Bile Duct Brushing Cytology

Statistical Analysis of Proposed Diagnostic Criteria

Andrew A. Renshaw, MD,[1] Rebecca Madge, CT(ASCP),[1] Michael Jiroutek, MS,[2] and Scott R. Granter, MD[1]

Key Words: Bile duct brushings; ERCP; Neoplasms; Cytology; Statistics

Abstract

Recent studies based on multivariate analysis have identified cytologic features that may be of value in the diagnosis of malignancy in bile duct brushings. We sought to assess the reproducibility and accuracy of these criteria. Three different observers used 4 sets of criteria that included 9 cytologic features to review 165 bile duct brushing specimens with available follow-up data. An overall assessment of malignancy and evaluation for the presence of chromatin clumping had excellent reproducibility (3-way κ values of 0.708 and 0.629, respectively). Evaluation for the presence of enlarged nuclei, increased nuclear:cytoplasmic ratio, nuclear molding, and loss of honeycombing showed moderate reproducibility. An overall assessment of malignancy was a better predictor of the presence of malignancy than any other criteria, with a sensitivity of 36.2% and a specificity of 95%. Retrospective analysis demonstrated that the criteria of chromatin clumping, increased nuclear:cytoplasmic ratio, and either nuclear molding or loss of honeycombing had similar sensitivity and specificity as the overall assessment of malignancy (sensitivity = 35.2%, specificity = 95%). An overall assessment of malignancy or the criteria of chromatin clumping, increased nuclear:cytoplasmic ratio, and either nuclear molding or loss of honeycombing are reproducible cytologic criteria that accurately predict malignancy in bile duct brushings.

Examination of bile duct fluid is a relatively noninvasive method for establishing a diagnosis of malignancy. Initially, secreted or aspirated bile was the only material available for cytologic examination. Currently, however, bile duct brushings are the favored method of epithelial cell sampling. The reported sensitivity for this test ranges from 33% to 80%,[1–18] and the specificity for a definitive diagnosis of carcinoma in most older series is 100%,[1–10] with only rare false-positive cases[11–15] (reviewed by Kurzawinski and associates[19]). Sensitivity varies with the location of the tumor, and brushings appear to be more sensitive than either washings or direct bile sampling. Flow cytometry provides only modest improvement in diagnostic accuracy.[13]

A variety of nonmalignant conditions may be an indication for bile duct brushing cytology. These include stones, surgical stenosis, pancreatitis, pancreatic cysts, sclerosing cholangitis, cirrhosis, jaundice, hydatid cyst, and acute cholangitis.[15,9,20,21] None of these conditions have specific cytologic findings. Although pancreatitis may be associated with numerous degenerate cells,[22] it is a common finding in pancreatic and bile duct carcinomas as well. Complicating the diagnostic evaluation further is the fact that some of these conditions are well known to either predispose to or frequently coexist with malignancy.

In an effort to improve the diagnostic accuracy of cytologic examination of bile duct brushings, several investigators have proposed specific diagnostic criteria for malignancy. Initial series emphasized disoriented or crowded cells in 3-dimensional groups, extreme nuclear enlargement, and nuclear contour irregularity,[8] but rigorous statistical analysis was not performed. More recent series have used multivariate analysis to derive their criteria. One of the first of these reports (the Japan criteria), performed on bile specimens rather than brushings, examined 32 criteria and determined that loss of honeycomb arrangement, enlarged nuclei, loss of polarity, bloody background, flat nuclei, and cell-in-cell arrangement were highly associated with the correct diagnosis of malignancy.[20] Another recent study (the Iowa criteria) used multivariate logistic regression analysis on bile duct brushings to analyze 18 variables and concluded that the criteria of nuclear molding, chromatin clumping, and increased nuclear:cytoplasmic ratio resulted in a sensitivity of 83% and a specificity of 98%.[21]

