Interobserver Reproducibility of Gleason Grading of Prostatic Carcinoma: Urologic Pathologists

WILLIAM C. ALLSBROOK, JR, MD, KATHY A. MANGOLD, PHD, MARIBETH H. JOHNSON, MS, ROGER B. LANE, MD, CYNTHIA G. LANE, MD, MAHUL B. AMIN, MD, DAVID G. BOSTWICK, MD, PETER A. HUMPHREY, MD, EDWARD C. JONES, MD, VICTOR E. REUTER, MD, WAEL SAKR, MD, ISABELL A. SESTERHENN, MD, PATRICIA TRONCOSO, MD, THOMAS M. WHEELER, MD, AND JONATHAN I. EPSTEIN, MD

Gleason grading is now the most widely used grading system for prostatic carcinoma in the United States. However, there are only a few studies of the interobserver reproducibility of this system, and no extensive study of interobserver reproducibility among a large number of experienced urologic pathologists exists. Forty-six needle biopsies containing prostatic carcinoma were assigned Gleason scores by 10 urologic pathologists. The overall weighted kappa coefficient κw for Gleason score for each of the urologic pathologists compared with each of the remaining urologic pathologists ranged from 0.56 to 0.70, all but one being at least 0.60 (substantial agreement). The overall kappa coefficient κ for each pathologist compared with the others for Gleason score groups 2-4, 5-6, 7, and 8-10 ranged from 0.47 to 0.64 (moderate-substantial agreement), only one less than 0.50. At least 70% of the urologic pathologists agreed on the Gleason grade group (2-4, 5-6, 7, 8-10) in 38 ("consensus" cases) of the 46 cases. The 8 "nonconsensus" cases included low-grade tumors, tumors with small cribriform proliferations, and tumors whose histology was on the border between Gleason patterns. Interobserver reproducibility of Gleason grading among urologic pathologists is in an acceptable range. HUM PATHOL 32:74-80. Copyright © 2001 by W.B. Saunders Company

Key words: **prostatic neoplasms, prostatic carcinoma, prostatic adenocarcinoma, grading, Gleason grading, interobserver reproducibility.**

Abbreviations: **κw, weighted kappa; κ, kappa.**

Gleason grading is the most widely used grading system for prostatic carcinoma in the United States.[1] However, there have been relatively few large studies of interobserver reproducibility of Gleason grading (reviewed in ref 2). Further, there has never been an extensive study of interobserver reproducibility of Gleason grading among a large number of pathologists who specialize in urologic pathology. The current study assesses interobserver reproducibility of Gleason grading among 10 urologic pathologists.

MATERIALS AND METHODS

Two of the authors (W.C.A. and J.I.E.) collected, from several sources, 46 hematoxylin and eosin-stained glass slides of prostatic needle biopsies containing prostatic carcinoma. The amount of tumor on each slide ranged from microscopic foci to extensive. Examples of the spectrum of Gleason scores were included. No effort was made to make the cases particularly difficult.

The slides were distributed for Gleason grading to 9 additional urologic pathologists (M.B.A., D.G.B., P.A.H., E.C.J., V.E.R., W.S. I.A.S., P.T., and T.M.W.). Their scores were analyzed, along with the scores jointly assigned by W.C.A. and J.I.E., giving 10 sets of data from urologic pathologists.

In addition to the assigned Gleason scores, the scores were categorized into 4 groups (Gleason scores 2-4, 5-6, 7, and 8-10) for analysis. The interobserver agreement between the pathologists for all 46 cases was calculated using simple kappa (κ) and weighted kappa (κw) coefficients and an overall κ and κw coefficient for each urologic pathologist, each being used as the reference standard.[3,4] Simple kappa (κ) is a measure of interobserver agreement. When the observed agreement exceeds chance agreement, κ is positive, with its magnitude reflecting the strength of the agreement. The κw uses weights to quantify the relative difference between categories. Close disagreement (eg, ±1 Gleason score) is not weighted as heavily as more serious disagreements. The overall κ coefficient combines all of the κ, in which a pathologist is considered the comparison pathologist against all the remaining pathologists, into an overall estimate of the value of κ. All analyses were performed using SAS/STAT Software (Release 6.12, 1998; SAS Institute, Inc, Cary, NC). Kappa 0.00 to 0.20 reflects slight agreement, 0.21 to 0.40 fair agreement, 0.41 to 0.60 moderate agreement, 0.61 to 0.80 substantial agreement, and 0.81 to 1.00 almost perfect agreement.

