PO Box 2345, Beijing 100023, China
www.wjgnet.com
wjg@wjgnet.com

World J Gastroenterol 2006 March 21; 12(11): 1713-1717
World Journal of Gastroenterology ISSN 1007-9327

Inter-observer variability in histopathological assessment of liver biopsies taken in a pediatric open label therapeutic program for chronic HBV infection treatment

Marek Woynarowski, Joanna Cielecka-Kuszyk, Andrzej Kałużyński, Aleksandra Omulecka, Maria Sobaniec-Łotowska, Julian Stolarczyk, Wojciech Szczepański

Marek Woynarowski, Department of Gastroenterology, Hepatology and Immunology,Children's Health Memorial Institute, Warsaw, Poland
Joanna Cielecka-Kuszyk, Department of Pathology, Children's Health Memorial Institute, Warsaw, Poland
Andrzej Kałużyński, Department of Pathology, Polish Mother Memorial Research Institute, Łódź, Poland
Aleksandra Omulecka, Department of Pathology, Medical University of Łódź, Poland
Maria Sobaniec-Łotowska, Department of Clinical Pathology, Medical University of Białystok, Poland
Julian Stolarczyk, Department of Pathology, Medical Academy of Gdańsk, Poland
Wojciech Szczepański, Department of Pathology, Jagiellonian University Medical College, Kraków, Poland
Correspondence to: Marek Woynarowski, MD, PhD, Department of Gastroenterology, Hepatology and Immunology, Children's Health Memorial Institute, Al. Dzieci Polskich 20, 04-730 Warsaw, Poland. m.woynarowski@med-net.pl
Telephone: +48-502-236654 Fax: +48-22-8157382
Received: 2005-10-20 Accepted: 2005-11-18

Abstract

AIM: To our knowledge, the inter-observer variability of the liver biopsy findings in HBV-infected children have not been studied as yet. Hence, we aimed to compare different pathologist's assessment of grading and staging in liver biopsies obtained from children prior to interferon treatment.

METHODS: We collected 920 biopsies from 11 medical centers. The biopsies were independently reviewed by 6 pathologists from academic centers who assessed Batts-Ludwig score for grading and staging. Satisfactory agreement among observers was defined as at least 60% of observers having the same opinion. Satisfactory dispersion between maximal and minimal score for the same biopsy specimen was defined as a maximum 1 point.

RESULTS: Satisfactory inter-observer agreement for grading was obtained in 51.6% and for staging in 75.7% of biopsies. Satisfactory dispersion for grading scores was observed in 44.5% and for staging in 72.7% of cases.

CONCLUSION: Our study demonstrates that: (1) pathologists differ in their assessment of grading and staging of liver biopsies; (2) inter-observer variability for staging is lower than that for grading; and (3) regardless of the inter-observer variability of assessments, the majority of children with chronic HBV infection have mild to moderate inflammation and mild to moderate fibrosis.

Key words: Grading; Staging; Type B Hepatitis; Children

Woynarowski M, Cielecka-Kuszyk J, Kałużyński A, Omulecka A, Sobaniec-Łotowska M, Stolarczyk J, Szczepański W. Inter-observer variability in histopathological assessment of liver biopsies taken in a pediatric open label therapeutic program for chronic HBV infection treatment. *World J Gastroenterol* 2006; 12(11): 1713-1717

http://www.wjgnet.com/1007-9327/12/1713.asp

INTRODUCTION

Many authors believe that a liver biopsy is a gold standard in hepatology[1]. However, there is considerable confusion in terminology and the methods of liver biopsy evaluation are still under discussion. These are the reasons why many different scoring systems for liver histological evaluation were proposed. The first scoring system developed by Knodell *et al*[2] has been used in clinical trials since 1981. This classification differentiated necroinflammatory changes and fibrotic liver damage. However, both features were combined in one histological activity index. This was the reason why the Knodell classification was replaced by other systems with clear differentiation between grading of inflammation and staging of fibrosis[3-6]. Nowadays, the scoring systems are widely used by pathologists and clinicians. It often happens that the grading and staging scores replace the descriptive evaluation of biopsy slides. In this situation, one can raise the question: which grading/staging classification is a true gold standard for a liver biopsy assessment? The other issues that can be discussed are results reproducibility and inter-observer variability of the grading and staging scores for the same biopsy. These problems are vital for the clinicians who receive the biopsy reports. However, they are not frequently addressed in the literature.

