Human Pathology (2021) 107, 29–38

Human
PATHOLOGY

www.elsevier.com/locate/humpath

Original contribution

Interobserver agreement in pathologic evaluation of bile duct biopsies☆

Yong-Jun Liu MD, PhD [a,b], Jessica Rogers MD [a], Yao-Zhong Liu MD, PhD [c], Xianyong Gui MD, PhD [a], Florencia Jalikis MD [a], Lisa Koch MD, PhD [a], Paul E. Swanson MD [a], Camtu D. Truong MD [a], Matthew M. Yeh MD, PhD [a,d,*]

[a] *Department of Laboratory Medicine and Pathology, University of Washington School of Medicine, Seattle, WA, 98195, USA*
[b] *Department of Pathology and Laboratory Medicine, University of Wisconsin School of Medicine and Public Health, WI, 53705, USA*
[c] *Department of Biostatistics and Data Science, Tulane University School of Public Health and Tropical Medicine, LA, 70112, USA*
[d] *Department of Medicine, University of Washington School of Medicine, Seattle, WA, 98195, USA*

Received 16 August 2020; revised 7 October 2020; accepted 12 October 2020

Available online 29 October 2020

Keywords:
Biliary strictures;
Bile duct biopsy;
Dysplasia;
Carcinoma;
Interobserver agreement

Summary Intraductal biopsy is commonly used for preoperative evaluation of the etiology of biliary strictures. Interpretation of intraductal biopsies is frequently challenging. The diagnosis often suffers from interobserver disagreement, which has not been studied in the literature. We sought to assess interobserver concordance in the interpretation of intraductal biopsies. Eighty-five biopsies were retrieved, falling into five diagnostic categories: negative for dysplasia (NED), indefinite for dysplasia (IND), low-grade dysplasia (LGD), high-grade dysplasia (HGD), and carcinoma (CA). Eight gastrointestinal pathologists blindly reviewed all the slides. Agreement among pathologists was analyzed using Fleiss κ and weighted concordance coefficient S^*. A face-to-face consensus/training session was held to discuss the classification criteria, followed by a second round review. The overall interobserver agreement was fair in the first round review (κ = 0.39; $S* = 0.56$) and improved to moderate in the second round review (κ = 0.48; $S* = 0.69$). The agreement before and after consensus meeting was substantial to nearly perfect for CA (κ = 0.65, $S* = 0.83$; and κ = 0.80, $S* = 0.91$), fair for HGD (κ = 0.28, $S* = 0.69$; and κ = 0.40, $S* = 0.63$), and moderate for NED (κ = 0.47, $S* = 0.50$; and κ = 0.47, $S* = 0.53$). Agreement improved from fair to moderate for LGD (κ = 0.36, $S* = 0.61$; and κ = 0.49, $S* = 0.71$) and slight to fair for IND (κ = 0.16, $S* = 0.51$; and κ = 0.33, $S* = 0.50$). Compared with Hollande's fixed specimens, the agreement was higher in almost all diagnostic categories in formalin-fixed biopsies. Overall, interobserver concordance was improved after a consensus/training session. Interobserver reproducibility was high at the

☆ Disclosures: All authors do not have conflict of interest or source of funding to declare.

* Corresponding author. Department of Laboratory Medicine and Pathology, University of Washington School of Medicine, NE140D, Seattle, WA, USA
E-mail address: myeh@uw.edu (M.M. Yeh).

https://doi.org/10.1016/j.humpath.2020.10.003

end of the diagnostic spectrum (CA) but fair to moderate for other diagnostic categories.

1. Introduction

Biliary strictures are a common diagnostic challenge in clinical practice. The most critical decision to guide clinical management is differentiating benign from malignant epithelial alterations. The majority of biliary strictures are malignant, primarily due to cholangiocarcinoma, pancreatic ductal adenocarcinoma, or ampullary/periampullary carcinoma [1]. Less common malignant causes include gallbladder adenocarcinoma, hepatocellular carcinoma, lymphoproliferative disorders, and metastatic disease [1]. Up to 30% of biliary strictures are associated with benign conditions, such as primary sclerosing cholangitis, IgG4-related sclerosing cholangitis, gallstones (Mirizzi syndrome or inflammatory stricture), chronic pancreatitis, iatrogenic bile duct injury after cholecystectomy, post-transplant strictures, and other less common conditions [1]. Studies have shown that as many as 25% of patients undergoing surgery for suspected malignant strictures turned out to have benign pathology [2–4]. Surgery for biliary strictures is associated with appreciable postoperative morbidity. Therefore, confirming diagnosis is crucial before considering aggressive surgical treatment.

Endoscopic retrograde cholangiopancreatography (ERCP) is usually performed for the evaluation of indeterminate biliary strictures. It allows tissue sampling by brushing cytology and/or intraductal biopsy. Brush cytology has a high specificity of nearly 100%, but the overall sensitivity and negative predictive value are only about 60% [5,6]. Intraductal biopsy was introduced to increase the detection of biliary malignancies. Compared with brush cytology, intraductal biopsy has been studied less extensively, and its sensitivity and specificity remain controversial [5,7,8]. Interpretation of bile duct biopsies can be challenging for various reasons. For instance, intraductal biopsy specimens are usually small. Fragmentation and superficial nature of the tissue may preclude complete evaluation of architectural features. A cautery or crush artifact may pose difficulties for interpretation. Moreover, in contrast to other common gastrointestinal (GI) disorders, such as Barrett esophagus and inflammatory bowel disease, reproducibility of morphologic features and diagnostic criteria of precancerous lesions of the biliary tract are relatively less well studied in biopsies. Misinterpretation by the pathologist is one of the common causes of incorrect diagnosis, which includes overinterpretation of low- or high-grade dysplasia and misinterpretation of reactive atypia in primary sclerosing cholangitis, choledocholithiasis, pancreatitis, reactive papillary changes, degenerative changes, intestinal metaplasia, and inflammatory/reactive changes from stenting or postsurgical stenosis [1].

In our practice, we routinely seek consensus when the diagnosis of dysplasia or carcinoma is considered for an intraductal biopsy. Through this informal consensus process, we have recognized that interobserver agreement is only fair in differentiating atypia from dysplasia and low-grade from high-grade dysplasia. The aims of this study were therefore to formally assess interobserver concordance in the interpretation of intraductal biopsies and to identify histologic features that lead to disagreement among pathologists.