Seven or more urologic pathologists agreed on Gleason

From The Medical College of Georgia, Augusta, GA; Emory University School of Medicine, Atlanta, GA; Mayo Clinic, Rochester, MN; Washington University Medical Center, St Louis, MO; Vancouver Hospital and Health Sciences Center, Vancouver, British Columbia, Canada; Memorial Sloan-Kettering Cancer Center, New York, NY; Harper Hospital, Detroit, MI; Armed Forces Institute of Pathology, Washington, DC; M.D. Anderson Cancer Center, Houston, TX; Baylor College of Medicine, Houston, TX; and The Johns Hopkins Medical Institutions, Baltimore, MD. Accepted for publication October 11, 2000.

Address correspondence and reprint requests to William C. Allsbrook, Jr, MD, Department of Pathology, Murphey Bldg, Room 210, School of Medicine, Medical College of Georgia, Augusta, GA 30912.

0046-8177/01/3201-0012$35.00/0
doi:10.1053/hupa.2001.21134

TABLE 1. Weighted Kappa, All Possible Pair Combinations of Urologic Pathologists, 46 Cases, Gleason Scores 2-10

Comparison Pathologist	Reference Pathologist									
	1	2	3	4	5	6	7	8	9	10
1		.63	.52	.70	.59	.58	.74	.66	.69	.84
2	.63		.59	.78	.74	.70	.67	.76	.68	.64
3	.52	.59		.55	.58	.61	.56	.58	.48	.56
4	.70	.78	.55		.70	.60	.70	.77	.72	.67
5	.59	.74	.58	.70		.65	.65	.65	.56	.62
6	.58	.70	.61	.60	.65		.66	.64	.65	.67
7	.74	.67	.56	.70	.65	.66		.63	.66	.75
8	.66	.76	.58	.77	.65	.64	.63		.70	.68
9	.69	.68	.48	.72	.56	.65	.66	.70		.68
10	.84	.64	.56	.67	.62	.67	.75	.68	.68	
Overall	.68*	.69	.56	.70	.64	.64	.68	.68	.66	.70*

*At least 1 κw in this group is significantly different from the others.

score group (2-4, 5-6, 7, and 8-10) in 38 of 46 cases. These were designated "consensus" cases. Kappa, κw, and overall κ and κw were not calculated for each urologic pathologist for the "consensus" cases because they would obviously be higher than those for the total 46 cases.

RESULTS

The overall κw for interobserver agreement for exact scores 2-10 for each of the urologic pathologists, used as the reference standard for each of the others (Table 1, pathologists not in alphabetical order), ranged from 0.56 to 0.70, with only one κw (0.56) less than 0.60 (substantial agreement). For pathologists 1 and 10, at least one of the κw combinations was significantly different from the others. κw combinations for each of the remaining pathologists were not significantly different.

For Gleason score groups of the 46 cases for all possible pair combinations of urologic pathologists, each used as the reference standard (Table 2, pathologists not in alphabetical order), κ ranged from 0.31 to 0.79. Kappa for 4 (4.4%) of 90 possible pair combinations was less than 0.40 (fair agreement), for 56 (62%) κ ranged from 0.41 to 0.60 (moderate agreement), and for 30 (33%) κ ranged from 0.61 to 0.80 (substantial agreement). Seventy-eight percent of the pair combinations had κ greater than 0.50 (midrange of moderate reproducibility). Overall κ for each of the 10 urologic pathologists compared with each of the other urologic pathologists ranged from 0.47 to 0.64, only 1 (0.47) less than 0.50 and 4 greater than 0.60.

One of the "consensus" cases was group 2-4, 12 were group 5-6, 9 group 7, the remaining 16 were group 8-10. Eleven (2.9%) of the responses for the 38 "consensus" cases were group 2-4. All possible pair combinations had, on average, perfect agreement on Gleason score in 24 of 46 (52%) cases and within ±1 Gleason score on an additional 16 cases (87% cumulative agreement for exact score ± 1), within ±2 scores on an additional 5 cases, and within ±3 scores on 1 case.