Table1 The inter-observers agreement definitions

Agreement rate Number of equal assessment/number of observations	Description	Percentage of agreement (%)
6/6, 5/5, 4/4	All observers gave the same score	100
3/4, 5/6, 4/5,	One observer scored differently than others	75-85
4/6, 3/5	Two observers scored differently than others	60-66
3/6. 2/4	Agreement of half of observers	50
2/6, 2/5, 1/4	Agreement of less than half of observers	25-40

We met the problem of liver biopsy interpretation when we analyzed the results of a nationwide program for interferon treatment of chronic HBV infection in children, conducted in Poland between 1993 and 1999. Twenty-six centers from all over the country participated in this program and 3700 children received treatment there. All centers used their local laboratory facilities, but the entry criteria and treatment regime (interferon 3 MU TIW for 20 wk) were the same. The data collected in the local centers was entered into CRF and sent for central analysis. The results published for 1688 children showed 51.5% of HBe clearance at one year after interferon discontinuation. The response was better in younger children and there was a positive correlation with ALT activity[7]. This observation clearly illustrates the problems with liver biopsy result interpretation that are often met by physicians who review the biopsy reports provided by different pathologists.

The liver biopsy specimens could be easily stored in paraffin blocks and as microscopical slides and used for microscopic evaluation many times. These gave us the opportunity to return to the biopsy slides taken from children with chronic HBV infection, treated with interferon in a Polish therapeutic program, and to re-evaluate them according to selected grading and staging classification. Therefore, we asked all participating centers to provide liver biopsy slides for central review. We wanted to see whether the liver biopsies obtained as part of the routine procedure were representative and whether the quality of slides used for diagnosis was satisfactory. We selected scoring system[6] which was used to evaluate the grading and staging of liver damage, and also performed inter-observer variability analysis. The biopsy results were compared with a patient's clinical and serological data to evaluate the role of liver biopsy in the decision-making process regarding children with chronic HBV infection.

The aim of this study was to present the inter-observer variability of liver biopsy assessment performed according to the Batts and Ludwig grading/staging system[6]. The other questions raised in our project will be discussed in further publications.

MATERIALS AND METHODS

The liver biopsy slides taken from 920 children were provided by 11 centers. The number of slides per one biopsy ranged from 1 to 6 (mean 2.36). Hematoxylin-eosin staining were available for all biopsies and for most of them Azan or PAS stains were available as well.

Six independent pathologists, from academic centers, participated in the study. All of them have been collaborating with hepatology units. Before the study was started, the consensus meeting had been organized and Bats-Ludwig liver biopsy assessment scale was chosen for the study as it is widely used in Poland and has a recommendation of Polish Association of the Study of the Liver. Small sample of biopsy slides was reviewed by every pathologist at this consensus meeting and methods of biopsy assessment and results recording were discussed.

The biopsies were divided into sets, which were circulated among observers. The original identification of the biopsies was withheld and pathologists had no additional clinical data on any particular biopsy. Each observer was asked to assess representativity (size of the sample and number of portal zones) and the technical accuracy (fixation and staining) of the slides and to score grading and staging. The assessments were recorded in standard CRF and stored in a central database. For each biopsy, at least four observers' assessments were obtained.

We analyzed the score distribution for every observer and the agreement rate between different observers for the same biopsy (Table 1). Agreement of 60-66% of observers was selected as the desired minimal level of inter-observer agreement.

For each biopsy, the dispersion between maximal and minimal score for grading and staging was calculated. Satisfactory dispersion between maximal and minimal score was defined as being not greater than 1 point.

All biopsies were taken as part of a routine diagnostic procedure and the parents were asked to sign the consent for liver biopsy. This project of biopsy retrospective assessment received the approval from the Ethics Committee of Children's Health Memorial Institute in Warsaw (decision No 185/KE/2000).

RESULTS

Biopsy representativity and technical accuracy assessment

A total of 5 295 opinions on biopsy specimen representativity were available. In 83.9% of the reports, the observers stated that the specimen was representative of the tissue and could be used for grading and staging evaluation. However, the rate of representative specimens varied among different observers from 65.9% to 93.3% (Figure 1). The agreement of at least 60% of observers was reached in 97% of cases (Figure 2).

