Virchows Arch (2004) 445:17–21
DOI 10.1007/s00428-004-1034-0

ORIGINAL ARTICLE

Axel Glaessgen · Hans Hamberg · Carl-Gustaf Pihl · Birgitta Sundelin · Bo Nilsson · Lars Egevad

Interobserver reproducibility of modified Gleason score in radical prostatectomy specimens

Received: 16 February 2004 / Accepted: 20 April 2004 / Published online: 20 May 2004

Abstract The Gleason score (GS) of prostate cancer is calculated by adding primary and secondary Gleason grades with patterns occupying less than 5% of the tumour often not included despite their probable prognostic significance. A modified Gleason score (mGS) comprising primary and tertiary patterns of higher grade has been proposed, but its interobserver variability has yet to be elucidated. Slides from 69 consecutive prostatectomy specimens were circulated among four genitourinary pathologists. GS and mGS were assessed, and results were compared in pairs. Mean weighted kappa for GS and mGS were 0.56 (range 0.52–0.66) and 0.58 (range 0.49–0.74), respectively. The difference between GS and mGS was 0, 1 and 2 score units in 66%, 26% and 8%, respectively, mean 0.41 score units (range 0.24–0.51). The increment was greater for transition-zone tumours than for peripheral-zone tumours (0.63 and 0.35 score units, respectively, P=0.002). An odd mGS (5, 7 or 9) was more often given than an odd GS (77% and 62%, respectively, P<0.001). Disagreement between observers greater than 1 score unit was more common with mGS than GS (18% and 4%, respectively, P<0.001). In conclusion, overall mean weighted kappa for interobserver reproducibility of mGS is at least as high as that of GS. However, there is a clustering of mGS in odd scores, and severe disagreement is more commonly observed than with GS. Training of mGS assessment would possibly improve agreement. Tertiary Gleason patterns need to be better defined.

Keywords Prostatic neoplasms/pathology · Prostatic neoplasms/Gleason grading · Prostatic neoplasms/classification · Prostatectomy, human

The study was approved by the ethics committee of The Karolinska Institute (01–443).

A. Glaessgen · H. Hamberg · B. Sundelin · L. Egevad (✉)
Department of Pathology and Cytology,
Karolinska Hospital,
Stockholm, Sweden
e-mail: Lars.Egevad@onkpat.ki.se
Tel.: +46-8-51775492
Fax: +46-8-331909

C.-G. Pihl
Department of Pathology,
Sahlgrenska University Hospital,
Gothenburg, Sweden

B. Nilsson
Department of Oncology and Pathology,
Karolinska Institute,
Stockholm, Sweden

Introduction

Few tumours display such remarkable morphological heterogeneity as prostate cancer. Usually more than one histological pattern is present in prostatectomy specimens, and cancer of a single grade is less common [1, 20]. The Gleason grading system is designed to take into account this complexity by including both the primary and secondary pattern (i.e. the most prevalent and the second most prevalent pattern) in the Gleason score (GS). However, in addition to the primary and secondary patterns, a tertiary pattern may be present. In a study from 1994, Aihara et al. found that over 50% of prostatectomy specimens contained cancer of at least three different Gleason patterns [1]. For convenience, the term tertiary pattern will be used in this article to designate only a grade that is higher than the primary and secondary patterns. Patterns occupying less than 5% of the tumour are often ignored when the GS is assessed [6]. Hence, if more than 95% of the tumour is occupied by a single pattern, GS is obtained by doubling that pattern. In those cases, the predominant pattern will be considered both primary and secondary, and a higher grade occupying less than 5% of the tumour will be referred to as tertiary.

Pan et al. showed that a tertiary grade has an adverse impact on prognosis after prostatectomy [19]. Men with a tertiary pattern 4 or 5 in a GS 5 or 6 tumour had a prognosis similar to those with a GS 7. The authors proposed that a tertiary component is reported together with the GS. However, it was also suggested that if future

studies would corroborate their findings, a modified Gleason score (mGS) calculated by adding primary and tertiary patterns might replace the GS. In a recent study on natural history of transurethral resection of the prostate (TURP)-detected prostate cancers, we demonstrated that mGS was a stronger predictor of disease-specific survival than GS and that both were independently significant prognosticators [8].

Before recommending mGS for general use, several issues have to be clarified, among them the reproducibility of this measure. Several studies on interobserver variability of Gleason grading have been published [3, 4, 7, 16], and, recently, we have shown that reproducibility of percentage Gleason grade 4/5 in prostatectomy specimens and prostate needle biopsies is at least as good as that of the GS [11, 12]. However, to our knowledge, reproducibility of the mGS has not been reported previously. In this study, interobserver reproducibility of this measure is investigated on a consecutive series of prostatectomy specimens.

Materials and methods

This study was based on a consecutive series of 69 radical prostatectomy specimens received from January 2000 to December 2000 at the Department of Pathology and Cytology at the Karolinska Hospital. The same set of slides was used in our recent study on reproducibility of percentage Gleason grade 4/5 [11].

The prostatectomy specimens were handled according to a standardised protocol previously described [11]. Briefly, the prostate was fixed overnight in 10% buffered formalin, inked and sliced horizontally at 4-mm intervals. The slices were cut in 2–6 segments, and the entire prostate was subsequently blocked in standard cassettes, sectioned at 4 μm and stained with haematoxylin and eosin. Cancer was outlined on the slides. The tumour originated from the peripheral zone in 54 cases and from the transition zone in 15 cases. Seminal vesicle invasion, extraprostatic extension and positive margins were present in 13 (19%), 41 (59%) and 33 (48%) cases, respectively.