The greatest variability in primary pattern assignment among the urologic pathologists was seen for pattern 2 and pattern 5. Only 3 of 460 total responses were assigned primary pattern 1. Of the 46 total cases and the 38 "consensus" cases, 9.8% and 5.6% of responses, respectively, were primary pattern 2. Of the 32 "consensus" cases with "consensus" primary patterns, 3.8% of the urologic pathologist responses were pattern

TABLE 2. Kappa, All Possible Pair Combinations of Urologic Pathologists, 46 Cases, 4 Grade Groups (2-4, 5-6, 7, and 8-10)

Comparison Pathologist	Reference Pathologist									
	1	2	3	4	5	6	7	8	9	10
1		.576	.435	.666	.541	.481	.694	.693	.643	.788
2	.576		.529	.709	.739	.587	.454	.679	.559	.487
3	.435	.529		.526	.574	.518	.379	.515	.308	.455
4	.666	.709	.526		.710	.495	.542	.777	.657	.576
5	.541	.739	.574	.710		.521	.443	.678	.471	.570
6	.481	.587	.518	.495	.521		.505	.524	.540	.572
7	.694	.454	.379	.542	.443	.505		.501	.525	.721
8	.693	.679	.515	.777	.678	.524	.501		.567	.601
9	.643	.559	.308	.657	.471	.540	.525	.567		.615
10	.788	.487	.455	.576	.570	.572	.721	.601	.615	
Overall	.628	.595	.470	.641	.591	.528	.539	.628	.546	.612

TABLE 3. Gleason Score Distribution for 8 Nonconsensus Cases, Urologic Pathologists

Gleason Score by Groups	Cases							
	A	B	C	D	E	F	G	H
2-4	**6** 6 (4)	**3** 1 (3) 2 (4)	**4** 1 (2) 3 (4)					
5-6	**4** 3 (5) 1 (6)	**6** 2 (5) 4 (6)	**6** 6 (5)	**6** 6 (6)	**4** 4 (6)			
7		**1** 1 (7)		4 4 (7)	5 5 (7)	5 5 (7)	5 5 (7)	6 6 (7)
8-10					**1** 1 (9)	**5** 2 (8) 3 (9)	**5** 3 (8) 2 (9)	**4** 2 (8) 2 (9)

NOTE. Agreement by less than 70% urologic pathologists for Groups 2-4, 5-6, 7, 8-10.
Bold represents number of urologic pathologists assigning this group. Number preceding parentheses is number of urologic pathologists assigning score. Number in parentheses is actual score.

2; none were pattern 1. For the "consensus" cases, 2.9% of urologic pathologist responses were group 2-4.

The Gleason score distribution for the 8 "nonconsensus" cases is shown in Table 3. There were 3 cases in which the disagreement was between groups 2-4 and 5-6 (Figs 1-3), 2 cases between 5-6 and 7 (Figs 4 and 5), and 3 cases between 7 and 8-10 (Figs 6-8).

Overall, for the 38 "consensus" cases, 20 (5.3%) of the 380 total scores were overscores, placing the case into the next highest group, and only 1 was by more than 1 group. Eight of the overscores were by 1 pathologist. Thirty-three (8.7%) of 380 scores were underscores, placing the case into the next lowest group, but none were by more than 1 group. These 38 "consensus" cases are reviewed in a subsequent study in which they were graded by general pathologists.[2]

Representative Cases

The 8 "nonconsensus" cases are reviewed in Table 3 and Figures 1 through 8.

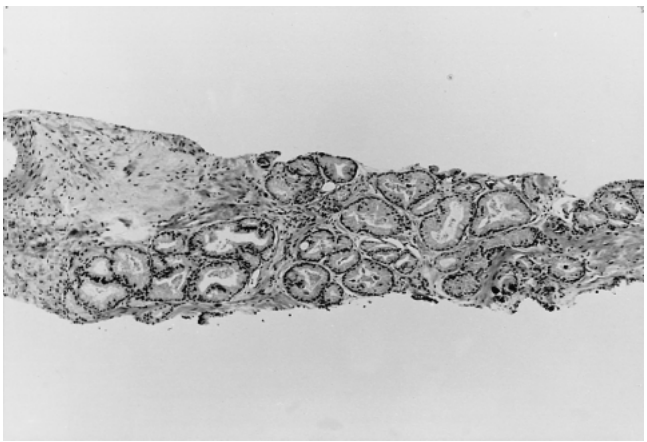

FIGURE 1. (Case A). Acini with variable size and shape and abundant pale staining cytoplasm. Crystalloids and eosinophilic secretions were present. Six pathologists assigned a score of 4, 3 assigned a score of 5, and 1 assigned a score of 6.