Nevertheless, although multivariate analysis is helpful in the analysis of these specimens, it is not without limits. The greatest concern regarding this technique is that, because a very large number of cytologic features were tested, it is highly probable that some combination will result in an apparent improvement in diagnostic accuracy that may in fact represent a chance occurrence rather than a reproducible result. Thus, the true sensitivity and specificity of the derived criteria can only be determined by applying the criteria to another group of cases; the sensitivity and specificity in the original series will almost certainly be better than in another unbiased population. Second, multivariate analysis does not provide any evaluation of the reproducibility of the derived criteria. If the criteria are difficult to reproduce, the resultant sensitivity and specificity will be as well. Third, the criteria used in the 2 studies based on multivariate analysis are similar, though not identical, and the accuracy of the individual cytologic features in each series that were eventually selected as being of great diagnostic importance was often similar to that of other cytologic features that did not achieve statistical significance. In fact, when the data from these 2 large series[20,21] are combined, a third set of criteria, namely those of chromatin clumping, loss of polarity, and nuclear molding (or cell-in-cell arrangement) actually appears to be better than the criteria used in either of the original papers.

Finally, these criteria are primarily qualitative rather than quantitative. Although individual criteria contain quantitative information (ie, a 3-fold variation in nuclear size), the final criteria are either present or absent. Other series have instead emphasized a gradation of atypia, including architectural, nuclear, and nucleolar atypia, in which increased degrees of atypia increase the likelihood of malignancy.[15] Although this approach is probably more familiar to cytologists than is the use of qualitative criteria, it is more difficult to measure. No series has attempted to compare this approach to the criteria derived from multivariate analysis.

To evaluate these cytologic criteria, we prospectively compared the accuracy and reproducibility of 4 different sets of criteria in a series of 165 bile duct brushings.

Methods

Cases were derived from the files of the Division of Cytology of the Department of Pathology, Brigham & Women's Hospital, Boston, Mass, from 1990 to 1997. Biliary duct brush specimens were obtained endoscopically for all cases and were delivered to the laboratory either as prepared air-dried slides or in sterile solution. A total of 215 cases were received in the department during the aforesaid period; adequate follow-up could not be obtained in 38 cases, leaving 177 cases for review. One islet cell tumor, 1 mucinous cystic neoplasm, 1 serous cystadenoma, 5 metastatic carcinomas, and 4 non-Hodgkin lymphomas were also excluded. The remaining 165 cases form the basis of this report.

Of those 165 cases, 92 (56%) had histologic (87 cases) or cytologic (5 cases) confirmation, and 73 (44%) had clinical or radiologic follow-up or a combination thereof. The cytologic confirmation consisted of material obtained by a route other than bile duct brushing, most commonly a percutaneous fine-needle aspiration. A diagnosis of malignancy required either pathologic confirmation or a radiologic mass consistent with a primary lesion in either the pancreas or hepatobiliary tree along with radiologic evidence of either metastatic disease or tumor progression. Of the 160 patients, 105 had a diagnosis of malignancy and 60 had a benign diagnosis. Of the malignant cases, 44 patients had tumors of the biliary tree and 61 had pancreatic tumors. The benign cases included 22 patients with biliary stricture of unknown cause, 2 with postoperative strictures, 4 with radiation-induced strictures, 21 with primary sclerosing cholangitis alone, 3 with primary sclerosing cholangitis and a stent, 1 with primary sclerosing cholangitis and a stone, 1 with a stent for stricture of unknown cause, 2 with acute cholangitis, 2 with pancreatitis, and 2 with ampullary masses.

Material included alcohol-fixed, Papanicolaou-stained direct smears in 117 cases (usually 2 per case), cytocentrifuged preparations in 5 cases, 44 cases with Thinprep preparations (Cytyc Corp, Boxborough, Mass) (usually 1 per case), and 28 cases with H&E-stained cell block material. All cases were reviewed individually by 3 of the authors (A.A.R, R.M, S.R.G.) without knowledge of the clinical outcome.

The presence or absence of 9 separate cytologic features was evaluated in each case. These features were derived from and described and illustrated in previous publications.[15,20,21] The features included the following: nuclear molding, cell-in-cell arrangement, or a combination thereof; chromatin clumping; increased nuclear:cytoplasmic ratio; nuclei enlarged to at least 3-fold the size of a normal nucleus; loss of polarity; loss of honeycombing (disordered sheets); bloody background; and flat nuclei (interpreted to mean nuclei with flattened nuclear outlines). In addition, an overall assessment of malignancy based on the degree of atypia (architectural, nuclear, and nucleolar) was performed. A cell-in-cell arrangement was interpreted as an extreme example of nuclear molding, and these 2 features were grouped together. Loss of polarity and loss of honeycombing were related but different features. Specifically, loss of polarity could be assessed in single (1-dimensional) strips of

