Interobserver variation in histopathological assessment of Helicobacter pylori gastritis

Ozlem Aydin, Reyhan Egilmez, Tuba Karabacak, Arzu Kanik

Ozlem Aydin, Reyhan Egilmez, Tuba Karabacak, Arzu Kanik, Department of Pathology and Biostatistics, Medical School, Mersin University, Mersin-Icel, Turkey
Correspondence to: Ozlem Aydin, M.D. Department of Pathology, Medical School, Mersin University, 33079 Mersin- Icel, Turkey. ozlemaydin66@hotmail.com
Telephone: +90-324-3374300/1192  Fax: +90-324-3374305
Received: 2003-07-04  Accepted: 2003-08-19

Abstract
AIM: Because the presence or absence of H pylori infection has important implications for therapeutic decisions based on histological assessment, the reproducibility of Sydney system is important. The study was designed to test the reproducibility of features of Helicobacter pylori gastritis, using the updated Sydney classification.

METHODS: Gastric biopsies of 40 randomly selected cases of H pylori gastritis were scored semiquantitatively by three pathologists. Variables analysed included chronic inflammation, inflammatory activity, atrophy, intestinal metaplasia, H pylori, surface epithelial damage. $\kappa$ values below 0.5 represented poor, those between 0.5 and 0.75 good and values over 0.75 excellent interobserver agreement.

RESULTS: The best interobserver agreement ($\kappa$=0.62) was present for intestinal metaplasia. The agreement was the poorest for evaluating atrophy ($\kappa$=0.31).

CONCLUSION: Although the results of this study were in accordance with some previous studies, an excellent agreement could not be reached for any features of H pylori gastritis. This low degree of concordance is assumed to be due to the personal evaluation differences in grading the features, the lack of standardized diagnostic criteria, and the ignorance to reach a consensus about the methods to be used in grading the features of H pylori gastritis before initiating the study.

Aydin O, Egilmez R, Karabacak T, Kanik A. Interobserver variation in histopathological assessment of Helicobacter pylori gastritis. World J Gastroenterol 2003; 9(10): 2232-2235
http://www.wjgnet.com/1007-9327/9/2232.asp

INTRODUCTION
Although gastritis was first interpreted to be due to aging and lifelong exposure to various insults, it is now clear that the most common cause of this inflammatory condition is infection with H pylori[9]. It has been shown[10] that this organism is strongly associated with chronic active gastritis as well as gastric adenocarcinoma and MALToma.

The Sydney system for grading and classifying chronic gastritis was devised to provide a standardized approach to the histologic interpretation of gastric biopsies in 1990[5]–[8], and it was later updated in 1994[9,10]. Although it was reported that the Sydney systems’ weakness was that it was used in complex descriptions rather than true diagnosis[7]. After the updating of the Sydney classification, several studies on interobserver variation for the assessment of H pylori gastritis have been reported[6,8-11]. The evaluation of interobserver agreement by using kappa ($\kappa$)-statistics has been accepted by pathologists for several years[9].

Although the histologic examination of gastric biopsy specimens is accepted as the gold standard[12,13] for the diagnosis of H pylori gastritis, it has not been demonstrated that histopathologic assessment is both accurate and reproducible[9].

The study was designed to test the reproducibility of the features of H pylori gastritis, using the updated Sydney classification by $\kappa$-statistics.

MATERIALS AND METHODS
Three pathologists participated in the study. One was a professor with primary interest in gastrointestinal pathology. The second was a 4-month assistant professor in pathology. An other was an 18-month pathology resident. The slides were examined independently, and also in combination with any clinical information by each of the pathologists.

Histologic evaluation
From 130 cases diagnosed as H pylori gastritis in our department (Department of Pathology, Medical School, Mersin University.) in a period of 17 months, 40 [22 (55.0 %) female, 18 (45.0 %) male] were randomly selected for study, their age ranged from 23 to 72 years, with a mean of 47.2. The specimens were excluded from the study because they were insufficient in mucosal thickness for proper assessment of atrophy and without surface epithilium before the selection. Slides were coded using a computer generated list of random numbers.

Biopsy samples from the antrum and body were formalin-fixed and paraffin-embedded and cut into 2-3 μm sections which were stained using hematoxylin and eosin (H&E), and alcian blue/PAS for intestinal metaplasia. Five H&E sections were examined for each case. The biopsies were scored semiquantitatively by three pathologists according to the updated Sydney classification[14].