2. Materials and methods

2.1. Case selection and slide review

The pathology database at the University of Washington was searched for ERCP-based intraductal forceps biopsies for biliary strictures from 2004 to 2018. More than 300 bile duct biopsies were identified, and we selected 85 cases for this study after careful re-review of the original pathology reports and slides. To be representative, we attempted to select an equal or a similar number of cases for each diagnostic category, although there were more cases diagnosed with negative for dysplasia or carcinoma. Biopsies suboptimal for evaluation were excluded, including specimens with insufficient tissue for assessment. Some specimens contained necrotic debris or fibroinflammatory material only without an epithelial component, which were also excluded. The original sign-out diagnoses for the 85 cases that were selected for this study included the following: negative for dysplasia (NED; n = 20), indefinite for dysplasia (IND; n = 20), low-grade dysplasia (LGD; n = 13), high-grade dysplasia (HGD; n = 15), and carcinoma (CA; n = 17). Few cases were descriptive in original diagnoses (eg, atypical, favor reactive changes). A specific category (eg, NED or IND) was assigned to these cases as part of case selection. Forty of these biopsies were processed using Hollande's fixative (2004–2010), whereas the remaining 45 were processed using 10% formalin fixative (2011–2018). The most representative slide from each of the cases was randomly renumbered from 1 through 85, and the original slide label was covered with non-transparent adhesive labels bearing the newly assigned number. Areas of interest that were previously marked on the slides were removed. Clinical information was not provided to the reviewers. Cases with insufficient tissue for histologic evaluation were excluded on re-review.

Eight pathologists participated in the slide review, including 6 pathologists whose primary clinical activity is participating in the University of Washington Medicine GI/hepatic/pancreatic pathology specialty sign-out service (X.G., F.J., L.K., P.E.S., C.D.T., and M.M.Y.) and 2 pathologists who were GI/liver pathology fellow trainees at the time of the study (Y.-J.L. and J.R.). The practicing experience of the 6 GI/liver pathologists ranges from 4 to 30 years (mean: 12.3 years). All of the 8 reviewers were familiar with the diagnostic criteria for dysplasia and carcinoma of the pancreatobiliary system as elaborated in the current World Health Organization (WHO) Classification of Tumors [9]. These criteria were used as a general basis for their interpretation of the biopsies in the first round. Without prior discussion of diagnostic criteria, the 85 glass slides and a scoring worksheet were circulated to the participating pathologists. Each pathologist was asked to assign a diagnosis by checking a box on the worksheet. The diagnostic choices were NED, IND, LGD, HGD, and CA. Definite classification was problematic for several cases, but the participating reviewers agreed to assign the most likely diagnostic choice for each of them.

2.2. Tutorial training

After the first round, a tutorial session was taken by each participating pathologist before the second round, organized by the lead investigator (M.M.Y.). Representative cases and images that cover all categories of diagnoses were reviewed during these sessions. The classification criteria were discussed and agreed upon by the participants and were applied to the second round of review on the original set of 85 slides. The only change in the second round was that the slides were again randomly renumbered so that the cases would be reviewed in a different order. The time interval between the first round review and tutorial training was about 3 months. Table 1 lists the criteria that were reviewed and agreed at consensus for NED, IND, LGD, HGD, and CA. The participating pathologists initiated the second round review at least 3 months after the tutoring training, ranging from 3 to 6 months.

All reviewers were blinded to clinical history of all cases for the first and second rounds of review.

2.3. Statistical analysis

General data handling and management was performed using R. Specific analyses related to interobserver agreement, including Fleiss κ [10] and S∗ statistic [11], were performed using an R package, *raters* (R package version 2.0.1, https://cran.r-project.org/web/packages/raters/index.html). The S∗ statistic was proposed as a modification of Fleiss κ in cases of nominal and ordinal variables [11], which was calculated as an index of inter-rater agreement among a set of raters using linear weights. The significance

Table 1 Criteria for grading dysplasia and carcinoma in intraductal biopsies (adapted from the studies by Zen et al [13,14]).

Negative for dysplasia (NED)	The mucosal architecture is preserved. Nuclear sizes and shapes are relatively uniform without nuclear pleomorphism. No increased nuclear-to-cytoplasm (N/C) ratio. Smooth nuclear membrane. Preservation of nuclear polarity. Nucleoli are not enlarged. Reactive or reparative features may be seen associated with inflammation, erosion, ulceration or changes from stents or postsurgical stenosis. In these settings, nuclear membranes should remain smooth, although the cells may display N/C enlargement and nucleoli may become more prominent but retain smooth contours.
Indefinite for dysplasia (IND)	Mucosal architecture may be mildly or moderately distorted. There are subtle cytological abnormalities such as mild nuclear hyperchromasia, slight irregularities of nuclear membrane and slight nuclear elongation or enlargement. There is minimal or mild variation in nuclear size or shape, but they do not vary significantly. Nuclear abnormalities are less marked than those seen in dysplasia. Diagnosis of IND is reserved for the situations when the changes are insufficient for the diagnosis of dysplasia but too significant for negative.
Low-grade dysplasia (LGD)	Flat, micropapillary, pseudopapillary or papillary architecture. There is nuclear pseudostratification, varying from lower two third of the epithelium to reaching the luminal surface. Focal loss of cellular polarity, but not a diffuse feature. Cytologically, there is nuclear abnormality, including nuclear hyperchromasia, elongation or enlargement. Variations in nuclear size and shape are present. Nuclear membrane is irregular. N/C ratio is increased. Mitoses are present but rare.
High-grade dysplasia (HGD)	Pseudopapillary, micropapillary or papillary architecture, and only rarely flat. There is architectural complexity. Cellular polarity diffusely and severely distorted with nuclei reaching and piling on the luminal surface. ‘Budding off’ of small clusters of epithelial cells into the lumen and cribriforming can be seen. There are cytologically malignant features with severe nuclear membrane irregularities, hyperchromasia or abnormally large nuclei. Cytologically resemble carcinoma, but invasion through the basement membrane are absent. Occasional mitoses are observed.
Carcinoma (CA)	Infiltrative tumor glands or large bizarre tumor cells with desmoplastic stromal response. Lymphovascular or perineural invasion may be present.

Table 2 Interobserver agreement in the diagnosis of dysplasia in the first round and second round review.