▮Table 1▮
Diagnostic Criteria for Bile Brushings

Name of Criteria	Criteria Content
Iowa*	Nuclear molding, chromatin clumping, and increased nuclear:cytoplasmic ratio
Japan†	3 or more of the following: loss of honeycombing, enlarged nuclei, loss of polarity, bloody background, flat nuclei, cell-in-cell arrangement
Boston	Chromatin clumping, loss of polarity, nuclear molding
Overall assessment of malignancy based on degree of atypia‡	Sufficient atypia to warrant a diagnosis of malignancy

*Cohen MB, Wittchow RJ, Johlin FC, et al. Brush cytology of the extrahepatic biliary tract: comparison of cytologic features of adenocarcinoma and benign biliary strictures. *Mod Pathol*. 1995;8:498–502.
†Nakajima T, Tajima Y, Sugano I, et al. Multivariate statistical analysis of bile cytology. *Acta Cytol*. 1994;38:51–55.
‡Layfield LJ, Wax TD, Lee JG, et al. Accuracy and morphologic aspects of pancreatic and biliary duct brushings. *Acta Cytol*. 1995;39:11–18.

▮Table 2▮
Three-way κ Analysis of Nine Cytologic Features in Bile Duct Brushings

Criteria	κ
Nuclear molding/cell-in-cell arrangement	0.414
Chromatin clumping	0.629
Increased nuclear:cytoplasmic ratio	0.503
Enlarged nuclei	0.569
Loss of polarity	0.204
Loss of honeycombing	0.587
Bloody background	0.191
Flat nuclei	0.082
Overall assessment of malignancy	0.708

▮Table 3▮
Sensitivity and Specificity of Nine Cytologic Features in Bile Duct Brushings for a Diagnosis of Malignancy

Criteria	Sensitivity (%)	Specificity(%)
Nuclear molding/cell-in-cell arrangement	35.2	83
Chromatin clumping	35.6	91.7
Increased nuclear:cytoplasmic ratio	41.0	88.3
Enlarged nuclei	48.6	68.3
Loss of polarity	36.2	96.7
Loss of honeycombing	51.0	73.3
Bloody background	15.2	88.3
Flat nuclei	6.7	100
Overall assessment of malignancy	36.2	95.0

cells and primarily represented a loss of the basal location of the nuclei. Loss of honeycombing could be assessed only in (2-dimensional) sheets of cells and consisted of a disorganized arrangement of nuclei in rows and columns. The overall assessment of atypia was somewhat vague by design and meant to test the observer's gestalt as to whether the degree of atypia was sufficient for a diagnosis of malignancy rather than objective assessment on any single or small set of features. These features were used to form 4 separate diagnostic criteria, detailed in **▮Table 1▮**. The Boston criteria shown in Table 1 were derived for this paper from data available in the Iowa and Japan papers.

Each observer's results were analyzed separately. In addition, the consensus of all 3 observers was also analyzed. Sensitivity and specificity were determined as usual. Reproducibility was determined by using the κ statistic, which was calculated between individual observers and as a 3-way analysis. Values of κ greater than 0.6 demonstrate excellent agreement, those between 0.4 and 0.6 demonstrate moderate agreement, and those less than 0.4 demonstrate poor agreement.

Results

Several observations were made when these slides were reviewed before any analysis was performed. All 3 observers believed that flat nuclei were poorly defined and difficult to identify. A bloody background could not be appreciated on Thinprep preparations. In addition, the material on Thinprep preparations was more atypical looking and appeared different than on the other preparations; cells were larger and more closely packed together, nuclei were larger and had more irregular borders, and nucleoli were more prominent. Also, all 3 observers agreed that fixation was critical. Air-drying artifact interfered with assessment of virtually all cytologic features, regardless of the diagnostic category of the specimen. Finally, several of the criteria appeared closely associated. For example, enlarged nuclei were often, though not always, associated with increased nuclear:cytoplasmic ratios. Similarly, loss of polarity was often, albeit not always, associated with loss of honeycombing. In fact, both enlarged nuclei and loss of polarity seemed to be less stringent versions of the criteria with which they were linked.