Case A. Most of the pathologists placed this at the lower end of the grading spectrum (Fig 1). The tumor was given a score of 4 by six pathologists, a score of 5 by 3, and a score of 6 by 1. Seven assigned a primary pattern of 2, and 3 assigned a primary pattern of 3.

Case B. Six pathologists assigned a primary pattern of 3 and 4 assigned a pattern of 2 for this tumor (Fig 2A and B). It was placed in group 2-4 by 3 pathologists; 1 assigned a score of 3 and 2 assigned a score of 4. Six pathologists placed the lesion in the 5-6 group; 2 assigned a score of 5 and 4 assigned a score of 6. One assigned a score of 7.

Case C. All urologic pathologists recognized that the tumor was lower grade (Fig 3). Eight pathologists assigned a primary pattern of 2 and 1 assigned a primary pattern of 1, only 1 assigning a primary pattern of 3. However, there was not a consensus between 2-4 and 5-6. This carcinoma was placed in group 2-4 by 4 urologic pathologists; 1 assigned score 2 and 3 assigned score 4. Six placed the lesion in group 5-6, all assigning a score of 5.

Case D. Nine of 10 pathologists assigned a primary pattern of 3, one assigning 4. Seven of 10 pathologists assigned a secondary pattern of 3 and 3 assigned a pattern of 4 (Fig 4). The tumor was assigned a score of 6 by 6 pathologists and a score of 7 by the remaining 4.

Case E. Seven pathologists assigned a primary pattern of 3 and 3 assigned 4 (Fig 5). Six assigned a secondary pattern of 3, 3 assigned a pattern of 4, and 1 assigned a pattern of 5. It was assigned a score of 6 by 4 pathologists. Five pathologists assigned a score of 7 and one assigned a score of 9.

Case F. All 10 pathologists assigned a primary pattern of 4. (Fig 6A and B). Five pathologists assigned a score of 7, 2 assigned a score of 8, and 3 assigned a score of 9.

Case G. Eight pathologists assigned a primary pattern of 4, 1 assigned 3, and 1 assigned 5 (Fig 7A and B). Five pathologists assigned a score of 7 and 5 placed the tumor in the 8-10 group, 3 assigned 8, and 2 assigned 9.

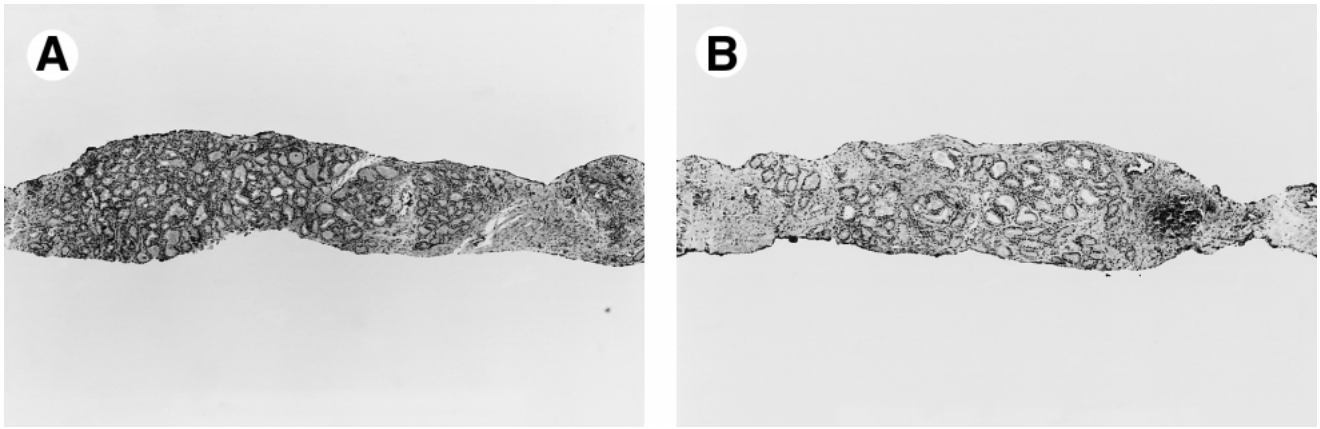

FIGURE 2. (A and B; Case B). Acini smaller and more crowded, and cytoplasm more amphophilic than Case A (A). Occasionally, acini more loosely arranged (B). Focus of more complex acini not shown. One pathologist assigned a score of 3, 2 assigned a score of 4, 2 assigned a score of 5, 4 assigned a score of 6, and 1 assigned a score of 7.