Diagnostic category	First round review						Second round review					
	Entire sample (n = 85)		Hollande's fixed (n = 40)		Formalin-fixed (n = 45)		Entire sample (n = 85)		Hollande's fixed (n = 40)		Formalin-fixed (n = 45)	
	Fleiss κ	S*	Fleiss κ	S*	Fleiss κs	S*	Fleiss κ	S*	Fleiss κ	S*	Fleiss κ	S*
Overall	0.39 [0.37−0.42]	0.56 [0.50−0.63]	0.28 [0.25−0.31]	0.47 [0.37−0.57]	0.49 [0.45−0.51]	0.66 [0.59−0.73]	0.48 [0.46−0.50]	0.69 [0.42−0.65]	0.43 [0.40−0.46]	0.68 [0.62−0.74]	0.53 [0.49−0.55]	0.71 [0.65−0.76]
NED	0.47 [0.43−0.51]	0.50 [0.40−0.60]	0.35 [0.29−0.41]	0.36 [0.23−0.50]	0.59 [0.53−0.65]	0.64 [0.50−0.78]	0.47 [0.42−0.51]	0.53 [0.44−0.63]	0.38 [0.32−0.44]	0.44 [0.31−0.57]	0.56 [0.50−0.61]	0.63 [0.48−0.77]
IND	0.16 [0.12−0.20]	0.51 [0.42−0.61]	0.09 [0.03−0.14]	0.44 [0.31−0.56]	0.25 [0.19−0.30]	0.60 [0.46−0.74]	0.33 [0.28−0.36]	0.50 [0.40−0.59]	0.30 [0.24−0.35]	0.41 [0.27−0.53]	0.34 [0.28−0.40]	0.59 [0.46−0.72]
LGD	0.36 [0.32−0.40]	0.61 [0.51−0.71]	0.11 [0.06−0.17]	0.60 [0.48−0.73]	0.49 [0.43−0.55]	0.63 [0.48−0.76]	0.49 [0.45−0.53]	0.71 [0.61−0.80]	0.36 [0.30−0.41]	0.73 [0.60−0.86]	0.56 [0.50−0.61]	0.69 [0.56−0.82]
HGD	0.28 [0.24−0.32]	0.69 [0.60−0.77]	0.23 [0.17−0.28]	0.68 [0.57−0.80]	0.33 [0.27−0.39]	0.70 [0.57−0.81]	0.40 [0.36−0.44]	0.63 [0.54−0.72]	0.47 [0.42−0.53]	0.77 [0.64−0.89]	0.32 [0.27−0.38]	0.69 [0.56−0.81]
CA	0.65 [0.62−0.70]	0.83 [0.75−0.90]	0.62 [0.56−0.70]	0.82 [0.70−0.91]	0.69 [0.63−0.75]	0.85 [0.74−0.94]	0.80 [0.76−0.84]	0.91 [0.85−0.96]	0.83 [0.78−0.89]	0.93 [0.85−1.00]	0.79 [0.72−0.84]	0.89 [0.80−0.96]
(CA + HGD) in one category	0.54 [0.50−.0.58]	0.64 [0.55−0.73]	0.47 [0.41−0.52]	0.59 [0.46−0.72]	0.61 [0.54−0.66]	0.68 [0.55−0.80]	0.73 [0.68−0.76]	0.78 [0.69−0.86]	0.77 [0.72−0.83]	0.83 [0.73−0.95]	0.67 [0.61−0.72]	0.73 [0.60−0.85]
(IND + LGD) in one category	0.22 [0.17−0.25]	0.28 [0.19−0.37]	0.05 [0.02−0.11]	0.18 [0.08−0.30]	0.35 [0.29−0.41]	0.38 [0.24−0.52]	0.35 [0.30−0.38]	0.36 [0.26−0.45]	0.29 [0.23−0.35]	0.31 [0.19−0.43]	0.39 [0.34−0.45]	0.41 [0.28−0.56]

NOTE. Interobserver agreement was measured in Fleiss κ and S* values, with 95% confidence interval in brackets. The level of interobserver agreement was graded based on the Landis and Koch scale (<0, poor agreement; 0.01−0.20, slight agreement; 0.21−0.40, fair agreement; 0.41−0.60, moderate agreement; 0.61−0.80, substantial agreement; and >0.80, nearly perfect agreement).
Abbreviations: NED, negative for dysplasia; IND, indefinite for dysplasia; LGD, low-grade dysplasia; HGD, high-grade dysplasia; CA, carcinoma.

level of Fleiss κ was calculated using the chi-square test, and the one for S∗ statistic was computed using a Monte Carlo algorithm with 1000 bootstrap simulations. The level of interobserver agreement was graded based on the Landis and Koch scale (<0, poor agreement; 0.01 to 0.20, slight agreement; 0.21 to 0.40, fair agreement; 0.41 to 0.60, moderate agreement; 0.61 to 0.80, substantial agreement; and >0.80, nearly perfect agreement) [10,12].

3. Results

3.1. Interobserver agreement: first round review

Table 2 illustrates the results (Fleiss κ and weighted S∗ values with 95% confidence interval) of interobserver agreement of the first and second rounds of review. In the entire sample (n = 85), the overall agreement was fair (κ = 0.39, S∗ = 0.56). Interobserver agreement was substantial for CA (κ = 0.65, S∗ = 0.83), moderate for NED (κ = 0.47, S∗ = 0.50), fair for HGD (κ = 0.28, S∗ = 0.69) and LGD (κ = 0.36, S∗ = 0.61), and slight for IND (κ = 0.16, S∗ = 0.51).

Reproducibility was also assessed using several clinically relevant grouping methods of diagnostic categories. When CA and HGD were grouped into one category at the high end of the spectrum, interobserver agreement was moderate (κ = 0.54, S∗ = 0.64). When IND and LGD were grouped in one category, interobserver agreement was fair (κ = 0.22, S∗ = 0.28).

Given the potential impact of alternative tissue fixation methods on histologic interpretation, interobserver agreement analysis was performed separately on Hollande's fixed specimens (n = 40) and formalin-fixed specimens (n = 45). Notably, interobserver agreement improved in almost all diagnostic categories in formalin-fixed specimens as compared with that of Hollande's fixed specimens, with most remarkable improvement made in the categories of LGD, IND, and NED.