The results of κ analysis are summarized in **▮Table 2▮**. We also performed κ analyses between each individual observer; the results were similar to each other and to the 3-way κ statistics reported in Table 2 except for loss of polarity, in which case the individual κ values were 0.420, 0.168, and 0.036, respectively. Overall assessment of malignancy was the most reproducible criterion, with an excellent κ value (0.708). Chromatin clumping was also very reproducible (κ =

▌Table 4▌
Sensitivity and Specificity of Loss of Polarity for Three Different Observers for a Diagnosis of Malignancy

Observer	Sensitivity (%)	Specificity (%)
1	51.4	76.7
2	31.4	96.7
3	18.1	91.5
Consensus	36.2	96.7

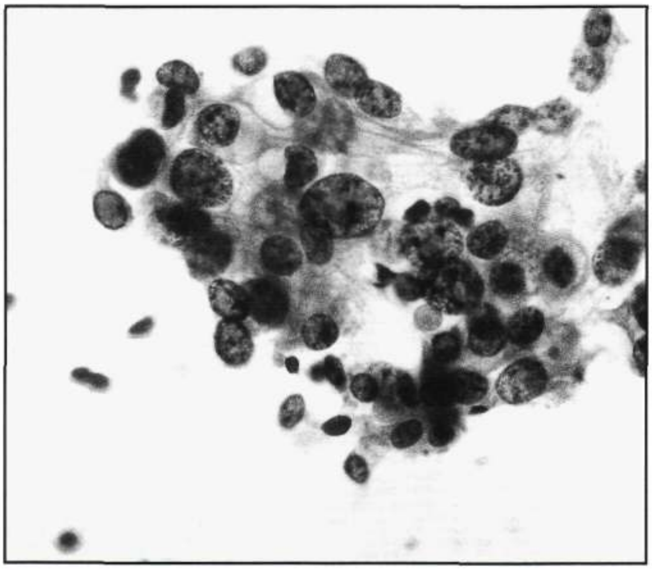
▌Table 5▌
Sensitivity and Specificity of Bloody Background for Three Different Observers for a Diagnosis of Malignancy

Observer	Sensitivity (%)	Specificity (%)
1	5.7	95.0
2	40.0	66.7
3	12.4	88.1
Consensus	15.2	88.3

▌Table 6▌
Sensitivity and Specificity of Four Different Diagnostic Criteria for a Diagnosis of Malignancy

Criteria	Sensitivity* (%)	Specificity* (%)
Iowa	24.8 (20.0–22.9)	98.3 (95.0–100)
Japan	39.0 (31.4–44.8)	86.7 (76.7–80.0)
Boston	21.9 (10.3–23.8)	100 (96.7–100)
Overall assessment of malignancy	36.2	95.0

*Result for consensus (range of individual results).



▌Image 1▌ Adenocarcinoma in bile duct brushings. The cells demonstrate chromatin clumping, an increased nuclear:cytoplasmic ratio, and loss of honeycombing but not nuclear molding (Papanicolaou, ×400).

0.629). On the other hand, loss of polarity, presence of a bloody background, and flat nuclei were not reproducible.

The sensitivity and specificity of each cytologic feature for the consensus diagnosis are summarized in **▌Table 3▌**. The sensitivity and specificity of each cytologic feature for each observer was similar to that of the consensus except for loss of polarity and bloody background, which are summarized in **▌Table 4▌** and **▌Table 5▌**, respectively.

The sensitivity of each set of diagnostic criteria using the consensus diagnoses are summarized in **▌Table 6▌**. The sensitivity of both the Japan criteria and the overall assessment of malignancy criteria were much higher than those of the Iowa and Boston criteria; however, the specificity of the Japan criteria were much lower than those of the other criteria. The results for the individual observers were similar.

After the aforesaid results had been obtained, those cases that were correctly identified as positive by the overall assessment of malignancy and missed by the Iowa criteria were reviewed. Of the 11 cases, none showed nuclear molding; however, all but 2 showed chromatin clumping and all but 1 had an increased nuclear:cytoplasmic ratio. In addition, all 11 cases showed both loss of honeycombing and enlarged nuclei. Next, we retrospectively tested the criteria of chromatin clumping, increased nuclear:cytoplasmic ratio, and either nuclear molding or loss of honeycombing **▌Image 1▌**. In that retrospective analysis, the sensitivity for a diagnosis of malignancy was 35.2% and the specificity was 95.0%. Substituting either enlarged nuclei or enlarged nuclei together with loss of honeycombing for loss of honeycombing did not improve the accuracy of the criteria.

