0022-5347/02/1685-2006/0
THE JOURNAL OF UROLOGY®

Vol. 168, 2006–2010, November 2002
Printed in U.S.A.
DOI: 10.1097/01.ju.0000034255.95952.df

INTEROBSERVER REPRODUCIBILITY OF PERCENT GLEASON GRADE 4/5 IN TOTAL PROSTATECTOMY SPECIMENS

AXEL GLAESSGEN, HANS HAMBERG, CARL-GUSTAF PIHL, BIRGITTA SUNDELIN, BO NILSSON AND LARS EGEVAD

From the Departments of Pathology and Cytology, and Oncology and Pathology, Karolinska Hospital, Stockholm, Department of Pathology, Central Hospital, Västerås and Department of Pathology, Sahlgrenska University Hospital, Gothenburg, Sweden

ABSTRACT

Purpose: Recently the percent Gleason grade 4/5 was proposed as a predictor of the outcome of prostate cancer and it has been shown that it adds prognostic information to that given by Gleason score. To our knowledge the interobserver variability of percent Gleason grade 4/5 has not yet been investigated. We assessed the percent Gleason grade 4/5, including the identification of high grade patterns and estimation of the percent tumor area, which is potentially more difficult than conventional Gleason grading.

Materials and Methods: A consecutive series of 69 total prostatectomy specimens was reviewed. A single slide per specimen was circulated among 4 genitourinary pathologists, who assessed Gleason score and the percent Gleason grade 4/5. Results were compared pairwise and a weighted κ was calculated for Gleason score and the percent Gleason grade 4/5.

Results: The 4 observers had a mean weighted κ for Gleason score and the percent Gleason grade 4/5 of 0.52 to 0.66 (overall mean 0.56) and 0.58 to 0.72 (overall mean 0.66), respectively. The best agreement for percent Gleason grade 4/5 was in 2 pathologists at the same department (weighted κ 0.86). Transition zone tumors had a lower weighted κ for Gleason score but a higher weighted κ for percent Gleason grade 4/5 than peripheral zone tumors. In cases of the greatest disagreement in the percent Gleason grade 4/5 crush artifact, cribriform cancer and high grade PIN within the tumor were significantly more common. An intraobserver reproducibility of weighted κ 0.91 was achieved for Gleason score and the percent Gleason grade 4/5.

Conclusions: Interobserver reproducibility of the percent Gleason grade 4/5 is substantial and at least as good as that of the Gleason score. Hence, concern about interobserver variability should not deter pathologists from using the percent Gleason grade 4/5 as a prognostic marker for prostate cancer.

KEY WORDS: prostate; prostatic neoplasms; prostatectomy; reproducibility of results; pathology, clinical

The Gleason score is a well established prognostic parameter for prostate cancer.[1,2] However, a weakness of the Gleason system is the strong tendency toward tumor clustering in the mid range of the scores. In some total prostatectomy series as many as 86%[3] to 89%[4] of tumors were Gleason score 6 or 7. To improve prognostication in this group Stamey et al proposed a new prognostic parameter, namely the percent Gleason grade 4/5, which is proportion of tumor occupied by high grade cancer.[5,6] The percent Gleason grade 4/5 predicts lymph node metastasis,[7] recurrence after total prostatectomy[5] and disease specific survival in patients on watchful waiting.[8] It has been shown that percent Gleason grade 4/5 conveys prognostic information that is independent of the Gleason score.[8] Others have reported the reproducibility of Gleason grading[9–13] but to our knowledge interobserver variability of the percent Gleason grade 4/5 has not yet been investigated. We assessed interobserver reproducibility of the percent Gleason grade 4/5 in total prostatectomy specimens.

MATERIAL AND METHODS

The series included a consecutive series of 69 total prostatectomy specimens received from January to December 2000 at the department of pathology and cytology at Karolinska Hospital, Stockholm. The study was approved by the independent ethics committee of Karolinska Hospital (01-443). Mean patient age at diagnosis was 61.4 years (range 50 to 73). Clinical stage was T1b, T1c, T2 and T3 in 3, 48, 17 and 1 cases, respectively. Mean preoperative serum prostate specific antigen was 10.2 ng./ml. (range 2.5 to 44). None of the patients received hormonal treatment or radiotherapy before prostatectomy.