3.2. Classification criteria for tutorial training

To better distinguish between different diagnostic categories, histologic features of each category were discussed and further clarified based on the current WHO Classification of Tumors and published literature [9,13–15]. Morphologic features used for classification include the following: (1) architecture of the glands; (2) cytologic featsures of the proliferating cells; (3) mitosis; and (4) associated inflammation, erosion, and ulceration. Fig. 1(A–E) shows the characteristic features of each diagnostic category. For this study, LGD is defined by features of biliary intraepithelial neoplasia (BillN)-1 and BillN-2 or intraductal papillary lesions showing low-grade dysplasia. HGD is defined by features of BillN-3 or intraductal papillary lesions showing high-grade dysplasia. The second round of blind scoring was conducted after multihead microscope training sessions.

3.3. Interobserver agreement: second round review

Compared with the first round review, interobserver agreement in the second round review improved in almost all the diagnostic categories (Table 2). In the entire sample, the overall agreement improved from fair to moderate ($\kappa = 0.48$, $S* = 0.69$). Interobserver agreement was nearly perfect for CA ($\kappa = 0.80$, $S* = 0.91$), fair for HGD ($\kappa = 0.40$, $S* = 0.63$), moderate for LGD ($\kappa = 0.49$, $S* = 0.71$) and NED ($\kappa = 0.47$, $S* = 0.53$), and fair for IND ($\kappa = 0.33$, $S* = 0.50$).

When CA and HGD were grouped into one category, interobserver agreement was substantial ($\kappa = 0.73$, $S* = 0.78$). When IND and LGD were grouped in one category, interobserver agreement was slight ($\kappa = 0.35$, $S* = 0.36$).

Separate analysis of Hollande's fixed specimens and formalin-fixed specimens showed a similar pattern as observed in the first round review. Agreement in formalin-fixed specimens remained higher in most diagnostic categories except for HGD, CA and grouping of HGD with CA.

The potential impact of experience of the participating pathologists (denoted by A, B, C, D, E, F, G, and H) was analyzed. Table 3 illustrates all κ value combinations with years of practice. Overall, interobserver agreement between the participating pathologists was fair to moderate. The performance of the two fellows was comparable with the other participating pathologists. The two pathologists with most years of experience had the highest agreement in diagnosis. However, correlation analyses showed no statistically significant association between high intraobserver agreement and yeas of practice. Although experience may partly explain the high agreement in diagnosis between the

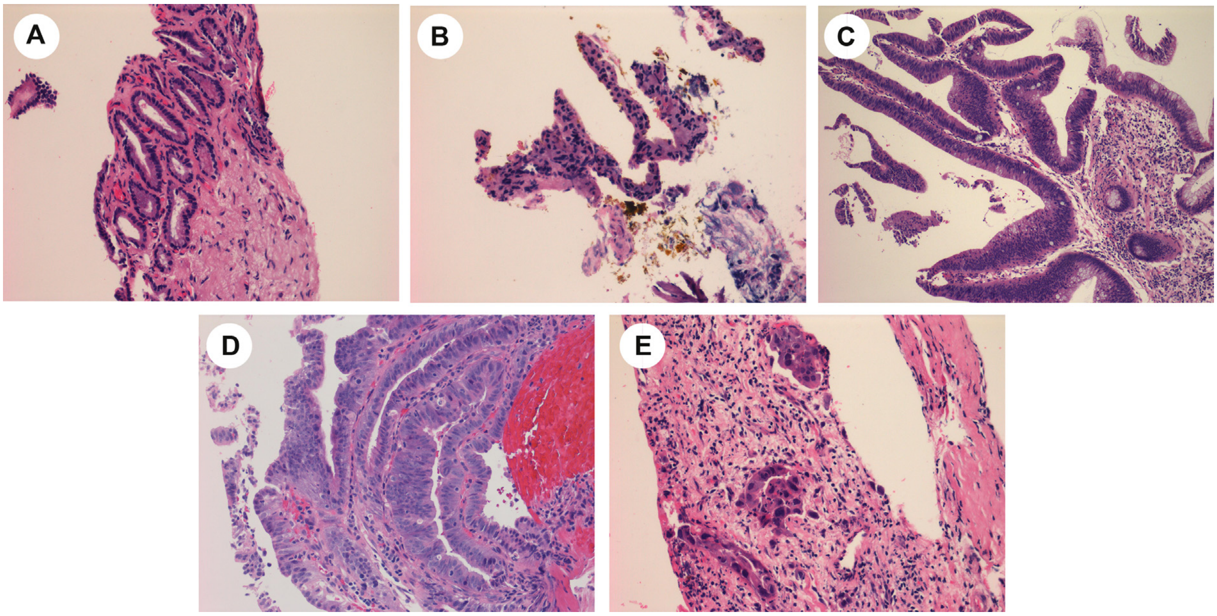

Fig. 1 **A, Negative for dysplasia (NED)**. Nuclei do not vary greatly in size or shape and are located basally. The N/C ratio is not increased. The nuclear envelope is generally smooth. The architecture is within normal limits. **B, Indefinite for dysplasia (IND)**. Subtle cytological abnormality. Nuclear sizes and shapes are relatively uniform. The architecture may be moderately distorted. The diagnosis is limited to cases in which the changes are too marked for negative but not sufficient for the diagnosis of dysplasia. **C, Low-grade dysplasia (LGD)**. This category covers BlIN-1 and BlIN-2. Nuclear pseudostratification, varying from lower two-thirds of the epithelium to reaching the luminal surface. Loss of cellular polarity is found, but it is not a diffuse feature. Mitoses are present but rare. **D, High-grade dysplasia (HGD)**. Cellular polarity is diffusely and severely distorted, with nuclei reaching and piling on the luminal surface. Cribriforming can be seen. Severe nuclear membrane irregularities, hyperchromasia, or abnormally large nuclei. Cytologically resembles carcinoma, but invasion through the basement membrane is absent. **E, Carcinoma (CA)**. Infiltrating single and small clusters of neoplastic cells.

two senior pathologists, long working relationship between them could also be a contributing factor.

3.4. Intraobserver agreement between the first round and second round review

Intraoberver agreement between the first round and second round review for all the participants (denoted by A, B, C, D, E, F, G, and H) was measured. Agreement of the 8 GI/liver pathologists was as follows: A (κ = 0.45; S∗ = 0.63), B (κ = 0.44; S∗ = 0.60), C (κ = 0.61; S∗ = 0.77), D (κ = 0.48; S∗ = 0.62), E (κ = 0.43; S∗ = 0.65), F (κ = 0.54; S∗ = 0.65), G (κ = 0.53; S∗ = 0.65), and H (κ = 0.78; S∗ = 0.82). It can be seen that intraobserver agreement was moderate for 6 participants and substantial for 2 participants.

