Clinical Urology
International Braz J Urol Vol. 33 (5): 639-651, September - October, 2007

Interobserver Agreement of Gleason Score and Modified Gleason Score in Needle Biopsy and in Surgical Specimen of Prostate Cancer

Sergio G. Veloso, Mario F. Lima, Paulo G. Salles, Cynthia K. Berenstein, Joao D. Scalon, Eduardo A. Bambirra

Section of Urology, Mario Penna Hospital, and Department of Pathology, School of Medicine, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

ABSTRACT

Introduction: Gleason score, which has a high interobserver variability, is used to classify prostate cancer. The most recent consensus valued the tertiary Gleason pattern and recommended its use in the final score of needle biopsies (modified Gleason score). This pattern is considered to be of high prognostic value in surgical specimens. This study emphasized the evaluation of the modified score agreement in needle biopsies and in surgical specimen, as well as the interobserver variability of this score.
Materials and Methods: Three pathologists evaluated the slides of needle biopsies and surgical specimens of 110 patients, reporting primary, secondary and tertiary Gleason patterns and after that, traditional and modified Gleason scores were calculated. Kappa test (K) assessed the interobserver agreement and the agreement between the traditional and modified scores of the biopsy and of the surgical specimen.
Results: Interobserver agreement in the biopsy was K = 0.36 and K = 0.35, and in the surgical specimen it was K = 0.46 and K = 0.36, for the traditional and modified scores, respectively. The tertiary Gleason grade was found in 8%, 0% and 2% of the biopsies and in 8%, 0% and 13% of the surgical specimens, according to observers 1, 2 and 3, respectively. When evaluating the agreement of the traditional and modified Gleason scores in needle biopsy with both scores of the surgical specimen, a similar agreement was found through Kappa.
Conclusion: Contrary to what was expected, the modified Gleason score was not superior in the agreement between the biopsy score and the specimen, or in interobserver reproducibility, in this study.

***Key words**: prostatic neoplasms; biopsy, needle; surgery; pathology*
Int Braz J Urol. 2007; 33: 639-51

INTRODUCTION

Prostate cancer tends to be morphologically heterogeneous (1), showing several patterns of differentiation, classified by Gleason system (2). Prostate needle biopsy provides random samples, which might not represent neoplasia in all its heterogeneity, generally downgrading the tumor (3-5). By clinical accompaniment, a worse prognosis was found in the

patients who had small proportions of Gleason patterns 4 and 5 tumors, which are not mentioned in the Gleason score (6-10). From this observation, the concept of modified Gleason score was created incorporating these small most aggressive patterns in the patient's score and being used in some prognostic nomograms (11,12), Figure-1.

Several studies deal with interobserver agreement of Gleason score, with all sorts of different results (13). Other studies deal with the agreement as regards modified Gleason score in slides (14). Recently, Helpap reported better association between needle biopsy and surgical specimen using the modified Gleason score (11). Gleason histopathological classification shows high level of subjectivity. Despite its undeniable clinical importance, as a diagnostic method, the Gleason score, more precisely the modified score, needs to be evaluated in relation to its reliability. Taking this into account, we tried to evaluate the interobserver agreement and the association between needle biopsy and the surgical specimen adopting Gleason and modified Gleason scores.

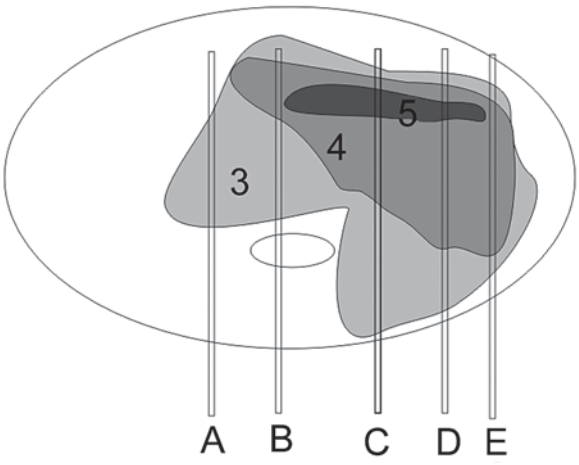

Figure 1 – *Differences between the traditional Gleason score (G) and the modified Gleason score in five hypothetical prostatic needle biopsies with prostate cancer.*
A) G = 3 + 3= 6, modified G = 3 + 3 = 6
B) G = 3 + 4 = 7, modified G = 3 + 4 = 7
C) G = 3 + 4 = 7, modified G= 3 + 5 = 8
D) G = 4 + 3 = 7, modified G = 4 + 5 = 9
E) G = 4 + 3 = 7, modified G = 4 + 3 = 7
The three shades of grey correspond to prostate cancer in Gleason patterns 3, 4 and 5, from lighter to darker respectively.

MATERIALS AND METHODS

A hundred and ten patients suffering from prostate cancer without any previous treatment and who would be referred to a radical prostatectomy agreed to participate in the research. They signed the consent term and sent their needle sextant biopsies, coming from different laboratories, to be reevaluated. Those biopsies had about two cores per sextant, mean total of 12 cores (range 6 to 24 cores). The surgical specimen was processed in the same laboratory, by partial sampling, producing about nine slides per surgical specimen (range 7 to 20), evaluating apex, distal third, mid third, proximal third, bladder neck, right and left seminal vesicles. Thus, the surgical specimen was not processed as a whole. All material was stained with hematoxylin-eosin. All the available slides of the needle biopsy and of the surgical specimens, with or without cancer, were evaluated by the observers.