Case H. Eight pathologists assigned a primary pattern of 4, 2 assigning 3 (Fig 8A-C). Six pathologists assigned a secondary pattern of 3 and 4 assigned a secondary pattern of 5. Six assigned a score of 7, 2 assigned a score of 8, and 2 assigned a score of 9.

DISCUSSION

There is some variability in interobserver agreement among urologic pathologists, but the overall κw for scores 2-10 are, with one exception (0.56), greater than 0.60 (substantial agreement). This is better agreement than most other studies in which κw was calculated. Some studies had similar or better exact or ± 1 agreement, but these had only a pair of participating pathologists or a prestudy primer or agreement on criteria (reviewed in ref 2). Further, the overall interobserver agreement is greater than the mid-portion of the moderate range when scores are separated into 4 groups. This, too, is substantially better agreement than in previous reports (reviewed in ref 2).

The greatest variability in pattern responses occurred for patterns 2 and 5. The latter is less of a problem because identifying pattern 5 as pattern 4 usually does not result in the tumor being removed from the score group 8-10.

The "nonconsensus" cases are instructive. There is a particular problem with grading prostatic carcinoma at the lower end of the Gleason spectrum. Two of the "nonconsensus" cases (A and C) are similar and both were recognized to be at the lower end of Gleason grade, 7 and 8 pathologists, respectively, assigning a primary pattern of 2. Six placed one of the tumors (case A) in the 2-4 group and 6 placed the other (case C) in the 5-6 group, all assigning a score

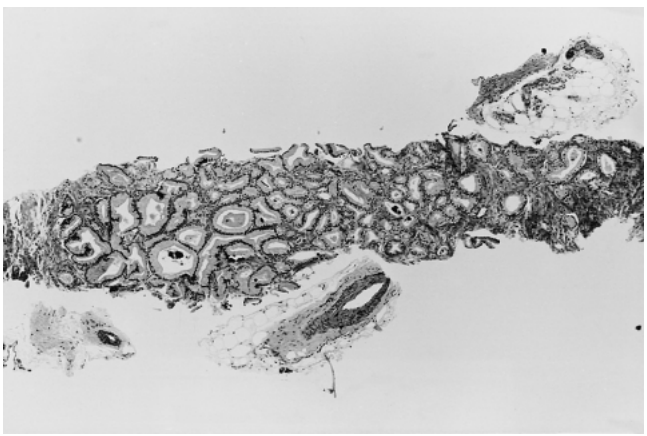

FIGURE 3. (Case C). Variability in acinar size and shape, some crowded, others more loosely arranged. The cytoplasm is abundant and pale. Intraluminal crystalloids are present. One pathologist assigned a score of 2, 3 assigned a score of 4, and 6 assigned a score of 5.

FIGURE 4. (Case D). In addition to the cribriform growth pattern, there are also scattered smaller acinar structures. Six pathologists assigned a score of 6 and 4 assigned a score of 7.

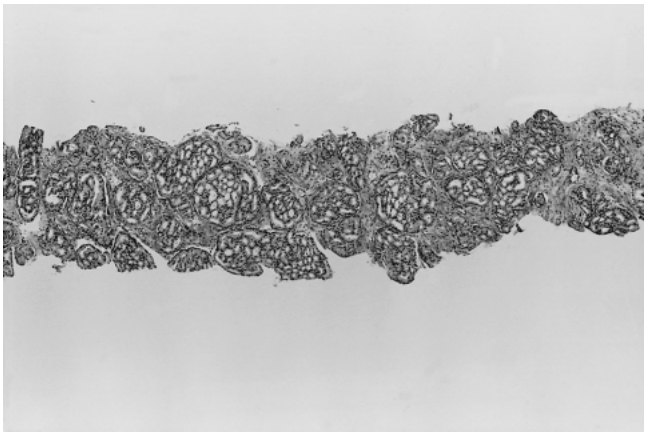

FIGURE 5. (Case E). This tumor also has a predominant cribriform growth pattern. Four pathologists assigned a score of 6, 5 assigned a score of 7, and 1 assigned a score of 9.

of 5. Case B was an even greater problem, with scores ranging from 3 to 7. It is obvious that the criteria for low-grade tumors must be more clearly defined, as even experienced urologic pathologists do not show good interobserver agreement. Another critical issue related to this group of tumors is the demonstration of excessive placement of tumors in Gleason group 2-4 (undergrading) by general pathologists.[2] One of the senior authors (J.I.E.) has recently recommended that Gleason's 2-4 prostatic carcinoma should not be diagnosed by needle biopsy.[5]