3.5. Analysis of discordant cases

The scoring results were reviewed in a tabulated format to identify cases without majority consensus in which at least 4 of the participants rendered different diagnoses in the first round and/or second round review. Morphologic features of these cases were analyzed. Some of these situations, with relevant clinical outcome data, are exemplified in Figs. 2—5 and Supplement Fig. 1—11. Table 4 shows the number of cases in agreement among pathologists in the first round and second round review. The numbers in agreement in 8, 7, 6, or 5 participating pathologists in each diagnostic category are provided. Consistent with the Fleiss κ and S∗ values observed in interobserver agreement analyses, the second round review shows improvement in overall agreement, particularly in the categories of HDG, LGD, and IND. Of the total 85 cases, complete agreement (in all 8 pathologists) or near-complete agreement (in 7 pathologists) was achieved in 15 and 10 cases, respectively, in the first round review, and in 16 and 10 cases, respectively, in the second round review. The majority of these cases were in the CA and NED categories. The number of cases in which the two most experienced pathologists were not in agreement was 37 in the first round and 32 in the second round, with major discrepancies caused by differentiating NED from IND and LGD from HGD and few discrepancies caused by differentiating NED from LGD. There were few cases with less than 50% concurrence in diagnosis (8 cases in the first round and 6 cases in the second round), most falling into categories IND and LGD.

Common factors causing difficulties in differentiating dysplasia from nondysplasia include Hollande’s fixation, crush/cautery artifacts, detachment of the epithelium, entrapped atypical glands in the stroma, and active inflammation. Distinction between LGD and HGD can sometimes be challenging owing to Hollande’s fixation, the micropapillary pattern of LGD, and the pathologist’s threshold of calling HGD, which could be somewhat subjective despite reference to established interpretative criteria and (before the second round review) a training tutorial re-emphasizing those criteria. Diagnosis of carcinoma is relatively straightforward if infiltrative glands or tumor cells are present in desmoplastic stroma. However,

Table 3 κ value combinations with years of practice of the participating pathologists.

Pathologist (years of practice)	A (<10 yrs)	B (>10 yrs)	C (>10 yrs)	D (<10 yrs)	E (>10 yrs)	F (>10 yrs)	G (<10 yr)	H (<10 yr)
A (<10 yrs)	-	0.30	0.50	0.27	0.45	0.44	0.50	0.36
B (>10 yrs)	0.28	-	0.31	0.34	0.37	0.32	0.31	0.33
C (>10 yrs)	0.41	0.43	-	0.18	0.55	0.48	0.46	0.44
D (<10 yrs)	0.54	0.48	0.37	-	0.31	0.24	0.28	0.29
E (>10 yrs)	0.51	0.21	0.60	0.40	-	0.51	0.49	0.38
F (>10 yrs)	0.40	0.41	0.45	0.57	0.43	-	0.47	0.36
G (<10 yr)	0.35	0.41	0.36	0.51	0.33	0.42	-	0.37
H (<10 yr)	0.28	0.40	0.56	0.33	0.29	0.34	0.38	-

Eight participating pathologists are denoted by A, B, C, D, E, F, G, and H.
κ values in the white zone are those of first round review; κ values in the gray zone are those of the second round review.

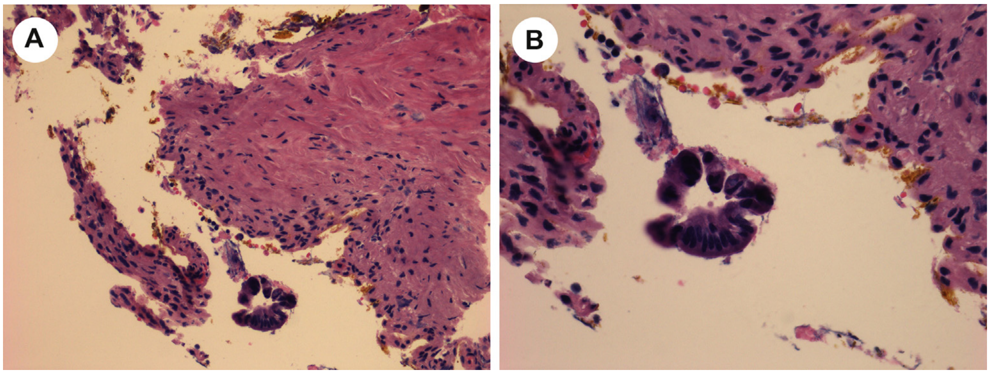

Fig. 2 First round: IND, 4; NED, 3; LGD, 1. Second round: IND, 6; NED, 2. This case shows detached strips of columnar epithelium with mild cytological atypia including nuclear hyperchromasia and elongation. The changes are insufficient for the diagnosis of LGD or HGD, but are too prominent for NED, better interpreted as IND. The patient was followed up clinically, with a final diagnosis leading to chronic cholecystitis. LGD, low-grade dysplasia; HGD, high-grade dysplasia; NED, negative for dysplasia; IND, indefinite for dysplasia.

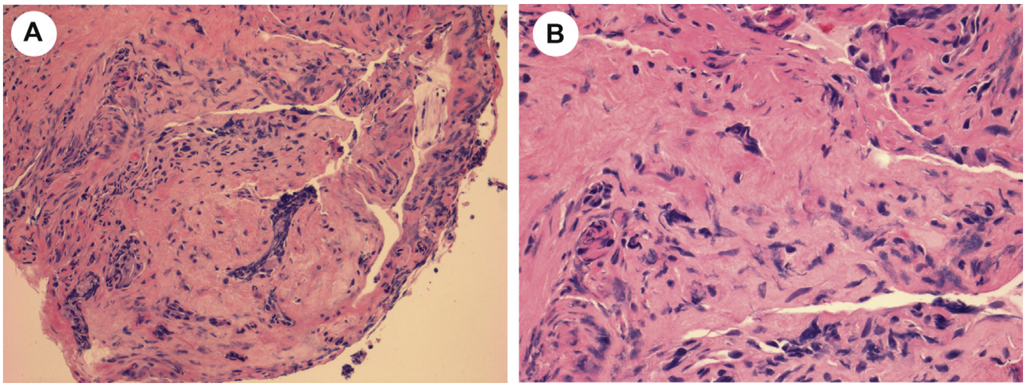

Fig. 3 First round: IND, 2; NED, 5; CA, 1. Second round: IND, 3; NED, 5. This biopsy features the crush/cautery artifact. Rare entrapped epithelial cells with irregular nuclear membrane are present in the fibrous stroma. On high magnification, these cells appear to retain a normal N/C ratio. Given the crush/cautery artifact, IND is favored, although NED is also an acceptable diagnosis with appropriate comments. Rebiopsy was recommended. The patient was followed up via rebiopsy, which turned out to be invasive cholangiocarcinoma. CA, carcinoma; NED, negative for dysplasia; IND, indefinite for dysplasia; N/C, nuclear-to-cytoplasm.