Three pathologists belonging to different services of Pathological Anatomy examined the slides of the needle and surgical samples of these patients. They did not know the clinical data nor did they know about the pairing between needle biopsy and surgical specimen. They filled in a protocol in which they should classify the primary, secondary and the most aggressive Gleason patterns of each examined area of both specimens. At the end of this task, Gleason score was calculated (the sum of primary and secondary patterns) of each sextant separately (12). The score of the specimen was the highest score found among the evaluated sextants, therefore, the global score was not calculated (4,5,7,15). From the most aggressive Gleason pattern, the tertiary pattern was determined, whenever it was possible. The modified Gleason score was calculated (the sum of primary and tertiary patterns) (12). Similarly, the highest modified score of the examined slides was adopted as the modified Gleason score of the specimen. The primary Gleason pattern was defined as the most frequent Gleason pattern of the sample. The secondary Gleason pattern was the second most frequent pattern, obligatory

higher than 5% of the tumor area (12). When the secondary pattern was less than 5%, the primary pattern was repeated. The tertiary Gleason pattern corresponded to the third Gleason pattern, necessarily more aggressive than the secondary pattern (12). In order to avoid terminology confusion, the Gleason score was called traditional Gleason score, being clearly differentiated from the modified Gleason score.

Data were collected in a data bank and statistically evaluated by Stata program version 9.1 (StatCorp. 4905 Lakeway Dr, College Station, USA). Kappa (K) and weighted Kappa test were used to evaluate the interobserver agreement and the agreement between the Gleason score of the two specimens. The interpretation of the agreement by Kappa value was done by the intervals: K < 0, poor; K = 0-0.2, slight; K = 0.2-0.4, fair; K = 0.4-0.6, moderate; K = 0.6-0.8, substantial; and K = 0.8-1.0, almost perfect (13). In the statistic inferences, in general, the level of significance of 5% was adopted and, consequently, a confidence level of 95% was used.

RESULTS

The samples' mean age was 63.5 +/- 7.7 years old (range 44 to 79 years old). The mean preoperative PSA was 10.2 +/- 8.2 ng/mL (range 1.2 to 53.4 ng/mL). The clinical tumor staging (digital rectal examination) was 46.7% of T1, 47.5% of T2 and 5.8% of T3. In the initial anatomic pathological test, extracapsular tumor extension was found in 17% (pT3a) and in seminal vesicles invasion in 11% (pT3b).

The three pathologists are specialized in the same university even though they nowadays work in different hospitals and laboratories. The experience of working in surgical pathology and the weekly amount of prostate tests, criteria adopted by Taille (13), allow us to classify the observers 1 and 2 as experienced and the observer 3 as less experienced.

Some slides, considered unsatisfactory, were rejected from the research.

In the biopsies, there was a predominance of Gleason pattern 3 in the primary pattern, range from 66% to 86%, and of secondary pattern, range from 63% to 71% among the observers. Similarly, in the surgical specimens, Gleason pattern 3 was more frequent in the primary pattern, being found from 75% to 81%, and as secondary pattern from 60% to 69% of the observations. There was an absolute predominance of Gleason grade 3 in the primary and secondary grade in both specimens. Gleason grade 6 was also predominant in the needle biopsy. In the surgical specimen there was a similar proportion of Gleason score 6 and 7. Table-1 shows the distribution of traditional and modified Gleason scores.

Table 1 *– Frequency of traditional and modified Gleason scores in the specimens according to observers, in percentage.*

Gleason Score	Observer 1				Observer 2				Observer 3			
	Needle Biopsy (N = 98)		Surgical Specimen (N = 109)		Needle Biopsy (N = 71)		Surgical Specimen (N = 85)		Needle Biopsy (N= 100)		Surgical Specimen (N = 110)	
	G	**ModG**	**G**	**ModG**	**G**	**ModG**	**G**	**ModG**	**G**	**ModG**	**G**	**ModG**
4	1%	-	-	-	-	-	-	-	-	-	-	-
5	-	-	-	-	-	-	-	-	-	-	1%	-
6	59%	58%	48%	41%	60%	60%	47%	47%	54%	52%	56%	50%
7	35%	32%	47%	51%	37%	37%	49%	49%	25%	27%	33%	35%
8	4%	6%	2%	3%	3%	3%	4%	4%	2%	2%	7%	9%
9	-	2%	1%	3%	-	-	-	-	1%	1%	2%	5%
10	1%	2%	2%	2%	-	-	-	-	-	-	1%	1%

G = percentage of traditional Gleason score; Mod G = percentage of modified Gleason score.

Interobserver agreement in needle biopsy as regards to primary Gleason grade was reasonable to moderate, according to Kappa. In the surgical specimen, the agreement was moderate to substantial. In the secondary Gleason pattern there was a divergence among the observers, the agreement was generally low, occasionally reasonable. As for the most aggressive Gleason pattern it was from reasonable to moderate (Table-2).