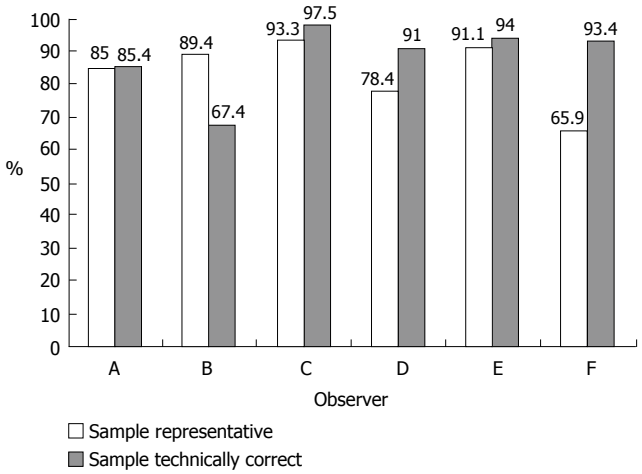

Figure 1 Observers' opinions on representativity and technical accuracy of liver biopsy specimens.

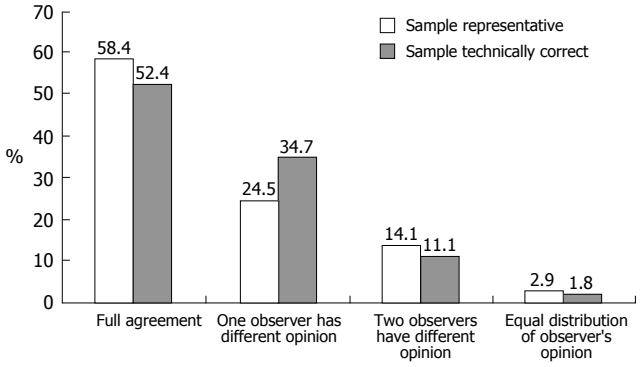

Figure 2 Inter-observer agreement in the assessment of liver biopsy specimen representativity and technical accuracy of the slides.

The opinions on technical accuracy of the biopsy slides were available in 5 296 reports and 88.1% of them showed that the slides were technically correct. The individual observers differed in their opinion on technical accuracy of the biopsy by between 67.4% and 97.5% (Figure 1). The desired level of agreement was reached in over 98% of cases (Figure 2).

Grading assessment

Four or more observers' assessments of biopsy grading were available for 844 biopsies and a total of 4 900 reports were analyzed. The frequency of different scoring varied among observers (Figure 3). The smallest differences in grading score frequencies among observers were noted for a score of 0 (0.3-17.2%, mean frequency for all observers: 8.1%) and 4 (0-7.8%, mean: 4.1%). The largest differences were noted for a score of 1 (13-50%, mean for all observers: 31%) and 3 (4-35.9%, mean: 22%).

The grading assessment agreement of more than 60% of the observers was obtained in 51.6% of cases, but full agreement of all observers reached in only 3.1% of cases. For 13.4% of biopsies, a large difference in observers' opinions was noted (Figure 4)

Different observers' grading score dispersion for the same biopsy equal or lower than 1 point was noted in 44.5% of cases and a dispersion of 2 points was noted in 47.3%. Substantial score dispersion among observers was noted in 8.2% of the cases (Figure 5).

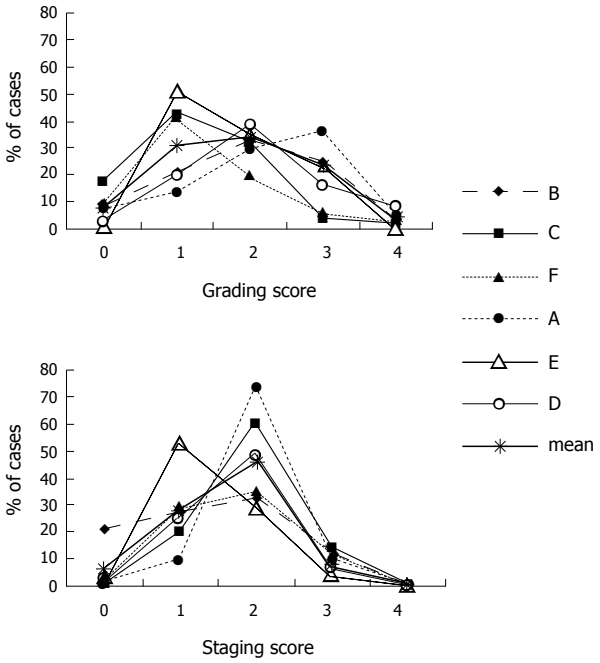

Figure 3 The frequency of different grading and staging scores for individual observer and the mean score frequency for all observers.