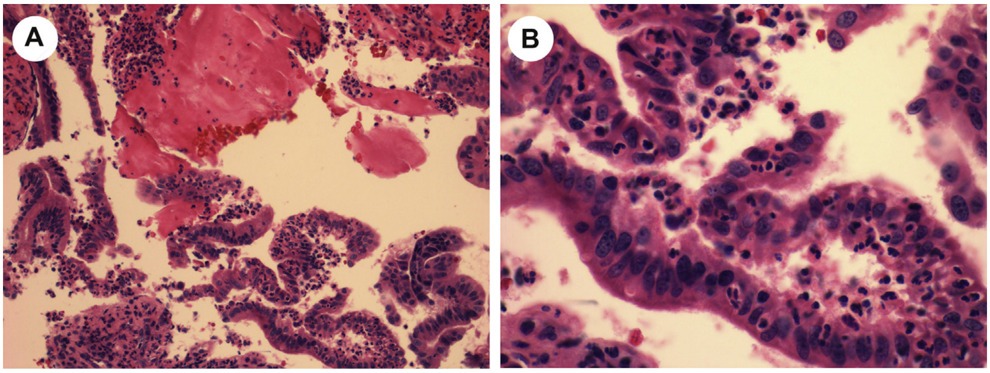

Fig. 4 First round: IND, 4; LGD, 1; HGD, 1; NED, 2. Second round: IND, 6; NED, 1; LGD, 1. Multiple detached superficial fragments of columnar epithelium are present with nuclear hyperchromasia and cytological atypia. Given the marked neutrophilic infiltrate in the surface epithelium, the changes favor reactive atypia associated with active inflammation. The patient had hepatic venous outflow obstruction treated with a TIPS procedure, and no carcinoma was developed in clinical follow-up. IND, indefinite for dysplasia; LGD, low-grade dysplasia; HGD, high-grade dysplasia; NED, negative for dysplasia; TIPS, transjugular intrahepatic portosystemic shunt.

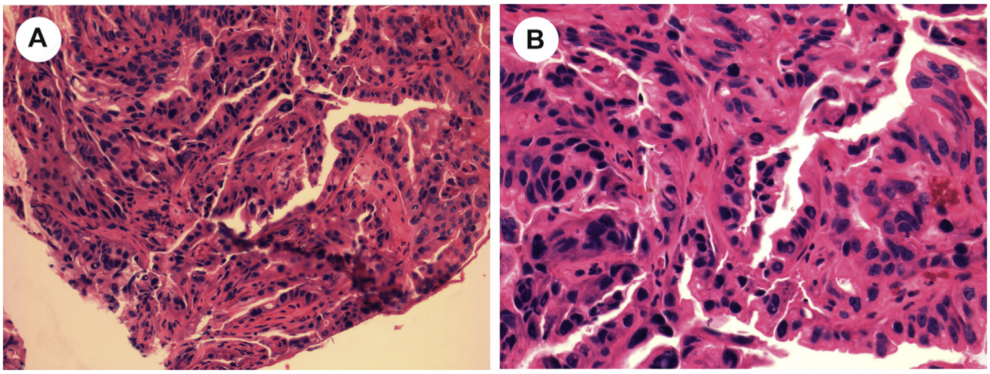

Fig. 5 First round: HGD, 3; CA, 3; LGD, 2. Second round: HGD, 4; CA, 4. The lesion features crowding of neoplastic glands, highly atypical epithelial cells, and loss of nuclear polarity. Most reviewers agreed the changes are those of at least HGD. Lack of stroma prevents evaluation of invasion. The patient had resected Klatskin perihilar cholangiocarcinoma. HGD, high-grade dysplasia; CA, carcinoma; LGD, low-grade dysplasia.

Table 4 The number of cases in agreement among pathologists in the first round and second round review.

Diagnostic category	Agreement in 8 pathologists	Agreement in 7 pathologists	Agreement in 6 pathologists	Agreement in 5 pathologists	Agreement in at least 5 pathologists
First round					
CA	3	3	3	2	11
HGD	0	0	4	1	5
LGD	3	1	1	4	9
IND	0	0	0	5	5
NED	9	6	6	6	27
Total	15	10	14	18	57
Second round					
CA	4	3	3	3	13
HGD	0	1	3	4	8
LGD	4	1	3	3	11
IND	0	1	0	5	6
NED	8	4	7	4	23
Total	16	10	16	19	61

Abbreviations: NED, negative for dysplasia; IND, indefinite for dysplasia; LGD, low-grade dysplasia; HGD, high-grade dysplasia; CA, carcinoma.

rare single tumor cells can be missed on low magnification, especially when the crush/cautery artifact presents.

4. Discussion

To our knowledge, this is the first study evaluating interobserver concordance in the diagnosis of intraductal biopsies for biliary strictures. Eight pathologists from the same institution participated in the slide review, including 6 GI pathologists with at least 4 years of practicing experience and 2 current GI/liver pathology fellows. The overall interobserver agreement in the first round review was fair, reaffirming the challenges in histologic interpretation of intraductal biopsies. The agreement in the second round review improved in almost all the diagnostic categories. Our data suggest that discussion and further clarification of diagnostic criteria could improve concordance among reviewers. It would be ideal to test the impact of training on a new set of biopsies in addition to the ones that had already been reviewed. This was not performed in the present study owing to the limitations in obtaining sufficient numbers of intraductal biopsies with optimal nature in each diagnostic category.