Interobserver agreement of traditional Gleason score in the needle samples was reasonable, with exact agreement among 60% to 68% and agreement +/- 1 Gleason score from 91% to 98%. In the specimens the agreement was from reasonable to moderate, with exact diagnosis from 66% to 71% and accepting difference of one unit from 96% to 99%. Modified Gleason grade presented similar agreement in both specimens, being reasonable to moderate. Exact diagnosis in the biopsy was from 58% to 69% and accepting agreement +/- 1 Gleason score from 86% to 97%. In the specimen the exact diagnosis was from 60% to 64%, accepting divergence of one unit chang-

Table 2 – *Interobserver agreement as regards primary, secondary and the most aggressive Gleason grade in needle biopsies and surgical specimens, considering each patient as an independent event.*

		Observers	Exact Agreement	Expected Agreement	Weighted Kappa	Kappa	Confidence Level	p Value	N
G1	Needle biopsy	1-2-3				0.4020			72
		1-2	85.51%	75.70%	0.4038	0.4036	0.1977 a 0.6095	0.0001	69
		1-3	75.00%	63.80%	0.3079	0.3094	0.1574 a 0.4614	0.0000	96
		2-3	84.29%	64.69%	0.5549	0.5549	0.3368 a 0.8917	0.0000	70
	Surgical specimen	1-2-3				0.5952			85
		1-2	86.90%	64.10%	0.6618	0.6352	0.4391 a 0.8313	0.0000	84
		1-3	85.32%	69.41%	0.5641	0.5201	0.3483 a 0.6919	0.0000	109
		2-3	90.59%	65.02%	0.7509	0.7309	0.5377 a 0.9241	0.0000	85
G2	Needle biopsy	1-2-3				0.1918			72
		1-2	62.32%	58.39%	0.0824	0.0944	-0.1351 a 0.3239	0.2101	69
		1-3	65.63%	55.20%	0.2567	0.2327	0.0399 a 0.4255	0.0090	96
		2-3	64.29%	56.57%	0.1648	0.1776	-0.0454 a 0.4006	0.0593	70
	Surgical specimen	1-2-3				0.2162			85
		1-2	59.52%	53.88%	0.1177	0.1223	-0.0784 a 0.3230	0.1223	84
		1-3	66.06%	51.26%	0.3464	0.3036	0.1363 a 0.4709	0.0002	109
		2-3	63.53%	54.63%	0.2054	0.1962	0.0943 a 0.3959	0.0272	85
Most aggressive G	Needle biopsy	1-2-3				0.4581			72
		1-2	73.91%	48.90%	0.5044	0.4895	0.2788 a 0.7002	0.0000	69
		1-3	72.92%	46.35%	0.4601	0.4951	0.3227 a 0.6675	0.0000	96
		2-3	68.57%	49.47%	0.3892	0.3780	0.1556 a 0.6004	0.0004	70
	Surgical specimen	1-2-3				0.4541			85
		1-2	69.05%	46.46%	0.4610	0.4219	0.2332 a 0.6106	0.0000	84
		1-3	71.56%	42.98%	0.5389	0.5012	0.3468 a 0.6556	0.0000	109
		2-3	68.24%	45.33%	0.4639	0.4190	0.2334 a 0.0646	0.0000	85

G1 = primary Gleason; G2 = secondary Gleason; most aggressive G = the regards most aggressive Gleason grade; weighted Kappa = with linear weight (disagreement by 1 category = 0.67 and disagreement by 2 categories = 0.33). Confidence level of 95% was used.

Table 3 – *Interobserver agreement as regards Gleason score (traditional) and modified Gleason score in needle biopsies and surgical specimens, considering each patient as an independent event.*

		Observers	Exact Agreement	Expected Agreement	Agreement ± 1	Weight Kappa	Kappa	Confidence Level	p Value	N
Gleason score (traditional)	Needle biopsy	1-2-3					0.3641			68
		1-2	68.12%	48.50%	98.55%	0.3996	0.3809	0.1732 a 0.5541	0.0002	69
		1-3	65.63%	42.19%	91.67%	0.4350	0.4054	0.2622 a 0.6676	0.0000	96
		2-3	60.00%	43.63%	92.86%	0.3215	0.2904	0.1119 a 0.4023	0.0007	70
	Surgical specimen	1-2-3					0.4616			84
		1-2	66.67%	45.96%	96.43%	0.4110	0.3832	0.1933 a 0.5765	0.0000	84
		1-3	71.56%	42.62%	99.08%	0.5911	0.5043	0.3566 a 0.6520	0.0000	109
		2-3	70.59%	42.02%	96.47%	0.5235	0.4927	0.3287 a 0.6567	0.0000	85
Modified Gleason score	Needle biopsy	1-2-3					0.3581			68
		1-2	69.57%	47.13%	97.10%	0.4444	0.4243	0.2248 a 0.6238	0.0000	69
		1-3	62.50%	40.61%	86.46%	0.3629	0.3685	0.2280 a 0.5090	0.0000	96
		2-3	58.57%	43.27%	92.86%	0.3080	0.2698	0.0913 a 0.4483	0.0015	70
	Surgical specimen	1-2-3					0.3615			84
		1-2	64.29%	44.52%	94.05%	0.3848	0.3563	0.1760 a 0.5366	0.0001	84
		1-3	63.30%	39.23%	95.41%	0.4901	0.3961	0.2580 a 0.5342	0.0000	109
		2-3	60.00%	40.55%	94.12%	0.3811	0.3271	0.1651 a 0.4891	0.0000	85

Weighted Kappa = with linear weight (disagreement by 1 category = 0.67 and disagreement by 2 categories =0.33). Confidence level of 95% was used.

ing from 94% to 95%. By adopting weighted Kappa, values similar to Kappa (not weighted) were found (Table-3).