Discussion

This report attempts to evaluate the diagnostic criteria for bile duct brushings proposed by others. We focused on 2 particular features of these criteria: reproducibility (as determined by κ values) and accuracy (as determined by sensitivity and specificity).

Surprisingly, our data clearly demonstrate that an overall assessment of malignancy based on the degree of atypia in a specimen was not only more reproducible than any other set of criteria but also resulted in a much higher sensitivity for malignancy with only a very small decrease in specificity. We suggest that this implies that cytologists are good at what they do; namely, cytologists develop a skill for assessing the overall degree of atypia or abnormality in a specimen that is a very good determinant of the presence of malignancy. The value of this skill in this setting has not previously been demonstrated.

It is no surprise that the sensitivities of both the Iowa and Japan criteria were lower than in the original articles. Although the practice of using the same cases from which a set of criteria was derived to test those same criteria is often reported in the cytology literature, verification of diagnostic utility requires testing criteria on a second set of subjects. This is especially true when the criteria in question were culled from a very large set of possible variables. The results will always be biased in favor of the new criteria.

However, the sensitivity of the criteria in our series is very low indeed, not even reaching 25%. Although there is a wide range of reported sensitivities for bile duct brushings,[1–18] this sensitivity is certainly in the lowest possible range. A review of the false-negative cases in this series revealed the majority to be due to sampling. One possible explanation for this fact is that the majority of patients in the series who had a malignant diagnosis had pancreatic masses, which have been previously shown to be more difficult to diagnose on the basis of bile duct brushings compared with biliary tree lesions.

The Boston and Iowa criteria (Table 5) have a similar level of accuracy. This is not surprising inasmuch as 2 of the 3 features for each criteria are the same. However, because the Boston criteria rely on loss of polarity, which does not appear to be very reproducible, we must conclude that the Iowa criteria are probably more reliable.

Nevertheless, with retrospective analysis it was possible to improve on the Iowa criteria and achieve results similar to those obtained with an overall assessment of malignancy by using a new set of modified qualitative criteria. These criteria suggest that malignancies in bile duct brushings have at least 2 different appearances: those that have nuclear molding and those that have loss of honeycombing. Obviously, the value of these new modified criteria needs to be assessed prospectively in a new and unbiased set of cases.

The current study has 2 limitations. First, an assessment of all of the criteria was made at the same time in each case. It is certainly possible that the exercise of evaluating for nuclear molding and so forth influenced our overall assessment of malignancy. Thus, our overall assessment of malignancy may have been influenced by the individual features contained in the other criteria, which may explain the ability to duplicate these results with a new set of modified criteria. The second limitation is that all 3 observers in this series are from the same institution and may share an approach to cytologic interpretation that is different from that at other institutions, even though the descriptions and illustrations of the cytologic features in the original articles were used to define each cytologic feature. This may be a particular problem with the assessment of overall malignancy inasmuch as, by design, this criterion was left somewhat vague. It would be of value to reproduce this study with a larger set of observers from a variety of practice settings.

In conclusion, we have compared the reproducibility and accuracy of 4 different sets of criteria for diagnosis of malignancy in bile duct brushings. We believe that assessment of the overall degree of malignancy is both more reproducible and more accurate than is reliance on any currently published set of qualitative criteria. However, the modified criteria of chromatin clumping, increased nuclear:cytoplasmic ratio, and either nuclear molding or loss of honeycombing may achieve a similar level of accuracy.

From the [1]Department of Pathology and the Division of Cytology, Brigham and Women's Hospital, the [2]Division of Biostatistics, Dana Farber Cancer Institute, Boston, Massachusetts, and [1,2]Harvard Medical School, Boston, Massachusetts.

Address reprint requests to Dr Renshaw: Department of Pathology, Brigham & Women's Hospital, 75 Francis St, Boston, MA 02115.