Although urologic pathologists recognize large cribriform sheets of tumor as pattern 4,[2] tumors composed of smaller circumscribed cribriform proliferations were another problem area (cases D and E). The issue in these cases is the lack of consensus in the diagnosis of cribriform Gleason pattern 3, possibly representing variability in interpretation of intraductal growth versus prostatic intraepithelial neoplasia, of invasive versus intraductal growth, and/or assessing degree of irregularity of some of the cribriform structures.[6-8]

The remaining 3 "nonconsensus" cases (F-H), along with case B, present 2 additional problems in grading. First, these cases had a large amount of tumor in the needle biopsies, and there were multiple patterns. There may be some subjective difficulty in quantitatively assessing the most dominant and next-most dominant pattern. Parenthetically, in these instances, assigning a primary and secondary pattern grade allows for some modulation of discrepancies between 2 assigned patterns, as opposed to assigning a single overall grade to the tumor. Second, these tumors had areas lying at the border between patterns (between patterns 2 and 3 in cases A-C, and between patterns 3 and 4, as well as 4 and 5, in cases F-H). Pathologists are faced with such "borderline" cases ("Is this a 'bad' 3 or a 'good' 4?") in all grading systems on a daily basis. With regard to patterns 3 and 4, the major problem appeared to be defining the limits of tiny, poorly defined acinar structures. In addition, at times, it may be difficult to determine whether the loss of acinar spaces is caused by compression artifact or by real inability to form spaces. With regards to patterns 4 and 5, the limits and proportions of tiny, poorly defined acinar structures versus cords and nests of cells appeared to be a problem. Finally, there may have been difficulty in defining the significance of cytoplasmic vacuoles as compared to true, albeit tiny, lumena. In cases F through H, there was essentially an even split between score group 7 and group 8-10. This is important because placing the tumor in the former would more likely lead to treatment by radical prostatectomy, and the latter group would more likely lead to treatment by radiation therapy.

In summary, for this series of cases, urologic pathologists have variable, but overall acceptable, interobserver reproducibility of Gleason grading, including actual Gleason score as well as Gleason grade groups. The interobserver reproducibility for any se-

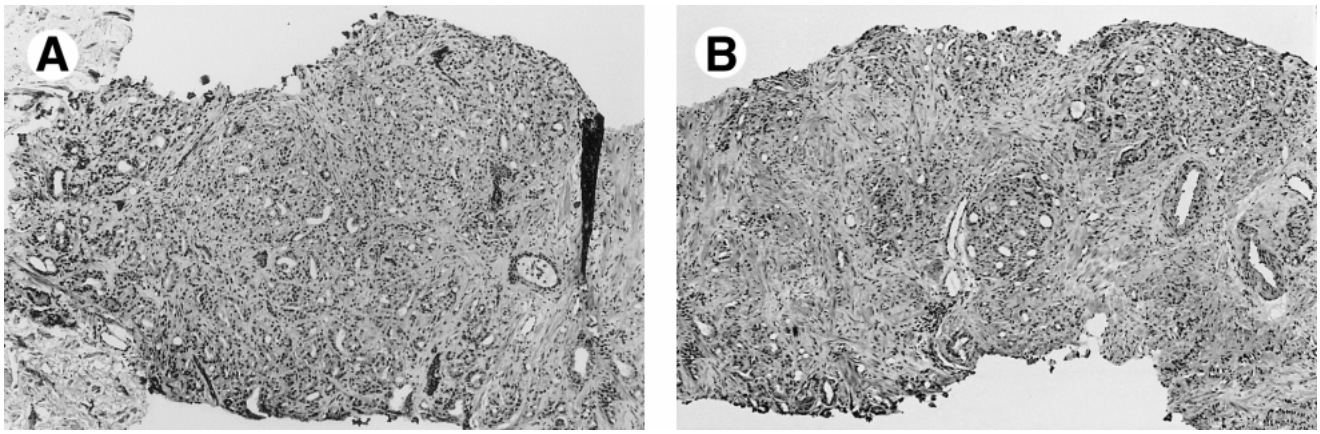

FIGURE 6. (A and B; Case F). Small acini with areas of poorly defined to absent lumena (A). Areas of classic Pattern 3 not shown. In other areas (B), nests or sheets of cells, at times with vacuolization but only rare true lumen formation. Five pathologists assigned a score of 7, 2 assigned a score of 8, and 3 assigned a score of 9.