Approximately half of the cases were Hollande's fixed specimens. In comparison with formalin-fixed slides, interobserver agreement was lower in almost all the categories before and after the consensus meeting. This observation stands in stark contrast to the original reason for using Hollande's fixative that the improved nuclear detail in picric acid mordants was intended to improve recognition and interpretation of nuclear detail. This may be simply because some of the participants were not familiar with the historical Hollande's fixed material and found them difficult to interpret. The nuclei appear to be larger and more prominent on Hollande's fixed slides than on formalin-fixed slides, which resulted in misinterpretation or overinterpretation of dysplasia. However, some ambiguities may be resolved by combining with other features such as the nuclear-to-cytoplasm ratio, nuclear polarity, growth pattern, and architecture of the glands. In fact, considerable improvement in agreement was achieved after the consensus meeting in almost all the diagnostic categories. In particular, a higher κ grade was achieved for IND, LGD, HGD, CA, and grouping categories.

Compared with Barrett esophagus and inflammatory bowel disease, reproducibility of morphologic features of dysplasia/neoplasia of the biliary tract in biopsy specimens is relatively less well studied, although the diagnostic criteria have been outlined in the WHO Classification of Tumors [9] and the literature [13,14]. This is partly because malignancy of the biliary tract is less common. The difficulty of accessing the biliary tract is also a contributing factor. Barrett esophagus and inflammatory bowel disease have well-established morphologic criteria for grading dysplasia, and there are effective surveillance mechanisms in the evaluation of the risk of progression from dysplasia to invasive carcinoma. Reproducibility of the diagnostic criteria has been extensively evaluated in biopsies [16−21]. However, such data are limited for the biliary tract. Two major forms of precancerous lesions of the biliary tract have been described lately, flat biliary dysplasia (BillN) and intraductal papillary neoplasms, analogous to pancreatic intraepithelial neoplasia and intraductal papillary mucinous neoplasm of the pancreas. Early studies reported moderate interobserver agreement on the proposed morphologic criteria of the three-grade classification system of BillN [13,14]. The data were based on the resection and explant material of patients with primary sclerosing cholangitis, choledochal cyst, and hepatolithiasis [13,14]. Reproducibility of these criteria in intraductal biopsies has not been evaluated and validated.

Biliary strictures are caused by a variety of benign and malignant disorders, including intrabiliary and extrabiliary tract lesions or conditions, which pose difficulties and add complexity to the histologic interpretation of biopsies. To better distinguish between different categories, we designed a classification system for dysplasia grading, which incorporates cytologic features, architecture, mitosis, associated inflammation, erosion, and ulceration. In particular, the morphologic features of IND, LGD, and HGD were clearly defined based on the current WHO classification [22]. BillN-1 and BillN-2 were included in the LGD category. Distinction between LGD and HGD largely relies on the degree of cytologic atypia and architectural complexity, which can be challenging as it is often a subtle microscopic difference, and evaluation may be affected by the reviewer's experience. This partly explains the fair interobserver agreement for LGD and HGD in the entire sample in the first round review. Despite further clarification and elaboration of the criteria in the session before the second round, the agreement was only moderate in the second round review.

The lowest agreement was achieved for IND, slight in the entire sample in the first round review and only fair after consensus meeting. The agreement was particularly low in Hollande's fixed slides. The agreement was fair when grouping IND and LGD into one category, although there was improvement after the consensus meeting. By definition, IND is defined by subtle or mild cytologic and/or architectural abnormalities and is reserved for the situations when the changes are insufficient for dysplasia but too significant for NED. Surface maturation remains a key feature of IND and for its distinction from LGD. However, separation of IND and LGD can be difficult, especially when there is coexisting active inflammation, erosion, or ulceration. Interestingly, poor to fair intraobserver and interobserver agreement for IND ($\kappa = 0.17$ and 0.25) was also reported in a recent study assessing dysplasia in ampullary mucosal biopsies even when clinical information was provided, reaffirming the challenges in reaching consensus in this diagnostic category [23]. Diagnosis of CA had substantial to nearly perfect agreement before and after the consensus meeting. High reproducibility in this category has clinical significance as surgical treatment is often the management plan.

Of note, clinical information was not provided to the reviewers in our study to avoid bias and its potential impact on reproducible histologic/morphologic interpretation, although we acknowledge clinical information would be helpful in real practice. In real practice, ancillary tests such as p53, Ki-67, and K-RAS immunohistochemical stains are sometimes used to help differentiate CA and HGD or CA/HGD and a marked reactive process. To our knowledge, there are still conflicting data about using these stains to make the diagnosis [24−26]. For instance, studies found that adding the information of p53 and K-RAS analysis did not improve the diagnosis in pancreatobiliary specimens [24,25]. A study reported low sensitivity of p53 immunohistochemical staining in diagnosis of cholangiocarcinoma, which limits its diagnostic utility [26]. Ki-67 was reported to facilitate distinction between benign and malignant lesions of the biliary duct; however, a low proliferation index (5−10%) cannot entirely exclude adenocarcinoma [26]. In our practice, we do not routinely use these ancillary tests, and thus, we have a limited number of cases to assess their value in diagnosis. Nevertheless, we acknowledge appropriate use and interpretation of these ancillary tests, in correlation with morphologic findings and the clinical setting, may help improve accurate diagnosis.

In this study, Fleiss κ statistics and a weighted interobserver concordance measure, $S*$ statistic, were used to assess interobserver concordance. Fleiss κ is widely used for measuring interobserver agreement in the literature. A potential limitation of κ is that it does not account for the graded scale of scoring (ordinal data). Our grade scoring data are ordinal data, which means a higher category has a higher grade of dysplasia. S∗ statistic, by accounting for a graded scoring scale, offers an additional useful parameter for agreement assessment. Interestingly, in our results (Table 2), S∗ statistic scores are higher than Fleiss κ scores for almost all interobserver agreement assessments (with a few exceptions of equal S∗ statistic and Fleiss κ scores), which testified the advantage of the S∗ statistic in capturing the interobserver agreement for ordinal data over the traditional Fleiss κ score. However, as a limitation with S∗measures, currently, there has been no well-accepted grading system for S∗ measures as widely adopted for Fleiss κ [27].