Tertiary Gleason pattern was diagnosed in 8%, 0% and 2% of the biopsies and in 8%, 0% and 13% of the surgical specimen according to observers 1, 2 and 3, respectively. Thus, traditional and modified Gleason scores, according to observer 1, were the same in 92% of both specimens. Observer 2 did not consider any pattern as tertiary, having 100% precision between the two Gleason scores. Examiner 3 had 98% of the needle biopsies and 87% of the surgical specimens with the same diagnosis between the two scores.

Traditional and modified Gleason scores were used to evaluate the association among their scores in both specimens by each observer. For observer 1, adopting the traditional score in needle biopsy and in surgical specimen K = 0.24 was found. Adopting the modified score in the biopsy and the traditional one in the specimen, we got K = 0.21. The same happened when using the modified score in the needle biopsy and in the surgical specimen. Examiner 2 did not find any difference in the association of scores between specimens (K = 0.26). When examiner 3 used the traditional score in the needle biopsy and in the surgical specimen, the value for Kappa was 0.18 and when using the modified score in the biopsy and the traditional one in the specimen, Kappa was 0.17. Adopting the traditional Gleason score in both specimens, lower downgrading in needle biopsy was found than by adopting the modified score in both samples (Table-4).

COMMENTS

The sample used reflects a group of patients referred to radical prostatectomy, in other words, young patients, with localized illness and generally low Gleason score. The three observers, also young, had similar academic and professional background and learned the Gleason system during medical residence

Table 4 *– Agreement between traditional and modified Gleason scores in needle biopsies and surgical specimens according to observers. considering each patient as an independent event. Number of needle biopsies whose Gleason score was downgraded or upgraded in relation to the surgical specimen.*

		Exact Agreement	Expected Agreement	Kappa	Confidence Level	p Value	N	Down	Over
Observer 1	Biopsy G x Surgical G	57.73%	44.11%	0.2438	0.0802 a 0.4074	0.0018	97	30 (31%)	11 (11%)
	Biopsy Gmod x Surgical G	54.64%	42.19%	0.2153	0.0609 a 0.3697	0.0031	97	28 (29%)	16 (16%)
	Biopsy Gmod x Surgical Gmod	52.58%	39.45%	0.2168	0.0728 a 0.3608	0.0016	97	33 (34%)	13 (13%)
Observer 2	Biopsy G x Surgical G	60.38%	46.17%	0.2639	0.0319 a 0.4959	0.0129	53	14 (26%)	7 (13%)
	Biopsy Gmod x Surgical G	60.38%	46.17%	0.2639	0.0319 a 0.4959	0.0129	53	14 (26%)	7 (13%)
	Biopsy Gmod x Surgical Gmod	60.38%	46.17%	0.2639	0.0319 a 0.4959	0.0129	53	14 (26%)	7 (13%)
Observer 3	Biopsy G x Surgical G	52.00%	40.79%	0.1893	0.0518 a 0.3268	0.0035	100	20 (20%)	28 (28%)
	Biopsy Gmod x Surgical G	51.00%	40.29%	0.1794	0.0415 a 0.3173	0.0054	100	20 (20%)	29 (29%)
	Biopsy Gmod x Surgical Gmod	48.00%	37.37%	0.1697	0.0361 a 0.3033	0.0064	100	26 (26%)	26 (26%)

Biopsy G = Gleason score in the needle biopsy; Surgical G = Gleason score in the surgical specimen; Biopsy Gmod = modified Gleason score in the needle biopsy; Surgical Gmod = modified Gleason score in the surgical specimen; Down = downgrading of the needle biopsy; Over = overgrading of the biopsy; Confidence level of 95% was used.

in the same institution. Therefore, a good agreement among them would be expected.

Higher agreement of primary Gleason pattern was found in the surgical specimen and not in the needle biopsy. By observing smaller areas, it is expected that more attention would be devoted to a specific area and higher agreement would happen. On the other hand, once the specimen is better represented in tissue extension, the suspected areas with borderline pattern were better examined, resulting in higher agreement. This reflects the difficulties in diagnosing secondary pattern, which besides involving the identification of Gleason patterns, demands tumor volume determination. As a rule, secondary Gleason pattern is the one that is more than 5% of the tumor area and with smaller extension than the primary pattern. Determining the tumor extension is not necessary for the diagnosis of the most aggressive Gleason pattern, the recognition of the worst pattern is sufficient. Glaessgen found a weak agreement as regards the diagnosis of the most aggressive patterns and considered that the difficulty in diagnosing them was big-

ger than in determining their volume (14). The experience did not influence the agreement much because it was not higher between the more experienced observers, what contradicts some authors (15,16).

Interobserver agreement of traditional Gleason score was slightly higher in the surgical specimen than in the needle biopsy. By adopting the modified Gleason score, the agreement was similar in needle biopsy and surgical specimen. In general, adopting weighted Kappa, the agreement values were a little higher, but without altering the previous relations. It is interesting to notice that the modified Gleason score did not show any superiority over traditional score, as Glaessgen reported (14). Evaluating the agreement in relation to the patterns, it is higher in the primary pattern and in the most aggressive one (this is intimately related to tertiary pattern) and too low in the secondary pattern. The modified score would be expected to obtain a higher agreement, but this did not happen. This fact might have happened due to the small number of tertiary pattern diagnosed and, as a result, the two scores were similar. However, this number is similar to the one found in Griffiths' study, where the diagnostic proportion of tertiary Gleason pattern was 6% for general pathologists and 9% for uropathologists, showing weak agreement in relation to tertiary pattern (17). This pattern, in general, refers to patterns 4 or 5, which can present borderline structures making the diagnosis more difficult (3). Generally, the studies regarding the use of tertiary pattern use it in the prognostic evaluation, in surgical specimens. (8-10) Mosse, when evaluating the prognosis of patients with tertiary pattern 5, found a worse prognosis in those with Gleason score 6 or 7 in the surgical specimens. (8) It is known that, statistically, those scores are the most frequent ones.