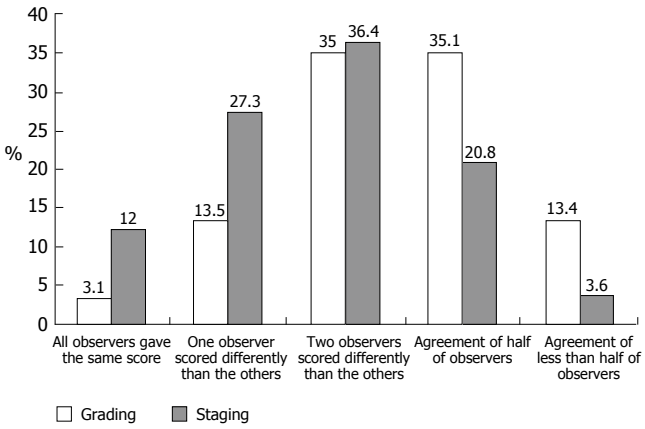

Figure 4 The agreement between different observers' grading and staging score for the same liver biopsy.

Staging assessment

Four or more observers' staging assessments were available for 843 biopsies and a total of 4 895 reports were analyzed. The frequency of different scoring varied among observers (Figure 3). The smallest differences in staging score frequency among observers were noted for a score of 4 (0-0.76%, mean frequency for all observers: 0.4%). The largest differences in score frequency among individual observers were noted for a score of 1 (9.6-52.7%, mean for all observers: 27.1%) and a score of 2 (28.4-72%, mean for all observers: 46.4%).

The staging assessment agreement of more than 60% of the observers was obtained in 75.7% of cases and

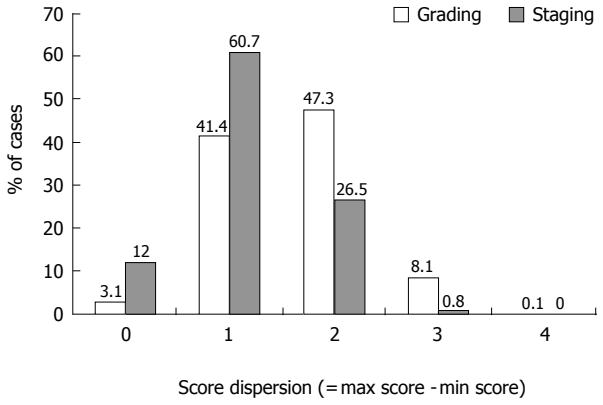

Figure 5 Grading and staging dispersion of assessments among different observers (maximal score - minimal score for the same biopsy).

full agreement of all observers reached in 12% of cases. Substantial differences in observers' opinion were noted in only 3.6% of biopsies (Figure 4).

Different observers' staging score dispersion for the same biopsy equal or lower than 1 point was noted in 72.7% of cases and dispersion of 2 points was noted in 26.5%. Substantial score dispersion among observers was noted very rarely, in only 0.8% of the cases (Figure 5).

DISCUSSION

Liver biopsy is a complex diagnostic procedure performed in many medical centers and evaluated in many laboratories. The aim of the biopsy is to obtain objective information about the condition of the liver tissue. However, there are many factors that influence the objectivity of this examination. Rousselet *et al*[8] have shown that the pathologists' experience is one of the most important factor influencing the inter-observer agreement on viral hepatitis liver biopsy assessment. The other factors that can be taken into account are: the place of liver biopsy sampling; the size of liver biopsy; fixation and staining of the liver tissue specimen; patient's age and reason for liver damage[9].

To our best of knowledge, there are no comprehensive studies addressing the inter-observer variability in liver biopsies taken from children with HBV infection. We, therefore, aimed to investigate the inter-observer error among pathologists who receive a series of biopsies taken in every day practice.