The prostate was fixed overnight in 10% buffered formalin. The specimen was inked and sliced horizontally at 4 mm. intervals. Slices were cut into 2 to 6 segments, usually quadrants, and the whole prostate was subsequently blocked in standard cassettes. The specimens were dehydrated, embedded in paraffin, sectioned at 4 μ., and stained with hematoxylin and eosin. The tumor and the prostate capsule were outlined on the slides. The main neoplasm originated from the peripheral zone in 54 men and from the transition zone in 15. We noted seminal vesicle invasion in 13 cases, extraprostatic extension in 41 and positive margins in 33.

A single slide per prostatectomy specimen, including the main tumor, was circulated among 4 pathologists specializing in genitourinary pathology (L. E., H. H., C. G. P. and B. S.) for a total of 276 responses and 414 pairwise comparisons. No consensus training on estimating the percent Gleason grade 4/5 preceded the study. Slides were reviewed, and the Gleason score and percent Gleason grade 4/5 were assessed. The percent Gleason grade 4/5 was estimated as 0%, focal (5% or less) and subsequently at 10% intervals (11% to 20%, 21% to 30% and so forth). Results were compared pairwise and a weighted κ was calculated for Gleason score, percent Gleason grade 4/5, and the distribution of tumors in

Accepted for publication June 28, 2002.

the categories Gleason scores 4 to 6, 7 and 8 to 10. For comparing the dichotomous distribution of lesions with or without Gleason grades 4 to 5 cancer the unweighted κ was used, including 0 to 0.2—slight, 0.21 to 0.4—fair, 0.41 to 0.6—moderate, 0.61 to 0.8—substantial and 0.81 to 1—almost perfect agreement.[12] Crush artifact, cribriform cancer, high grade PIN within the tumor and mucinous cancer were graded 0—absent, 1—focal, 2—more than focal but not extensive and 3—extensive. The Mann-Whitney U test was used for comparing these estimates and the chi-square test was used for comparing proportions with $p < 0.05$ considered significant.

RESULTS

The Gleason score determined by the pathologists was 5 to 9 (mean 6.35 to 6.88, table 1). The average Gleason score in peripheral and transition zone tumors was 6.83 and 6.13, respectively. The observers assigned a Gleason score of 6 or 7 in 81% to 90% of cases (mean 86%). The average percent Gleason grade 4/5 of the observers was 14% to 41% (mean 30%, figs. 1 and 2). The average percent Gleason grade 4/5 in peripheral and transition zone tumors was 33% and 17%, respectively. The percent Gleason grade 4/5 was between 0% and 50% in 76% of the 276 responses (table 2).

The observers were compared pairwise and the Gleason score concordance within each of the 6 pairs was calculated. Exact agreement was observed in 44% to 70% of cases (mean 57%), while there was disagreement of 1 Gleason score in 26% to 51% (mean 39%) and disagreement of 2 in 1% to 6% (mean 4%). The mean weighted κ for the Gleason score estimated by each observer versus each of the other 3 observers was 0.52 to 0.66 (overall mean 0.56, table 3). Three pathologists showed moderate agreement and 1 showed substantial agreement. When the results were categorized as Gleason scores 4 to 6, 7 and 8 to 10, a mean weighted κ of 0.53 (range 0.45 to 0.61) was obtained.

All responses were compared pairwise and differences in the percent Gleason grade 4/5 were calculated (fig. 3). In 113 of the 414 pairs (27%) there was exact agreement, in 164 (40%) the disagreement was 5% or less, in 252 (61%) it was 15% or less and in 35 (8%) it was more than 50%. The mean weighted κ for the percent Gleason grade 4/5 estimated by each observer against each of the other 3 observers was 0.58 to 0.72 (overall mean 0.66, table 3). Three pathologists showed substantial agreement and 1 showed moderate agreement. The best agreement was achieved by 2 observers at the same department (weighted κ 0.86). Results were categorized according to any Gleason grades 4 to 5 cancer and unweighted kappa was calculated. The mean kappa of all observers was 0.49 (range 0.34 to 0.56). A single observer (L. E.) reviewed the slides a second time after 10 months. Intra-observer reproducibility was almost perfect for Gleason score and the percent Gleason grade 4/5 (weighted κ 0.91).