Considering that prostate cancer is heterogeneous and multicentric (1), it is assumed that the biopsy, which samples a small portion of it, might not represent it efficiently (3-5). Traditionally it is believed that Gleason score in needle biopsy tends to downgrade the surgical specimen, because a less differentiated pattern may not have been sampled in the biopsy (4,12). That was observed by observers 1 and 2. Taking the downgrading concept as a starting-point, some authors suggest the use of modified Gleason score, which would better reflect the real tumor characteristics for it values the most aggressive small patterns (6). The International Society of Urological Pathology (ISUP) on Gleason grading recommends the inclusion of tertiary pattern (modified Gleason score) in needle biopsies. (12) In the surgical specimens, however, it is still recommended to mention the tertiary pattern, whenever it is present, without including it in the score (Gleason score). (12) Considering that the needle biopsy downgrades the score, it was expected that the modified Gleason score would have a better agreement with the traditional Gleason score in the surgical specimen. However this fact could not be demonstrated. The modified score in needle biopsy compared to traditional and modified score in the specimen, presented the same Kappa values or even slightly inferior ones when adopting the traditional score in the biopsy. The best representation in the biopsy was not proved when adopting the modified Gleason score. This fact, as previously mentioned, might have happened due to the low diagnosis of tertiary pattern. Helpap, on the contrary, evaluating slides of 368 patients, found improvement of the exact agreement between the two specimens using the modified Gleason score instead of the traditional score, ranging from 58% to 78% (8). However, he did not use the Kappa test to evaluate the real agreement, nor reported the diagnostic proportion of tertiary pattern.

CONCLUSIONS

In this study, the modified Gleason score did not prove to be superior in reproducibility compared to the traditional Gleason score, both in the needle biopsy and in the surgical specimen. Contrary to what was expected, the use of the modified score in the biopsy was not superior to the traditional score, comparing to the Gleason scores of the specimen. Within the aim of the study, the modified Gleason score was not superior to the traditional one. These conclusions might be due to the methodology used, as well as to the observers involved. Isolated morphological analysis is based in criteria of low reproducibility. It is necessary to reevaluate the association between the two Gleason scores, using different samples with a higher amount of tertiary pattern.

ACKNOWLEDGMENT

Dr. Sergio G. Veloso has a CNPq Grant, Ministry of Technology, Brazil.

CONFLICT OF INTEREST

None declared.

REFERENCES

1. Aihara M, Wheeler TM, Ohori M, Scardino PT: Heterogeneity of prostate cancer in radical prostatectomy specimens. Urology. 1994; 43: 60-6; discussion 66-7.
2. Billis A, Pompeo AC: Adenocarcinoma da próstata. Int Braz J Urol. 2003; 29 (suppl 1): 27-34.
3. Montironi R, Mazzuccheli R, Scarpelli M, Lopez-Beltran A, Fellegara G, Algaba F: Gleason grading of prostate cancer in needle biopsies or radical prostatectomy specimens: contemporary approach, current clinical significance and sources of pathology discrepancies. BJU Int. 2005; 95: 1146-52.
4. Lopez-Beltran A, Mikuz G, Luque RJ, Mazzucchelli R, Montironi R: Current practice of Gleason grading of prostate carcinoma. Virchows Arch. 2006; 448: 111-8.
5. Egevad L, Allsbrook WC Jr, Epstein JI: Current practice of Gleason grading among genitourinary pathologists. Hum Pathol. 2005; 36: 5-9.
6. Pan CC, Potter SR, Partin AW, Epstein JI: The prognostic significance of tertiary Gleason patterns of higher grade in radical prostatectomy specimens: a proposal to modify the Gleason grading system. Am J Surg Pathol. 2000; 24: 563-9.
7. Rioux-Leclercq NC, Chan DY, Epstein JI: Prediction of outcome after radical prostatectomy in men with organ-confined Gleason score 8 to 10 adenocarcinoma. Urology. 2002; 60: 666-9.
8. Helpap B, Egevad L: The significance of modified Gleason grading of prostatic carcinoma in biopsy and radical prostatectomy specimens. Virchows Arch. 2006; 449: 622-7.
9. Epstein JI, Allsbrook WC Jr, Amin MB, Egevad LL; ISUP Grading Committee: The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma. Am J Surg Pathol. 2005; 29: 1228-42.
10. De la Taille A, Viellefond A, Berger N, Boucher E, De Fromont M, Fondimare A, et al.: Evaluation of the interobserver reproducibility of Gleason grading of prostatic adenocarcinoma using tissue microarrays. Hum Pathol. 2003; 34: 444-9.
11. Glaessgen A, Hamberg H, Pihl CG, Sundelin B, Nilsson B, Egevad L: Interobserver reproducibility of modified Gleason score in radical prostatectomy specimens. Virchows Arch. 2004; 445: 17-21.
12. Hollenbeck BK, Bassily N, Wei JT, Montie JE, Hayasaka S, Taylor JM, et al.: Whole mounted radical prostatectomy specimens do not increase detection of adverse pathological features. J Urol. 2000; 164: 1583-6.
13. Kunz GM Jr, Epstein JI: Should each core with prostate cancer be assigned a separate gleason score? Hum Pathol. 2003; 34: 911-4.
14. Glaessgen A, Hamberg H, Pihl CG, Sundelin B, Nilsson B, Egevad L: Interobserver reproducibility of percent Gleason grade 4/5 in total prostatectomy specimens. J Urol. 2002; 168: 2006-10.
15. Allsbrook WC Jr, Mangold KA, Johnson MH, Lane RB, Lane CG, Epstein JI: Interobserver reproducibility of Gleason grading of prostatic carcinoma: general pathologist. Hum Pathol. 2001; 32: 81-8. Erratum in: Hum Pathol 2001; 32: 1417.
16. Allsbrook WC Jr, Mangold KA, Johnson MH, Lane RB, Lane CG, Amin MB, et al.: Interobserver reproducibility of Gleason grading of prostatic carcinoma: urologic pathologists. Hum Pathol. 2001; 32: 74-80.
17. Griffiths DF, Melia J, McWilliam LJ, Ball RY, Grigor K, Harnden P, et al.: A study of Gleason score interpretation in different groups of UK pathologists; techniques for improving reproducibility. Histopathology. 2006; 48: 655-62.

