mean of the difference between visual-CPA and DIA-CPA for the assessments by pathologist 1 was 0.49 and for the assessments pathologist 3 it was 1.78, both of which were not significantly different from 0, indicating good agreement (Table 5).

Intra-class correlation coefficients for CPA assessments (n=20).

The ICCC between the assessments of pathologists 1 and 3 for the visual-CPA was 0.78 and for DIA-CPA 0.97.

Discussion

The study evaluated inter-observer variability of quantitative (DIA-CPA and visual-CPA) and categorical scores (Ishak and NASH CRN) for histological assessment of liver fibrosis, in biopsies from patients with mixed liver disease aetiologies. DIA-CPA was the most reproducible score with the highest ICCC (0.97), followed by visual-CPA (ICCC: 0.78). Both categorical scores perform less well (best weighted kappa between pathologist pairings were 0.67 for Ishak and 0.59 for NASH CRN). Furthermore, Bland Altman analysis showed good agreement between all pairs of assessors for DIA-CPA, but in none of the pairs for visual CPA. DIA-CPA also had a lower “within slide” variability (4.6) compared to visual-CPA (21)

These finding becomes even more significant when training and experience are considered. All the pathologists that took part in the study have been trained and have extensive experience in the use of the categorical scores, which are indeed routinely used at our centre. In contrast, they only received a 45 minute training session in the use of the DIA-CPA technique and were given the chance to practice only in 5 slides. No specific training was given in the reporting of visual-CPA.

The high inter-observer agreement for DIA-CPA reported here, is in keeping with other studies[3, 10, 12]. However, comparisons between DIA-CPA techniques and categorical scores, have produced some conflicting results. For example, Pilette *et al*[12], found DIA-CPA to have an ICCC of 0.996 compared to kappa values ranging from 0.29 – 0.87 for the Knodell scoring system[13]. However, in another study, categorical scores were reported to have better inter-observer agreement than DIA-CPA methods[14]. The most likely explanation for this discrepancy is the difference

methodology used in the study by Wright *et al* [14], where different sections from the same biopsy core were used to assess observer dependent agreement and this may have introduced bias due to the quality of staining between the two sections. Despite these conflicting results, we feel that our results in the context of the published literature suggest that DIA-CPA is the most robust and reproducible method for histological liver fibrosis assessment and should be used where the resources are available.

There are however, considerable hurdles that need to be overcome if this technique is to see widespread use. The DIA-CPA techniques depend on the quality of the digital images for the analysis. To achieve the necessary image quality in this study, a slide scanner was used, together with freely available imaging analysis software. Other DIA-CPA methods rely on both proprietary image acquisitions systems and software costing up to US$ 6850[3, 10]. Furthermore, DIA-CPA would require additional time for the preparation and reporting of digital images. We did not assess the time aspects of DIA-CPA analysis but previous studies have reported that it would take an additional 5-10 minutes per biopsy for the image analysis alone[14].

Several aspects of visual-CPA assessment make it appealing for further evaluation as an alternative to DIA-CPA in routine clinical practice. Visual-CPA can be done quickly and requires no extra equipment. Furthermore, in this study visual-CPA assessment achieved lower inter-observer variability than the routinely used categorical scores, and we found good agreement between DIA-CPA and visual-CPA in this study. More experience and training may therefore improve the performance of visual-CPA further. It is also possible for visual-CPA to be used

alongside routine categorical scores and this may provide additional information that could be used clinically.

There are very limited published data for CPA techniques in NAFLD. For example, in the study by Hall *et al*, evaluating DIA CPA in explanted livers[15], patients with NAFLD were not included as it was difficult to establish any contribution to disease from alcohol. Validation of CPA techniques in prospective NAFLD cohorts is therefore required.

Study limitations.

The study evaluated inter-observer agreement of different histological fibrosis scoring systems in a cohort of patients with mixed liver disease aetiologies. The categorical score we chose to assess were not necessarily the ones designed for each specific aetiology. For example the Ishak scoring system[1] was designed for patients with chronic hepatitis C while here it was applied to patients of all aetiologies. However, both the categorical scores are descriptive and rely on the recognition of specific features by the reporting pathologists (e.g. cirrhotic nodules or fibrotic bridges between portal tracks). The recognition of these features should therefore be possible on biopsies from any liver disease aetiology.

Furthermore, there were no minimum biopsy quality criteria despite evidence showing that accuracy depends on biopsy length and number of portal tract present[16, 17]. Despite this, the biopsies included in our study are representative of

biopsies in routine clinical care, where biopsies are rarely repeated for reasons of adequacy.

In conclusion, our data suggest that collagen proportionate area measured using digital imaging analysis techniques (DIA –CPA) is the most reproducible method of histological liver fibrosis assessment, and should be performed where the necessary resources are available. Furthermore, visual-CPA, is an attractive technique that warrants further evaluation as it could be easily implemented alongside traditional reporting. Both CPA techniques would provide adjunctive data to semi quantitative scores, however, further studies are needed to examine whether quantitative techniques could replace categorical reporting in routine practice.