The updated Sydney system has a scale of 0-3 for scoring the features of chronic gastritis[14]. In order to improve assessment of minor degrees of alteration, a detailed histopathological classification was used, which also provides numerical data for statistical analysis[15]. At first, each variable was divided into seven subcategories, resulting in a score on a scale of 0-6. But the $\kappa$ values could not be calculated using this classification. The 6 subcategories (excluding 0, none) were then amalgamated by pairs (none, 0; mild, 1-2; moderate, 3-4; severe, 5-6), but the calculation of the $\kappa$ values was again impossible for the majority of variations using this classification, and the calculated values were found to be low. So, we came to a conclusion that the agreement between pathologists could be improved when a different amalgamated 3-point scale classification was used for each variable (Table 1).
Interobserver agreement was analysed with the use of \( \kappa \) statistics (BMDP software: Cork, Ireland). The benchmarks suggested by Svanholm et al.\[16\] were accepted. Values below 0.5 represented poor, those between 0.5 and 0.75 good and values over 0.75 excellent interobserver agreement. Only values greater than 0.5 were considered good enough for diagnostic reliability. Confidence interval was calculated for only statistically significant values.

**RESULTS**

\( \kappa \) values and their 95 % confidence intervals between three pathologists for \textit{H pylori} gastritis are shown in Table 2. On
blinded review of the coded slides the best interobserver agreement (κ=0.62; CI: 0.40-0.85) was present for intestinal metaplasia. The good agreement was reached in the assessment of the grade of *H pylori*, with κ value of 0.56 (CI: 0.28-0.84). The interobserver agreement was the poorest for evaluating atrophy (κ=0.31; CI:0.13-0.56). Following atrophy, the two variables with poor agreement were chronic inflammation (κ=0.49; CI: 0.13-0.85) and inflammatory activity (κ=0.44; CI: 0.13-0.71).

There was an agreement among the three observers for only evaluating intestinal metaplasia and the grade of *H pylori*. There was no interobserver agreement among the three pathologists for the assessment of surface epithelial damage. An excellent agreement could not be reached in any features of *H pylori* gastritis in our study.

**DISCUSSION**

Correct and reliable histological diagnosis of *H pylori* gastritis has a great influence on clinical practice as an indicator for therapy. Reliability in assessing intestinal metaplasia and atrophy in histological specimens was especially important because these changes were associated with an increased risk of gastric cancer[12,17-19]. Andrew et al[8] and Tepes et al[12] held that histopathology was a reliable diagnostic method for *H pylori* gastritis based on their results.

The best interobserver agreement was reached for intestinal metaplasia. The κ values were 0.51-0.62 (CI: 0.40-0.85). As in our study, others have also shown a good agreement for scoring intestinal metaplasia, with κ values varying from 0.54 (CI: 0.31-0.77) in the study by Tepes et al[12] to 0.73 in the study by Andrew et al[8]. However, our κ values were lower than those reported by Fiocca et al[20] (κ=0.75-0.92). Although, the H&E stain has been the standard basis for recognition of intestinal metaplasia[21], we based our observations on the alcian blue/PAS in addition to H&E because of ease to identify the goblet cells.

In the present study, the grading of *H pylori* reached good reproducibility, with κ value of 0.56 (CI: 0.28-0.84). This result was consistent with the study of Fiocca et al[20] (κ=0.62), Andrew et al[8] (κ=0.74) and Tepes et al[12] (κ=0.43), but lower than the value reported by El-Zimaity et al[9] (κ=0.90). Our results have also confirmed that H&E was an adequate stain for the detection of *H pylori*. There was no need for an additional staining like Warthin-Starry to identify the organism.

The lack of explicit criteria for the diagnosis of normal gastric mucosa when mononuclear cells were present, made grading difficult[22]. Therefore, the κ value for assessment of the degree of chronic inflammation (κ=0.49; CI:0.13-0.85) using semiquantitative scoring was lower than that for intestinal metaplasia and for the grading of *H pylori* in the present study. Tepes et al[12], also found a κ value for chronic inflammation ranged from 0.39 to 0.53. Our result is also in accordance with those of Fiocca et al[20], who reported κ values ranging from 0.49 to 0.82 and Andrew et al[8] who reported κ value of 0.58.