Accepted after revision:
August 8, 2007

Correspondence address:
Dr. Sergio Geraldo Veloso
Rua Henrique Benfenatti, 237
São João del-Rei, MG, 36307-042, Brazil
Fax: + 55 32 3371-8003
E-mail: velososg@ig.com.br

EDITORIAL COMMENT

At a consensus conference organized in 2005 by the International Society of Urological Pathology (ISUP), the Gleason grading system underwent its first systematic revision (1). The purpose of the meeting was to standardize both the perception of histological patterns and how the grade information is compiled and reported. One of the decisions of the ISUP working group was that high-grade tumor of any quantity on needle biopsy should be included in the Gleason score. The ISUP recommendations contribute to a general shift upwards of the Gleason scores and it may be necessary to re-iterate some previous studies on grading of prostate cancer. Helpap et al. recently compared conventional and modified Gleason grading in radical prostatectomy specimens and preoperative biopsies and reported on the distribution of modified Gleason score and its correlation with other prognostic factors such as age, stage and serum PSA (2-4). Few studies have been performed on interobserver reproducibility of this new variant of Gleason grading.

In a study by Glaessgen et al., the reproducibility of modified Gleason grading among four genitourinary pathologists was analyzed using a set of 69 consecutive radical prostatectomy specimens (5). Mean weighted kappa for conventional and modified Gleason score were 0.56 (range 0.52-0.66) and 0.58 (range 0.49-0.74), respectively. This study was carried out before the ISUP consensus meeting was held and only addressed the effect of inclusion of tertiary patterns of higher grade in the Gleason score. Hence, recent changes in pattern recognition were not taken into account. Furthermore, the ISUP recommendations to include tertiary higher patterns in the score pertained to needle biopsies, while the study by Glaessgen et al. was done on radical prostatectomy specimens only (5).

Veloso et al., in this paper, present a similar study on the reproducibility of a modified Gleason grading, now done on both needle biopsies and radical prostatectomy specimens. Again, only the effect of inclusion of tertiary higher patterns was studied. In needle biopsies a weighted kappa of 0.36 was reached both with conventional and modified Gleason grading. In radical prostatectomy specimens, the weighted kappa was 0.46 and 0.36, respectively. This interobserver agreement was slightly lower than that of previous studies. For example, in a biopsy study on conventional Gleason score by Glaessgen et al., a weighted kappa of 0.48 to 0.55 (mean 0.51) was reached among 4 genitourinary pathologists using a consecutive series of needle biopsies from 69 men (279 glass slides) (6). Allsbrook et al. circulated 46 needle biopsies containing prostatic carcinoma among 10 genitourinary pathologists (6). The weighted kappa for Gleason score ranged from 0.56 to 0.70. However, the biopsies of this series were selected rather than consecutive which may lead to a better reproducibility.

From studies performed so far, it seems that the interobserver reproducibility of the Gleason grading remains essentially the same with modified Gleason grading and results are probably more influenced by the study design.

Revision of a grading system may be necessary when we gain new knowledge of the biology of cancer. However, it must also be remembered that a revision has consequences in terms of modified prognostic impact of a certain grade and also warrants new studies to verify the value of the novel grading system (7). Whether modified Gleason grading of needle biopsies is superior as predictor of prognosis remains to be seen.