Gleason score concordance was lower in transition zone than in peripheral zone tumors (weighted κ 0.43 and 0.54), while the opposite was true for the percent Gleason grade 4/5 (weighted κ 0.74 and 0.64, respectively, table 3). Of the 54 peripheral zone tumors 4 (7%) had more than 1 Gleason score difference in the lowest and highest Gleason score of the 4 observers, as did 4 of the 15 transition zone tumors (27%) ($p = 0.039$). Of the peripheral zone tumors 16 (30%) had more than a 50% difference in the lowest and highest percent Gleason grade 4/5 of the 4 observers compared with 1 transition zone tumor (7%) ($p = 0.068$). In the 17 cases with more than a 50% disagreement in the percent Gleason grade 4/5 crush artifact, cribriform cancer and high grade PIN within the tumor were significantly more common than in the other cases, while mucinous cancer was only nonsignificantly more common (table 4). In peripheral zone tumors cribriform cancer and high grade PIN were significantly more common than in transition zone tumors, while crush artifact and mucinous cancer were only nonsignificantly more common (table 4).

TABLE 1. *The distribution of Gleason scores among the 4 observers*

	Gleason Score					
	5	6	7	8	9	Av.
Observer No.:						
1	6	15	41	3	4	6.77
2	5	15	45	3	1	6.71
3	4	42	20	1	2	6.35
4	1	16	42	10	0	6.88
No. tumors (%)	16 (5.8)	88 (31.9)	148 (53.6)	17 (6.2)	7 (2.5)	

There were no Gleason scores 2 to 4 or 10 tumors.

DISCUSSION

To be clinically useful a histopathological grading system must provide significant prognostic information, be reasonably easy to use and reproducible. The value of the percent Gleason grade 4/5 as a predictor of prognosis has already been demonstrated[5,7,8] but to our knowledge we report the first investigation of the reproducibility of this measure.

Histopathological grading of cancer is a subjective translation of a continuous spectrum of morphological patterns into discrete numerical data. This translation is rendered even more difficult because tumor dedifferentiation is a nonlinear process that results in a complex array of morphological variants. Grading prostate cancer is particularly difficult because of the pronounced morphological heterogeneity of this tumor. More than 50% of prostatectomy specimens contain cancer of at least 3 Gleason grades.[14] Inevitably any grading system is flawed by some degree of interobserver variability. Interobserver variability of the Gleason system has been investigated by several groups. A detailed review was provided by Allsbrook et al.[11] For many reasons the results of previous studies are difficult to compare. The number of observers, and the number and type of specimens differed. Some studies were preceded by a tutorial, while in others specimens were selected and not consecutive. Svanholm and Mygind reported a Gleason score weighted κ of 0.7 but only 2 observers were involved, which may have overestimated agreement compared with a larger group of observers.[9] Lessells et al used 12 observers for a weighted κ of 0.45.[10] In a study of Allsbrook et al an overall mean weighted κ of 0.66 (range 0.56 to 0.7) was achieved by 10 observers and a set of 46 select biopsy specimens.[12] The overall mean weighted κ for Gleason score in our study was 0.56 (range 0.52 to 0.66) with 4 observers. Specimens were consecutive to avoid biased results because of under or over representation of difficult cases. In a previous study we noted that the Gleason score of prostate cancer could be grouped into 4 categories with significantly different prognoses, namely 4 to 5, 6, 7 and 8 to 10.[15] However, in this study no Gleason score 4 or 10 tumors were reported and only a few were assigned a Gleason score of 9. Hence, grouping tumors in the categories 4 to 5, 6, 7 and 8 to 10 would only have differed marginally from ungrouped Gleason scores. Therefore, for studying the effect on interobserver variability tumors were grouped into the 3 categories 4 to 6, 7 and 8 to 10. The reproducibility of grouped Gleason scores was almost the same as that of ungrouped scores (weighted κ 0.53 and 0.56, respectively).

Assessment of the percent Gleason grade 4/5 includes identifying high grade tumor and estimating the percent of tumor area involved with such patterns, which is potentially more difficult than conventional Gleason grading. However, the current study shows that reproducibility of the percent Gleason grade 4/5 is at least as good as that of the Gleason score. There are several possible explanations for this finding. The assessment of Gleason score includes some difficulties that are