The interobserver agreement was poor with κ value of 0.44 (CI: 0.13-0.71) for scoring neutrophil infiltration in gastric mucosa. This result was consistent with those of Tepes et al[12] (κ=0.28-0.41) and Andrew et al[8] (κ=0.69). But the interobserver agreements of the studies of El-Zimaity et al[9] (κ=0.80) and Fiocca et al[20] (κ=0.58-0.77) were better than ours. Inflammatory activity and *H pylori* infection were present together and when only neutrophils were discovered in the tissue specimen the pathologists should intensively search for some residual *H pylori*.[22]

Recently, it has been shown in several studies that even experienced gastrointestinal pathologists had poor interobserver agreement over the assessment of gastric atrophy of *H pylori* gastritis[8,9-11]. In the present study, the interobserver agreement for the grade of atrophy was lower than that for the other gastritis features. As in our study (κ=0.31, CI:0.13-0.56), others have also shown the lowest agreement for scoring atrophy, with κ values varying from 0.42 in the study of Fiocca et al[20] to 0.51 in the study of Andrew et al[8]. Tepes et al[12] also found the lowest interobserver agreement for atrophy (κ=0.17-0.57). Although El-Zimaity et al[9] also found the poorest agreement for atrophy, with κ value ranged from 0.08 to 0.29, the agreement in our study for the evaluation of atrophy was still better.

Among the similar previous studies, the surface epithelial damage in *H pylori* gastritis has been evaluated in only the study of Chen et al[5]. They reached good to excellent reproducibility in grading this feature, with weighed κ values of 0.6 and 0.73. But there was no interobserver agreement between the three pathologists for the assessment of surface epithelial damage in our study. Although the Sydney classification has been used routinely, the surface epithelial damage in *H pylori* gastritis have not been evaluated in our department until the present study was designed. It is suggested that the reason of this disagreement may be the lack of our experience in evaluating epithelial damage.

The results of this study suggest that assessment of many histopathologic features of *H pylori* infection have a low degree of concordance. Interobserver variation has been rather high in this study as in some other studies[8,12,23]. This may be due to the discrepancies in the interpretation of the features of *H pylori* gastritis, or due to the observations of the pathologists. Essentially, a perfect agreement by pathologists was practically impossible because pathology results were based on subjective interpretation of different features and classification, and numerous studies on the reproducibility of histopathologic data have reached similar conclusion. Pathologists could usually agree in the presence or absence of a particular histological characteristic, but were seldom consistent when they estimated its degree[24-27].

In the present study, the best interobserver agreement was reached between the assistant professor and the pathology resident, suggesting that the scale of the score is more important than experiences.

Because of the level of agreement in the presence or absence of *H pylori* infection had important implications for therapeutic decisions based on histological assessment[9], reproducibility of Sydney system is important. The updated Sydney system for scoring *H pylori* gastritis is useful and reproducible, but it needs to be improved in the criteria for grading the histologic features[25]. The lack of standardized diagnostic criteria is likely to have contributed significantly to the poor interobserver agreement found in certain features such as atrophy[9] as in our study. More exact criteria will probably further improve the interobserver agreement in assessing the histologic features, but some interobserver variability will probably persist because of the subjectivity that has been part of all semiquantitative grading systems[12]. The point that where cases were reviewed and numerical parameters were established was the best strategy to improve diagnostic concordance between pathologists[28].

Although, the results of this study were in accordance with some previous studies, an excellent agreement could not be reached for any features of *H pylori* gastritis. In conclusion, this unexpectedly low degree of concordance is assumed to be due to the personal evaluation differences in grading the features, and the lack of the standardized diagnostic criteria, as well as the ignorance to reach a consensus about the methods to be used in grading the features of *H pylori* gastritis before initiating the study.

**REFERENCES**

1. Soll AH. Gastritis and *Helicobacter pylori*. In: Goldman L,
Bennett JC, eds. Cecil Textbook of Medicine, 21st ed. Philadelphia: Saunders 2000: 643-767

Petersen WL, Graham DY. Helicobacter pylori. In: Feldman M, Scharschmidt BF, Sleisenger MH, eds. Sleisenger & Fordtran’s Gastrointestinal and Liver Disease: Pathophysiology (Diagnosis) Management, 6th ed. Philadelphia: 1998: 604-620

Misiewicz JJ, Tytgat GNJ, Goodwin CS. The Sydney system: a new classification of gastritis. J Hepatol Gastroenterol 1991; 6: 209-222

Owen DA. The Stomach. In: Sternberg SS, ed. Diagnostic Surgical Pathology. 3rd ed. Philadelphia: Lippincott Williams & Wilkins 1999: 1311-1349

Genta RM, Dixon MF. The Sydney system revisited: the Houston international gastritis workshop. Am J Gastroenterol 1995; 90: 1039-1041

Genta RM. Helicobacter pylori, inflammation, mucosal damage, and apoptosis: pathogenesis and definition of gastric atrophy. Gastroenterology 1997; 113(Suppl): 551-555