REFERENCES

1. Epstein JI, Allsbrook WC Jr., Amin MB, Egevad L: The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason grading of prostatic carcinoma. Am J Surg Pathol. 2005; 29: 1228-42.
2. Helpap B, Egevad L: Correlation of modified Gleason grading with pT stage of prostatic carcinoma after radical prostectomy. Anal Quant Cytol Histol (in press).
3. Helpap B, Egevad L: Correlation of modified Gleason grading of prostate carcinoma with serum PSA, age and tumor extent in needle biopsy specimens. Anal Quant Cytol Histol (in press).
4. Helpap B, Egevad L: The significance of modified Gleason grading of prostatic carcinoma in biopsy and

radical prostatectomy specimens. Virchows Arch. 2006; 449: 622-7.
5. Glaessgen A, Hamberg H, Pihl CG, Sundelin B, Nilsson B, Egevad L: Interobserver reproducibility of modified Gleason score in radical prostatectomy specimens. Virchows Arch. 2004; 445: 17-21.
6. Glaessgen A, Hamberg H, Pihl CG, Sundelin B, Nilsson B, Egevad L: Interobserver reproducibility of percent Gleason grade 4/5 in prostate biopsies. J Urol. 2004; 171: 664-7.
7. Albertsen PC, Hanley JA, Barrows GH, Penson DF, Kowalczyk PD, Sanders MM, et al.: Prostate cancer and the Will Rogers phenomenon. J Natl Cancer Inst. 2005; 97: 1248-53.

Dr. Lars Egevad
Department of Pathology & Cytology
Karolinska Hospital
Stockholm, Sweden
E-mail: lars.egevad@ki.se

EDITORIAL COMMENT

This paper by Dr Veloso et al. deal with interobserver agreement of Gleason score and modified Gleason score (1) in needle biopsy and in surgical specimen of prostate cancer. This group of authors found that the modified Gleason score was not superior in the agreement between the biopsy score and the specimen, or in interobserver reproducibility.

The Gleason grading system is a powerful tool to prognosticate and aid in the treatment of men with prostate cancer. The needle biopsy Gleason score correlates with virtually all other pathologic parameters, including tumor volume and margin status in radical prostatectomy specimens, serum PSA levels and many molecular markers. The Gleason score assigned to the tumor at radical prostatectomy is the most powerful predictor of progression following radical prostatectomy. However, there exist significant deficiencies in the practice of this grading system. Not only does there exist problems among practicing pathologists but also a relative lack of interobserver reproducibility among experts.

Correlation

There have been several studies addressing the correlation between Gleason scores in needle biopsies and corresponding radical prostatectomy specimens. Although earlier studies used the thicker (14-gauge) needle biopsies (2,3), more recent series based on thin-core (18-gauge) needles used in conjunction with biopsy guns attached to transrectal ultrasound. Sextant or other modes of systematic sampling are typically performed in the more current series. In a recent compilation of data on 3,789 patients from 18 studies, exact correlation of Gleason scores was found in 43% of cases and correlation plus or minus one Gleason core unit in 77% of cases (4). Under-grading of carcinoma in needle biopsy is the most common problem, occurring in 42% of all reviewed cases. Importantly, over-grading of carcinoma in needle biopsies may also occur, but this was only found in 15% of cases. In general, adverse findings on needle biopsy accurately predict adverse

findings in the radical prostatectomy specimen, whereas favorable findings on the needle biopsy do not necessarily predict favorable findings in the radical prostatectomy specimens in large part due to sampling error.

Sources of Discrepancies

Sampling error

Perhaps the most important factor is sampling error, which relates to the small amount of tissue removed by thin-core needle biopsies. The average 20-mm, 18-gauge core samples approximately 0.04% of the average gland volume (40 cc). The most common type of sampling error occurs when there is a higher grade component present within the radical prostatectomy specimen, which is not sampled on needle biopsy (5). This typically occurs when a needle biopsy tumor is graded as Gleason score 3 + 3 = 6. In the radical prostatectomy, there exists a Gleason pattern 4, which was not sampled on the biopsy, resulting in a prostatectomy Gleason 3 + 4 = 7.

In some instances, under-grading results from an attempt to grade very tiny areas of carcinoma, so-called minimal or limited adenocarcinoma (6). Scores of minimal adenocarcinoma in needle biopsies show a reasonably strong correlation with radical prostatectomy scores, but the Gleason scores do not have the same power to predict extra-prostatic extension and positive margin status as they do in non-minimal carcinomas (6).

Over-grading can result from sampling error in cases where the high-grade pattern is selectively represented in needle biopsy. It may only represent a very minor element in the radical prostatectomy specimen. Even the same cancer focus may have different grades depending on the area sampled.

Borderline cases

The other source of discrepancy between biopsy and radical prostatectomy is borderline cases. In the description of the Gleason grading system, there are some cases that are right at the interface between two different patterns where there will be inter-observer variability and possible even intra-observer variability (7).

Pathology error

Pathology error is most frequently seen when pathologists assigned a Gleason score of ≤ 4 on a needle biopsy, which in fact was Gleason score 5-6. Many pathologists under-grade needle biopsies by confusing quantitative changes with qualitative changes. When there is a limited focus of small glands of cancer on needle biopsy, by definition this is a Gleason pattern 3. Gleason pattern 3 consists of small glands with an infiltrative pattern. Biopsying truly low-grade adenocarcinoma of the prostate could not result in just a few neoplastic glands but rather would be more extensive, as low-grade adenocarcinoma grows as nodules of closely packed glands rather than infiltrating in and amongst normal glands.

Under-grading may result from difficulty in recognizing an infiltrative growth pattern or failing to recognize the presence of small areas of gland fusion (7).

Pathologists' education and experience

The pathologists' experience in grading thin-core needle biopsies can also influence overall correlation with radical prostatectomy results. With experience, pathologists recognize grading pitfalls; in particular, the fact that Gleason scores of 4 and lower are almost non-existent in needle biopsy situation. Furthermore, small areas of fusion in the presence of a predominantly grade 3 background are recognized and will yield a Gleason score of 7, which often correlates well with radical prostatectomy results (8).