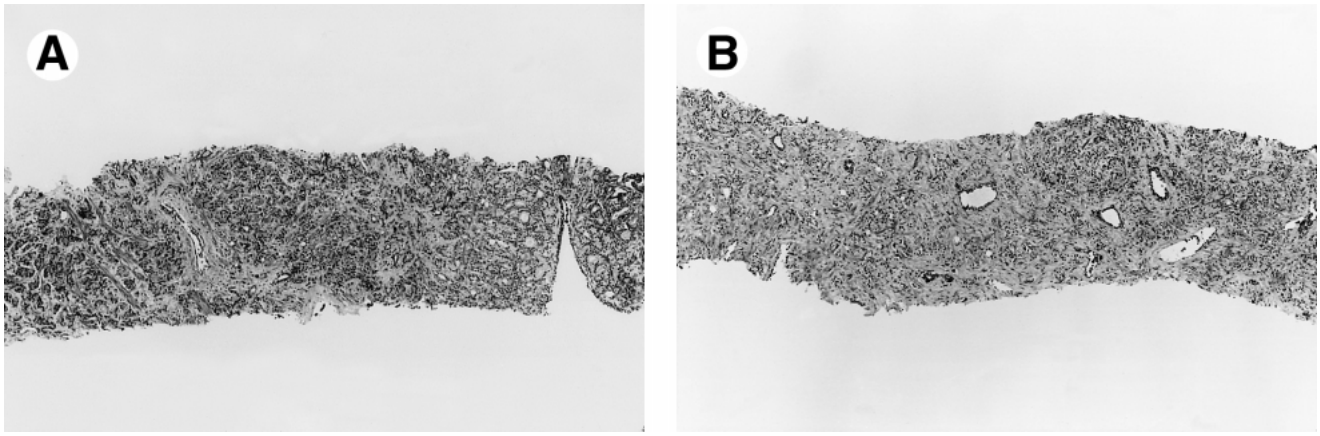

FIGURE 7. (A and B; Case G). Mostly infiltrating nests of cells with only occasional lumena (A, left). In other areas, well-defined acini with rare cribriforming (A, right). Another area (B) with smaller nests and cords with only occasional, and at times no, lumena. Five pathologists assigned a score of 7, 3 assigned a score of 8, and 2 assigned a score of 9.

ries of cases will depend on the number of problem cases studied. Problem cases include low-grade tumors, tumors with small circumscribed cribriform structures, and tumors borderline between classic Gleason patterns. The problem of low-grade tumors can be eliminated by following the recommendation discussed previously not to diagnose them by needle biopsy, at least until more precise criteria are defined that correlate with prognosis.[5] Cribriform tumors are the subject of a number of recent studies, and hopefully a clearer understanding of them will lead to more precise grading.[6-8] Finally, there are the tumors with "borderline" patterns, at times, with large amounts of tumor. It is not possible to determine the actual percentage of these tumors in day-to-day practice, but in our experience they are a minority. We