In summary, this single-center study highlights the challenges in histologic evaluation of intraductal biopsies for biliary strictures. The common morphologic features and factors that lead to interobserver disagreement were described and discussed. Out data suggest that seeking second opinion, obtaining consensus among groups, and gathering previous cases for the training set among groups in academic and/or community practice would be helpful to improve our agreement in clinical practice. Given that pathologic diagnosis of biliary dysplasia/neoplasia is critical to subsequent clinical therapy, further expanded studies involving more pathologists and other institutions are warranted to better define the diagnostic criteria of biliary dysplasia/neoplasia and to improve reproducibility of diagnosis.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.humpath.2020.10.003.

References

[1] Bowlus CL, Olson KA, Gershwin ME. Evaluation of indeterminate biliary strictures. Nat Rev Gastroenterol Hepatol 2016;13:28−37.

[2] Corvera CU, Blumgart LH, Darvishian F, Klimstra DS, DeMatteo R, Fong Y, et al. Clinical and pathologic features of proximal biliary strictures masquerading as hilar cholangiocarcinoma. J Am Coll Surg 2005;201:862−9.

[3] Clayton RA, Clarke DL, Currie EJ, Madhavan KK, Parks RW, Garden OJ. Incidence of benign pathology in patients undergoing hepatic resection for suspected malignancy. Surgeon 2003;1:32−8.

[4] Gerhards MF, Vos P, van Gulik TM, Rauws EA, Bosma A, Gouma DJ. Incidence of benign lesions in patients resected for suspicious hilar obstruction. Br J Surg 2001;88:48−51.

[5] Navaneethan U, Njei B, Lourdusamy V, Konjeti R, Vargo JJ, Parsi MA. Comparative effectiveness of biliary brush cytology and intraductal biopsy for detection of malignant biliary strictures: a systematic review and meta-analysis. Gastrointest Endosc 2015;81: 168−76.

[6] Moreno Luna LE, Kipp B, Halling KC, Sebo TJ, Kremers WK, Roberts LR, et al. Advanced cytologic techniques for the detection of malignant pancreatobiliary strictures. Gastroenterology 2006;131: 1064−72.

[7] Ponchon T, Gagnon P, Berger F, Labadie M, Liaras A, Chavaillon A, et al. Value of endobiliary brush cytology and biopsies for the diagnosis of malignant bile duct stenosis: results of a prospective study. Gastrointest Endosc 1995;42:565−72.

[8] Schoefl R, Haefner M, Wrba F, Pfeffel F, Stain C, Poetzi R, et al. Forceps biopsy and brush cytology during endoscopic retrograde cholangiopancreatography for the diagnosis of biliary stenoses. Scand J Gastroenterol 1997;32:363−8.

[9] WHO classification of tumours of the digestive system. 5th ed. World Health Organization; 2019.

[10] Chmura Kraemer H, Periyakoil VS, Noda A. Kappa coefficients in medical research. Stat Med 2002;21:2109−29.

[11] Marasini D, Quatto P, Ripamonti E. Assessing the inter-rater agreement for ordinal data through weighted indexes. Stat Methods Med Res 2016;25:2611−33.

[12] McHugh ML. Interrater reliability: the kappa statistic. Biochem Med 2012;22:276−82.

[13] Zen Y, Aishima S, Ajioka Y, Haratake J, Kage M, Kondo F, et al. Proposal of histological criteria for intraepithelial atypical/proliferative biliary epithelial lesions of the bile duct in hepatolithiasis with respect to cholangiocarcinoma: preliminary report based on interobserver agreement. Pathol Int 2005;55:180−8.

[14] Zen Y, Adsay NV, Bardadin K, Colombari R, Ferrell L, Haga H, et al. Biliary intraepithelial neoplasia: an international interobserver agreement study and proposal for diagnostic criteria. Mod Pathol 2007;20:701−9.

[15] Torbenson M, Zen Y, Yeh MM. Tumors of the liver, AFIP Atlas of tumor pathology series 4. American Registry of Pathology; 2018.

[16] Montgomery E, Bronner MP, Goldblum JR, Greenson JK, Haber MM, Hart J, et al. Reproducibility of the diagnosis of dysplasia in Barrett esophagus: a reaffirmation. Hum Pathol 2001;32:368−78.

[17] Kerkhof M, Kusters JG, van Dekken H, Kuipers EJ, Siersema PD. Biomarkers for risk stratification of neoplastic progression in Barrett esophagus. Cell Oncol 2007;29:507−17.

[18] Salomao MA, Lam-Himlin D, Pai RK. Substantial interobserver agreement in the diagnosis of dysplasia in Barrett esophagus upon review of a patient's entire set of biopsies. Am J Surg Pathol 2018;42:376−81.

[19] Allende D, Elmessiry M, Hao W, Dasilva G, Wexner SD, Bejarano P, et al. Inter-observer and intra-observer variability in the diagnosis of dysplasia in patients with inflammatory bowel disease: correlation of pathological and endoscopic findings. Colorectal Dis 2014;16:710−8.

[20] Odze RD, Goldblum J, Noffsinger A, Alsaigh N, Rybicki LA, Fogt F. Interobserver variability in the diagnosis of ulcerative colitis-associated dysplasia by telepathology. Mod Pathol 2002;15:379−86.

[21] Eaden J, Abrams K, McKay H, Denley H, Mayberry J. Inter-observer variation between general and specialist gastrointestinal pathologists when grading dysplasia in ulcerative colitis. J Pathol 2001;194: 152−7.

[22] Nagtegaal ID, Odze RD, Klimstra D, Paradis V, Rugge M, Schirmacher P, et al. The 2019 WHO classification of tumours of the digestive system. Histopathology 2020;76:182−8.

[23] Allard FD, Goldsmith JD, Ayata G, Challies TL, Najarian RM, Nasser IA, et al. Intraobserver and interobserver variability in the assessment of dysplasia in ampullary mucosal biopsies. Am J Surg Pathol 2018;42:1095−100.

[24] Ponsioen CY, Vrouenraets SM, van Milligen de Wit AW, Sturm P, Tascilar M, Offerhaus GJ, et al. Value of brush cytology for dominant strictures in primary sclerosing cholangitis. Endoscopy 1999;31:305−9.

[25] Stewart CJ, Burke GM. Value of p53 immunostaining in pancreaticobiliary brush cytology specimens. Diagn Cytopathol 2000;23:308−13.

[26] Tsokos CG, Krings G, Yilmaz F, Ferrell LD, Gill RM. Proliferative index facilitates distinction between benign biliary lesions and intrahepatic cholangiocarcinoma. Hum Pathol 2016;57:61−7.

[27] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159−74.