Intra-observer and interobserver variability

Reproducibility studies can be categorized as intra-observer and interobserver. For investigations of intra-observer agreement of Gleason grades, exact agreement was reported in 43% to 78% of cases (8,9), and agreement within plus or minus one Gleason score unit was reported in 72% to, 87% of cases. Gleason wrote tat he duplicated exactly his previous histologic scores approximately 50% of times. Highly variable levels of interobserver agreement on Gleason scores have also been reported, with range of 36% to 81% for exact agreement and 69% to 86% observers within plus or minus one Gleason score unit. Improvements in Gleason grading reproducibility can be achieved by

recognizing problematic areas and educating physicians via meetings, courses, website tutorials, and publications that specifically focus on the Gleason grading system (10).

REFERENCES

1. Epstein JI, Allsbrook WC Jr., Amin MB, Egevad LL, ISUP Grading Committee: The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma. Am J Surg Pathol. 2005; 29: 1228-42.
2. Garnett JE, Oyasu R, Grayhack JT: The accuracy of diagnostic biopsy specimens in predicting tumor grades by Gleason's classification of radical prostatectomy specimens. J Urol. 1984; 131: 690-3.
3. Mills SE, Fowler JE Jr: Gleason histologic grading of prostatic carcinoma. Correlation between biopsy and prostatectomy specimens. Cancer. 1986; 57: 346-9.
4. Humphrey PA: Prostate Pathology. Chicago, ASCP Press. 2003; p. 138.
5. Algaba F, Chivite A, Santaularia JM, Oliver A: Evidence of the radical prostatectomy Gleason score in the biopsy Gleason score. Actas Urol. Esp. 2004; 28: 21-6.
6. Rubin MA, Dunn R, Kambham PA, Misick CP, O'Toole KM: Should a Gleason score be assigned to a minute focus of carcinoma on prostate biopsy? Am J Surg Pathol. 2000; 24: 1634-40.
7. Steinberg DM, Sauvageot J, Piantadosi S, Epstein JI: Correlation of prostate needle biopsy and radical prostatectomy Gleason grade in academic and community setting. Am J Surg Pathol. 1997; 21: 566-76.
8. Cintra ML, Billis A: Histologic grading of prostatic adenocarcinoma. Intra-observer reproducibility of the Mostofi, Gleason, and Böcking grading systems. Int Urol Nephrol. 1991; 23: 449-54.
9. Özdamar SO, Sarikaya S, Yildiz L, Atilla MK, Kandemir B, Yildiz S: Intra-observer and interobserver reproducibility of WHO and Gleason histologic grading systems in prostatic adenocarcinomas. Int Urol Nephrol. 1996; 28: 73-7.
10. Egevad L, Allsbrook WC, Epstein JI: Current practice of Gleason grading among genitourinary pathologists. Hum Pathol. 2005; 36: 5-9.

Dr. Rodolfo Montironi
Institute of Pathological Anatomy
Polytechnic University of the Marche Region
Ancona, Italy
E-mail: r.montironi@univpm.it

Dr. Liang Cheng
Dept. of Pathology and Laboratory Medicine
Indiana University School of Medicine
Indianapolis, IN, USA
E-mail: liang_cheng@yahoo.com

EDITORIAL COMMENT

In the original Gleason system, the most common and second most common grade patterns are added to arrive at the Gleason score with tertiary patterns not factored in. For example, in a needle biopsy with Gleason score 3 + 4 = 7, a smaller tertiary component of very high grade pattern 5 tumor would not be factored in. In the Consensus Conference on Updating the Gleason grading system, it was recommended that a tertiary component of higher grade tumor on biopsy be included within the Gleason

score by adding the most common and highest grade patterns. In the above example, this would result in a Gleason score of 3 + 5 = 8. This study by Veloso et al. found that the interobserver reproducibility for the modified biopsy Gleason score was not superior to the routine Gleason score and was also not more accurate in predicting radical prostatectomy Gleason score. The major limitation of their study, as they acknowledge, is the limited number of cases with a tertiary pattern on biopsy, ranging from 0%, 2%, to 8% amongst the three observers out of a total of 110 patients. With such small numbers, it would be impossible to show any differences between the routine and modified Gleason score. In a recent paper on 2,370 men with prostate cancer, Patel et al. also found that Gleason score 7 with tertiary pattern 5 was uncommon, occurring in 1.5% of cases (1). However, they documented that Gleason score 7 tumor on biopsy with tertiary pattern 5 has the same prognosis as Gleason score 8 tumor when treated by radiotherapy or radical prostatectomy. These findings are in concert with several studies that have documented the same adverse prognostic significance of tertiary pattern 5 in radical prostatectomy specimens. The growing body of evidence suggests that Gleason score 3 + 4 with tertiary pattern 5, whether on biopsy or radical prostatectomy, should be considered as Gleason score 8.

REFERENCE

1. Patel AA, Chen M, Renshaw AA, Da Amico AV: PSA failure following definitive treatment of prostate cancer having biopsy Gleason score 7 with tertiary grade 5. JAMA 2007; 298: 1533-38.

Dr. Jonathan I. Epstein
Dept of Pathology, Urology and Oncology
The Johns Hopkins Hospital
Baltimore, MD, USA
E-mail: jepstein@jhmi.edu