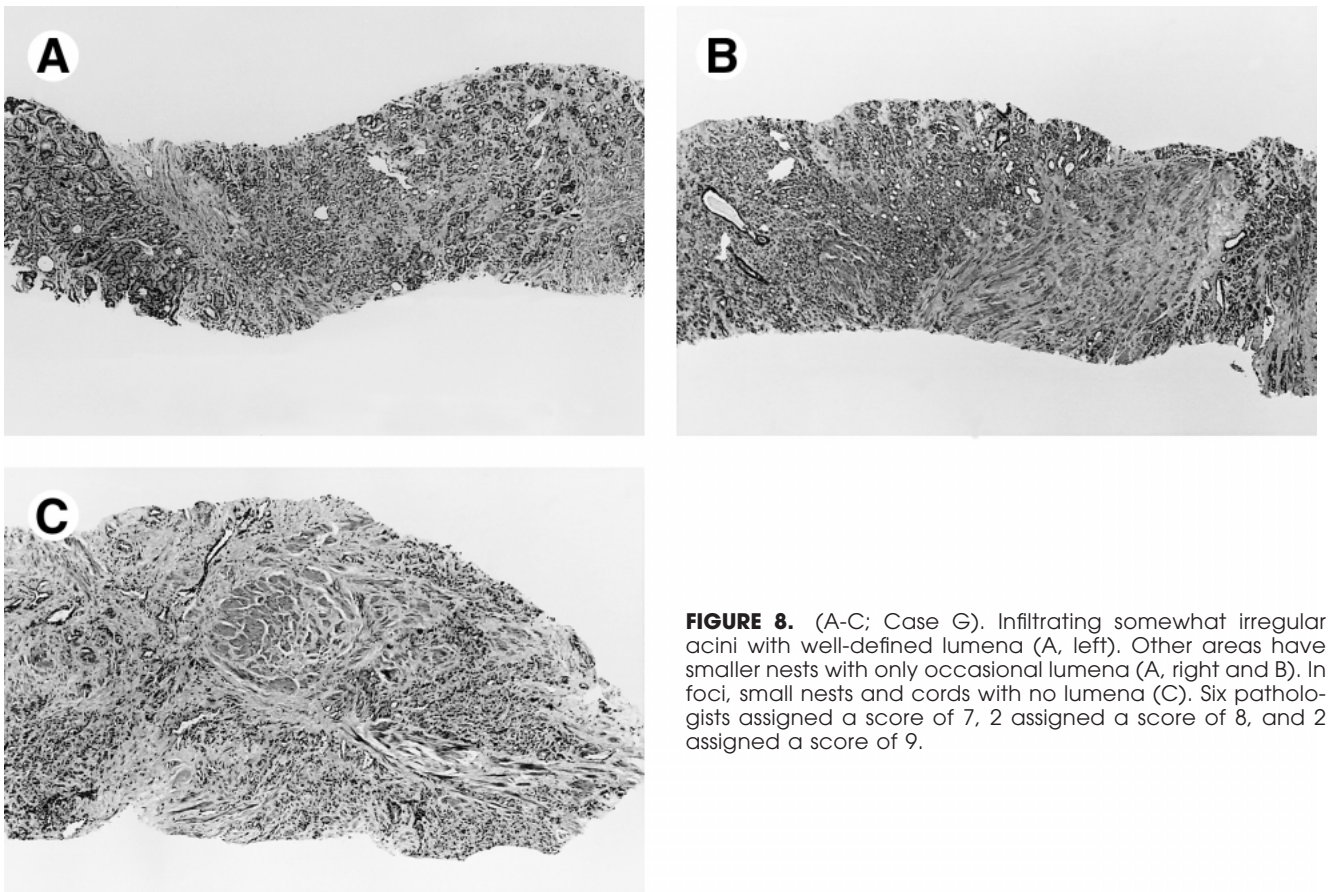

FIGURE 8. (A-C; Case G). Infiltrating somewhat irregular acini with well-defined lumena (A, left). Other areas have smaller nests with only occasional lumena (A, right and B). In foci, small nests and cords with no lumena (C). Six pathologists assigned a score of 7, 2 assigned a score of 8, and 2 assigned a score of 9.

anticipate that these cases would give rise to similar problems in any grading system.

Acknowledgment. The authors thank Michelle Page and JoAnn Higdon for secretarial assistance, Laura McKie for preparing the Tables, and Cheryl Nichols for photographic assistance.

REFERENCES

1. Allsbrook WC, Jr, Mangold KA, Yang X, et al: The Gleason grading system: An overview. J Urol Pathol 10:141-157, 1999

2. Allsbrook WC, Jr, Mangold KA, Johnson MH, et al: Interobserver reproducibility of Gleason grading of prostatic carcinoma: General pathologists. HUM PATHOL 32:81-88, 2001

3. Cohen J: A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20:37-46, 1960

4. Fleiss JL: Statistical Methods for Rates and Proportions (ed 2). New York, NY, Wiley, 1981

5. Epstein JI: Gleason score 2-4 adenocarcinoma of the prostate on needle biopsy: A diagnosis that should not be made. Am J Surg Pathol 24:477-478, 2000

6. Amin MB, Schultz DS, Zarbo RJ: Analysis of cribriform morphology in prostatic neoplasia using antibody to high-molecular-weight cytokeratins. Arch Pathol Lab Med 118:260-264, 1994

7. McNeal JE, Yemoto CEM: Spread of adenocarcinoma within prostatic ducts and acini. Morphologic and clinical correlations. Am J Surg Pathol 20:802-814, 1996

8. Rubin MA, de La Taille A, Bagiella E, et al: Cribriform carcinoma of the prostate and cribriform prostatic intraepithelial neoplasia. Incidence and clinical implications. Am J Surg Pathol 22:840-848, 1998