References

1. Floyd WN, Cobb C. Cholangiography and bile cytopathology in the diagnosis of biliary tract obstruction. *S Med J*. 1985;78:134–137.
2. Hatfield ARW, Smithies A, Wilkins R, et al. Assessment of endoscopic retrograde cholangiopancreatography (ERCP) and pure pancreatic juice cytology in patients with pancreatic disease. *Gut*. 1976;17:14–21.
3. Bourke JB, Swann JC, Brown CL, et al. Exocrine pancreatic function studies, duodenal cytology, and hypotonic duodenography in the diagnosis of surgical jaundice. *Lancet*. 1972;i:605–608.
4. Roberts-Thomson IC, Hobbs JB. Cytodiagnosis of pancreatic and biliary cancer by endoscopic duct aspiration. *Med J Aust*. 1979;1:370–372.
5. Nishimura A, Den N, Sato H, et al. Exfoliative cytology of the biliary tract with the use of saline irrigation under choledochoscopic control. *Ann Surg*. 1973;80:594–599.
6. Nieburgs HE, Dreiling DA, Rubio C, et al. The morphology of cells in duodenal-drainage smears: histologic origin and pathologic significance. *Am J Dig Dis*. 1962;7:489–505.
7. Kline TS, Joshi LP, Goldstein F. Preoperative diagnosis of pancreatic malignancy by the cytologic examination of duodenal secretions. *Am J Clin Pathol*. 1978;70:851–854.
8. Mitchell ML, Carney CN. Cytologic criteria for the diagnosis of pancreactic carcinoma. *Am J Clin Pathol*. 1985;83:171–176.
9. Farrari AP, Lichtenstein DR, Slivka A, et al. Brush cytology during ERCP for the diagnosis of biliary and pancreatic malignancies. *Gastrointest Endosc*. 1994;40:140–145.
10. Cobb CJ, Floyd WN. Usefulness of bile cytology in the diagnostic management of patients with biliary tract obstruction. *Acta Cytol*. 1985;29:93–100.
11. Kameya S, Kuno N, Kasugai T. The diagnosis of pancreatic cancer by pancreatic juice cytology. *Acta Cytol*. 1981;25:354–360.
12. Yamada T, Murohisa B, Muto K, et al. Cytologic detection of small pancreaticoduodenal and biliary cancers in the early developmental stage. *Acta Cytol*. 1984;28:435–442.
13. Ryan ME, Baldauf MC. Comparison of flow cytometry for DNA content and brush cytology for detection of malignancy in pancreaticobiliary strictures. *Gastrointest Endosc*. 1994;40:133–139.
14. Rupp M, Hawthorne CM, Ehya H. Brushing cytology in biliary obstruction. *Acta Cytol*. 1990;34:221–226.

15. Layfield LJ, Wax TD, Lee JG, et al. Accuracy and morphologic aspects of pancreatic and biliary duct brushings. *Acta Cytol*. 1995;39:11–18.
16. Kocjan G, Smith AN. Bile duct brushings cytology: potential pitfalls in diagnosis. *Diagn Cytopathol*. 1997;16:358–363.
17. Zafar N, Mensi DW, Moinuddin SM. Cytoradiologic diagnostic correlation of pancreaticobiliary disease: a retrospective review of findings in 94 cases [abstract]. *Acta Cytol*. 1997;41:1547.
18. Vadmal MS, Semmelmeier S, Smilari TF, et al. Brushing cytology of biliary tree: a retrospective analysis of 175 cases [abstract]. *Acta Cytol*. 1997;41:1547–1548.
19. Kurzawinski T, Deery A, Davidson BR. Diagnostic value of cytology for biliary stricture. *Br J Surg*. 1993;80:414–421.
20. Nakajima T, Tajima Y, Sugano I, et al. Multivariate statistical analysis of bile cytology. *Acta Cytol*. 1994;38:51–55.
21. Cohen MB, Wittchow RJ, Johlin FC, et al. Brush cytology of the extrahepatic biliary tract: comparison of cytologic features of adenocarcinoma and benign biliary strictures. *Mod Pathol*. 1995;8:498–502.
22. Smithies A, Hatfield ARW, Brown BE. The cytodiagnostic aspects of pure pancreatic juice obtained at the time of endoscopic retrograde cholangio-pancreatography (E.R.C.P.). *Acta Cytol*. 1977;21:191–195.