Fenoglio-Preiser CM, Noffsinger AE, Stemmermann GN, Lantz PE, Listrom MB, Rilke FO. The Nonneoplastic Stomach. In: Fenoglio-Preiser CM, Noffsinger AE, Stemmermann GN, Lantz PE, Listrom MB, Rilke FO, eds. Gastrointestinal Pathology An Atlas and Text, 2nd ed. Philadelphia: Lippincott-Raven 1999: 153-237

Andrew A, Wyatt JJ, Dixon MF. Observer variation in the assessment of chronic gastritis according to the Sydney system. Histopathol 1994; 25: 317-322

el-Zimaty HM, Graham DY, al-Assi MT, Malaty H, Karttunen TJ, Graham DP, Huberman RM, Genta RM. Interobserver variation in the histopathological assessment of Helicobacter pylori gastritis. Hum Pathol 1996; 27: 35-41

Alhomsi MF, Adeyemi EO. Grading Helicobacter pylori gastritis in dyspeptic patients. Comp Immunol Microbiol Infect Dis 1996; 19: 147-154

van Grieken NC, Weiss MM, Meijer GA, Bloemena E, Lindeman J, Offerhaus GJ, Meuwissen SG, Baak JP, Kuipers EJ. Rapid quantitative estimation of gastric corpus atrophy in tissue sections. J Clin Pathol 2001; 54: 63-69

Tepes B, Ferlan-Marolt V, Jutersek A, Kaucic B, Zaletel-Kragelj L. Interobserver agreement in the assessment of gastritis reversibility after Helicobacter pylori eradication. Histopathology 1999; 34: 124-133

Genta RM. Pathology of Helicobacter pylori infection. In: Weinstein RS, ed. Advances in Pathology an Laboratory Medicine. St. Louise, Mosby 1994: 443-465

Dixon MF, Genta RM, Yardley JH, Correa P. Classification and grading of gastritis, the updated Sydney system. Am J Surg Pathol 1996; 20: 1161-1181

Chen XY, Huist RWM, Bruno MJ, van der Ende A, Xiao SD, Tytgat GN, Ten Kate FJ. Interobserver variation in the histopathological scoring of Helicobacter pylori related gastritis. J Clin Pathol 1999; 52: 612-615

Svanholm H, Starklint H, Gundersen HJ, Fabricius J, Bartebo H, Olsen S. Reproducibility of histomorphologic diagnoses with special reference to kappa statistics. APMIS 1989; 97: 689-698

Correa P. Human gastric carcinogenesis: a multistep and multifactorial process. First american cancer society award lecture on cancer epidemiology and prevention. Cancer Res 1992; 52: 6735-6740

Sipponen P. Gastric cancer-a long term consequence of Helicobacter pylori infection? Scand J Gastroenterol Suppl 1994; 201: 24-27

Meining A, Stolle M. Close correlation of intestinal metaplasia and corpus gastritis in patients infected with Helicobacter pylori. Z Gastroenterol 2002; 40: 557-560

Fiocca R, Villani L, Cornaggia M. Interobserver variation in the assessment of H pylori gastritis (abstract). Gut 1996; Suppl 2: A104-A105

Segura D, Montero C. Histochemical characterization of different types of intestinal metaplasia in gastric mucosa. Cancer 1983; 52: 498-503

Genta RM, Lew GM, Graham DY. Changes in the gastric mucosa following eradication of Helicobacter pylori. Mod Pathol 1993; 6: 281-289

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

Riddell RH, Goldman H, Ransohoff DF, Appelman HD, Fenoglio CM, Haggitt RC, Aheren C, Correa P, Hamilton SR, Morson BC, Sommers SC, Yardley JH. Displasia in inflammatory bowel disease: Standardized classification with provisional clinical application. Hum Pathol 1983; 14: 931-968

Reid BJ, Haggitt RC, Rubin CE, Roth G, Surawicz CM, Van Belle G, Lewin K, Weinstein WM, Antonioli DA, Goldman H, Mac Donald W, Owen D. Observer variation in the diagnosis of dysplasia in Barrett’s oesophagus. Hum Pathol 1988; 19: 166-178

Dawson A, Ibrahim NB, Gibbs AR. Observer variation in the histopathological classification of thymoma: Correlation with prognosis. J Clin Pathol 1994; 47: 519-523

Sorensen JB, Hirsch FR, Gazdar A, Olsen JE. Interobserver variability in histopathologic subtyping and grading of pulmonary adenocarcinoma. Cancer 1993; 71: 2971-2976

Guarnier J, Herrera-Goepfert R, Mohar A, Sanchez L, Halperin D, Ley C, Parsonnet J. Interobserver variability in application of the revised Sydney classification for gastritis. Hum Pathol 1999; 30: 1431-1434

Edited by Wang XL