When analyzing the inter-observer variability, one must take into account the definition of agreement. Peutz *et al*[10] showed that the tolerance ±1 point markedly increases the reproducibility of staging results on the Ishak scale. The complexity of assessment scale and number of observers can also influence the variability of diagnoses. Gronback *et al*[11], in their report on a series of 46 liver biopsies analyzed independently by 5 observers, showed that the more complex the scale, the bigger the variability of results. The variability in different pathologists' assessment of the liver biopsy may lead to variability of therapeutic decisions based on liver biopsy reports. Thus, in our study, we compared the results obtained from different pathologists for the same biopsies. To provide a practical value of the results for every day life, we have resigned from sophisticated statistics. The questions that we asked were very simple but vital for the clinician who collaborates with different independent pathologists.

In our biopsy series, we demonstrated that inter-observer variability depended on the scale complexity and agreement cut-off. When the same opinion of at least 60% of observers (≥4 out of 6, or ≥3 out of 5 observers) was used as a cut-off point, the inter-observer agreement for representativity reached 97% and for technical accuracy of the biopsy reached 98%. Even with the higher cut-off, the inter-observer agreements for these two parameters were satisfactory. These positive results could be obtained in a simple scale with a two-option (yes/no) selection[12].

The Batts-Ludwig grading and staging scale is more complex. With the same cut-off as specified above (agreement of at least 60% of observers), the inter-observer agreement for grading reached in 51.6% of cases and for staging 75.7% of cases. A large number of observers was probably the reason why the agreement with the higher cut-off level was very low.

Moreover, a large number of observers was also the reason why the dispersion between the maximal and minimal score for single biopsies was high. In a four-step scale, a maximal difference of 1 point should probably be used as acceptable tolerance. This level of dispersion was achieved for grading in 44.5% of cases and for staging in 72.7%. We found a higher inter-observer agreement and a lower score dispersion for staging than for grading, showing that staging assessment is more reliable and probably more reproducible than grading assessment. This has also been confirmed by other studies[1,8].

The analysis of inter-observer grading and staging agreements and score dispersion showed that despite the same assessment scale being used by pathologists, their reports differed. This is because grading and staging assessments are essentially subjective[13] and this is not a unique situation for liver biopsy only. Similar discrepancies have been reported for other diagnostic techniques in hepatology as well[14]. However, despite the inter-observers variability, the numbers of children with extremely low and extremely high grading or staging were small. Our data showed that the vast majority of children with chronic HBV infection had mild to moderate inflammatory changes and mild to moderate fibrosis. The risk of biopsy-proven HBV-related cirrhosis in this age group is minimal. This observation confirms the clinical observations of Bortolotti *et al*[15] who could not find a progression to liver cirrhosis in children with HBV infection. On the other hand, our data indicated that the chance of completely normal liver tissue in children with chronic HBV infection is small. These observations suggest that the approach to liver biopsy should be changed. Liver biopsy should not be a mandatory examination in all children with active HBV infection, but it should be performed only in selected patients: those with persistently abnormal ALT despite HBeAg clearance, or those with initial signs of liver destruction. In everyday practice, the grading and staging scores should not replace the descriptive biopsy report, as scoring systems are not reliable and they reduce the amount of information on liver tissue. Grading and staging scores become important information in a liver biopsy

report when the series of liver biopsies for the same patient are analyzed for research or follow-up reasons, providing, however, that all the patients' biopsies are reviewed by the same pathologist on the same occasion. It seems important that both clinician and pathologist are aware of the liver biopsy limitations and they closely collaborate with each other.

In conclusion, our study shows that pathologists differ in their assessment of grading and staging of liver biopsy, but the inter-observer variability for staging is lower than that for grading. The majority of children with chronic HBV infection have mild to moderate inflammation and mild to moderate fibrosis. Thus, the value of liver biopsy as a guide for current therapeutic decision in children with chronic HBV infection is limited. The liver biopsies should rather be used as a tool for research or long-term disease dynamics assessment.