Acknowledgements

This work was supported by grants from the Oxford NIHR Biomedical Research Centre, and the British Heart Foundation.

MP designed the study, developed the DIA-CPA technique locally, analysed data and drafter the manuscript. JB designed and conducted the statistical analysis. EF, DD, NS and KF analysed data. RB designed the study and analysed data. SN and EB designed the study. LMW designed the study, developed the DIA-CPA technique and analysed data. All authors have revised the manuscript for important intellectual content and have read and approved this submission.

References

1. Ishak K, Baptista A, Bianchi L *et al.* Histological grading and staging of chronic hepatitis. *J Hepatol* 1995;**22**;696-699.
2. Kleiner DE, Brunt EM, Van Natta M *et al.* Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology. 2005 Jun;41(6):1313-21.*
3. Calvaruso V, Burroughs AK, Standish R *et al.* Computer-assisted image analysis of liver collagen: Relationship to ishak scoring and hepatic venous pressure gradient. *Hepatology* 2009;**49**;1236-1244.
4. Huang Y, de Boer WB, Adams LA, MacQuillan G, Bulsara MK, Jeffrey GP. Image analysis of liver biopsy samples measures fibrosis and predicts clinical outcome. *J Hepatol* 2014;**61**;22-27.
5. Tsochatzis E, Bruno S, Isgro G *et al.* Collagen proportionate area is superior to other histological methods for sub-classifying cirrhosis and determining prognosis. *J Hepatol* 2014;**60**;948-954.
6. Tsochatzis EA, Crossan C, Longworth L *et al.* Cost-effectiveness of noninvasive liver fibrosis tests for treatment decisions in patients with chronic hepatitis c. *Hepatology* 2014;**60**;832-843.
7. Intraobserver and interobserver variations in liver biopsy interpretation in patients with chronic hepatitis c. The french metavir cooperative study group. *Hepatology* 1994;**20**;15-20.
8. Goldin RD, Goldin JG, Burt AD *et al.* Intra-observer and inter-observer variation in the histopathological assessment of chronic viral hepatitis. *J Hepatol* 1996;**25**;649-654.
9. Gronbaek K, Christensen PB, Hamilton-Dutoit S *et al.* Interobserver variation in interpretation of serial liver biopsies from patients with chronic hepatitis c. *J Viral Hepat* 2002;**9**;443-449.
10. Campos CF, Paiva DD, Perazzo H *et al.* An inexpensive and worldwide available digital image analysis technique for histological fibrosis quantification in chronic hepatitis c. *J Viral Hepat*;**21**;216-222.
11. Banerjee R, Pavlides M, Tunnicliffe EM *et al.* Multiparametric magnetic resonance for the non-invasive diagnosis of liver disease. *J Hepatol* 2014;**60**;69-77.
12. Pilette C, Rousselet MC, Bedossa P *et al.* Histopathological evaluation of liver fibrosis: Quantitative image analysis vs semi-quantitative scores. Comparison with serum markers. *J Hepatol* 1998;**28**;439-446.
13. Knodell RG, Ishak KG, Black WC *et al.* Formulation and application of a numerical scoring system for assessing histological activity in asymptomatic chronic active hepatitis. *Hepatology* 1981;**1**;431-435.
14. Wright M, Thursz M, Pullen R, Thomas H, Goldin R. Quantitative versus morphological assessment of liver fibrosis: Semi-quantitative scores are more robust than digital image fibrosis area estimation. *Liver international : official journal of the International Association for the Study of the Liver* 2003;**23**;28-34.
15. Hall A, Germani G, Isgrò G, Burroughs AK, Dhillon AP. Fibrosis distribution in explanted cirrhotic livers. *Histopathology* 2012;**60**;270-277.
16. Pierre B, Delphine D, Valerie P. Sampling variability of liver fibrosis in chronic hepatitis c. *Hepatology* 2003;**38**;1449-1457.
17. Andrew Rennie H, Emmanuel T, Richard M, Andrew Kenneth B, Amar Paul D. Sample size requirement for digital image analysis of collagen proportionate area in cirrhotic livers. *Histopathology* 2013;**62**;421-430.

Figure legends

Figure 1. Digital Imaging Analysis for estimation Collagen Proportionate Area.

The image in (a) shows the liver biopsy image produced using a high definition slide scanner and (b) demonstrates the manual editing step to crop out artefacts in the background and any non-liver tissue. The image is then split into the RGB components and the green channel (c) is selected for further analysis. Manual thresholding is used to select the collagen in the slide (d) and the number of pixels in this selection is then automatically counted by the software. A further step allows the collagen selection (yellow outline) to be superimposed onto the original image (e and magnified section in f), which allows the reporting pathologist to validate the collagen selection. The total area of the biopsy slide can also be estimated using this technique, and collagen proportionate area is calculated as:

(number of pixels in the collagen area / number of pixels in the whole biopsy)x100.