ACKNOWLEDGMENTS

The authors thank the 11 centers that participated in the Polish Pediatric Interferon Program for HBV therapy: Children's Health Memorial Institute in Warsaw, Medical Academy of Łódź, Medical Academy of Bydgoszcz, Infectious Hospital in Bydgoszcz, Regional Hospital in Toruń, Medical Academy of Gdańsk, Silesian Medical Academy, Medical Academy of Białystok, Medical Academy of Warsaw, Kraków Specialist Hospital, Infectious Hospital in Gdańsk. The authors also thank the heads of the teams: Prof. Jerzy Socha, Prof. J. Kuydowicz, Prof. M. Czerwionka-Szaflarska, Dr. E. Smukalska, Dr. E. Strawińska, Dr. A. Liberek, Prof. K. Karczewska, Dr. D. Lebensztejn, Dr. B. Kowalik-Mikołajewska, Dr. A. Gorczyca, Dr Z. Michalska and all their colleagues who participated in the program for providing the biopsy slides for this study. The authors thank Dr. Flavia Bortolotti from University of Padua for remarks and advices for the manuscript preparation.

REFERENCES

1 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

2 **Knodell RG,** Ishak KG, Black WC, Chen TS, Craig R, Kaplowitz N, Kiernan TW, Wollman J. Formulation and application of a numerical scoring system for assessing histological activity in asymptomatic chronic active hepatitis. *Hepatology* 1981; **1**: 431-435

3 **Scheuer PJ.** Classification of chronic viral hepatitis: a need for reassessment. *J Hepatol* 1991; **13**: 372-374

4 **Desmet VJ,** Gerber M, Hoofnagle JH, Manns M, Scheuer PJ. Classification of chronic hepatitis: diagnosis, grading and staging. *Hepatology* 1994; **19**: 1513-1520

5 **Ishak K,** Baptista A, Bianchi L, Callea F, De Groote J, Gudat F, Denk H, Desmet V, Korb G, MacSween RN. Histological grading and staging of chronic hepatitis. *J Hepatol* 1995; **22**: 696-699

6 **Batts KP,** Ludwig J. Chronic hepatitis. An update on terminology and reporting. *Am J Surg Pathol* 1995; **19**: 1409-1417

7 **Woynarowski M**, Socha J. Results of interferon alpha treatment in children with chronic type B hepatitis. Polish centers experience between 1990 and 1997. *Ped Pol* 1998; **73:** 1031-1041

8 **Rousselet MC,** Michalak S, Dupre F, Croue A, Bedossa P, Saint-Andre JP, Cales P. Sources of variability in histological scoring of chronic viral hepatitis. *Hepatology* 2005; **41**: 257-264

9 **Regev A,** Berho M, Jeffers LJ, Milikowski C, Molina EG, Pyrsopoulos NT, Feng ZZ, Reddy KR, Schiff ER. Sampling error and intraobserver variation in liver biopsy in patients with chronic HCV infection. *Am J Gastroenterol* 2002; **97**: 2614-2618

10 **Petz D,** Klauck S, Rohl FW, Malfertheiner P, Roessner A, Rocken C. Feasibility of histological grading and staging of chronic viral hepatitis using specimens obtained by thin-needle biopsy. *Virchows Arch* 2003; **442**: 238-244

11 **Gronbaek K,** Christensen PB, Hamilton-Dutoit S, Federspiel BH, Hage E, Jensen OJ, Vyberg M. Interobserver variation in interpretation of serial liver biopsies from patients with chronic hepatitis C. *J Viral Hepat* 2002; **9**: 443-449

12 **Goldin RD,** Goldin JG, Burt AD, Dhillon PA, Hubscher S, Wyatt J, Patel N. Intra-observer and inter-observer variation in the histopathological assessment of chronic viral hepatitis. *J Hepatol* 1996; **25**: 649-654

13 **Scheuer PJ,** Standish RA, Dhillon AP. Scoring of chronic hepatitis. *Clin Liver Dis* 2002; **6**: 335-47, v-vi

14 **Winkfield B,** Aube C, Burtin P, Cales P. Inter-observer and intra-observer variability in hepatology. *Eur J Gastroenterol Hepatol* 2003; **15**: 959-966

15 **Bortolotti F,** Jara P, Crivellaro C, Hierro L, Cadrobbi P, Frauca E, Camarena C, De La Vega A, Diaz C, De Moliner L, Noventa F. Outcome of chronic hepatitis B in Caucasian children during a 20-year observation period. *J Hepatol* 1998; **29**: 184-190

S- Editor Guo SY **L- Editor** Kumar M **E- Editor** Ma WH