Similar to our previous study, a single slide from the main tumour focus of each prostatectomy specimen was selected by one of the authors (L.E.) [11]. One of the slides with the greatest amount of cancer was chosen. When multiple slides with a substantial amount of cancer were available, the slide that most closely represented the GS of the main tumour was used. The glass slides were circulated among four pathologists who specialised in genitourinary pathology (L.E., H.H., C.G.P. and B.S.) for a total of 276 responses. No consensus training preceded the study. Slides were reviewed, and a mGS was calculated by adding primary (dominating) Gleason grade and tertiary Gleason pattern of higher grade when present. When no tertiary component was found, mGS was identical to conventional GS. The GS and percentage Gleason grade 4/5 given in our previous study were used [11]. Slides were circulated twice for the two studies, but GS and percentage Gleason grade 4/5 were only assessed at the first slide review, and mGS only at the second. By comparing the responses of each of the four participants with those of the other three, six pairwise comparisons were obtained for each of the 69 cases, for a total of 414 comparisons. Results were compared in pairs, and weighted kappa values were calculated for GS and mGS.

Results

The GSs given by the four observers ranged from 5 to 9 (mean 6.35–6.88, overall mean 6.68, Table 1). A GS of 6 or 7 was assigned to 81–90% of cases (mean 86%). Average GS was higher in peripheral-zone than in transition-zone tumours (6.83 and 6.13, respectively, $P<0.001$).

The mGSs given by the observers ranged from 5 to 9 (mean 6.86–7.35, overall mean 7.09, Table 1). The mean difference between mGS and GS was 0.24–0.51 (overall mean 0.41). A mGS of 6 or 7 was assigned to 70–90% of cases (mean 78%). Average mGS was higher in peripheral-zone than in transition-zone tumours (7.18 and 6.77, respectively, $P=0.015$).

The score distribution of GS and mGS differed significantly ($P<0.001$, Fig. 1). Of all responses, 6%, 32%, 54%, 6% and 3% were GSs 5, 6, 7, 8 and 9, respectively. For mGS, this distribution was 3%, 15%, 63%, 8% and 11%, respectively.

In 183 (66%) of the responses, GS and mGS were identical (Table 2). There was an increment from GS to mGS of 1 or 2 units in 72 (26%) and 21 (8%) responses, respectively. The increment was greater for transition-zone tumours than for peripheral-zone tumours (0.63 and 0.35, respectively, $P=0.002$). Of GSs 5, 6, 7 and 8 tumours, 56% (9 of 16 tumours), 56% (49 of 88), 20% (29 of 148) and 47% (8 of 17), respectively, had a tertiary pattern and, thus, a higher mGS than GS. Tertiary patterns 4 and 5 were reported 56 (20%) and 37 (13%) times, respectively. In peripheral-zone tumours, tertiary patterns 4 and 5 were reported in 17% and 13%, respectively, and in transition-zone tumours, these patterns were reported in 33% and 13%, respectively. When the increment from GS

Table 1 Distribution of Gleason scores and modified Gleason scores

Observer no.	5	6	7	8	9	Average
Gleason score						
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. of tumours (%)	16 (6)	88 (32)	148 (54)	17 (6)	7 (3)	276
Modified Gleason score						
1	5	11	39	6	8	7.01
2	1	9	45	7	7	7.14
3	2	14	48	2	3	6.86
4	1	7	41	7	13	7.35
No. of tumours (%)	9 (3)	41 (15)	173 (63)	22 (8)	31 (11)	276

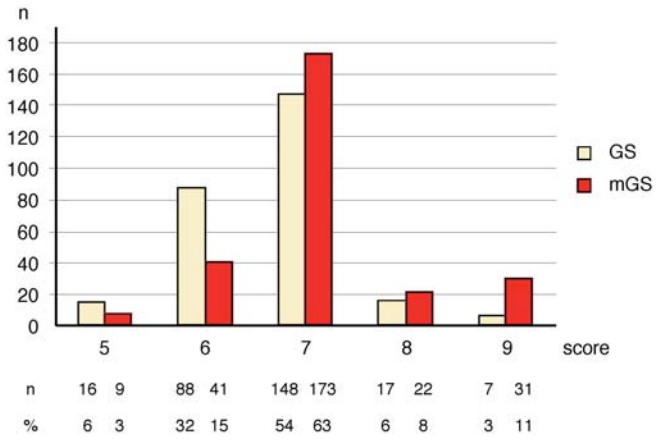

Fig. 1 Distribution of Gleason scores (GS) and modified Gleason scores (mGS)

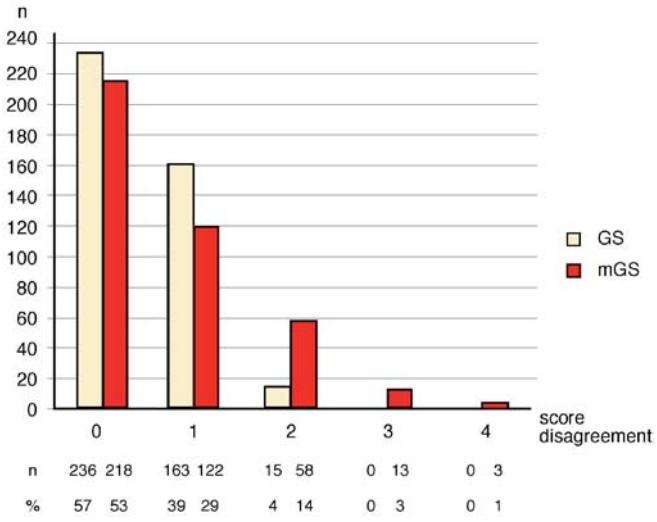

Fig. 2 Score disagreement in pairwise comparison of Gleason scores (GS) and modified Gleason scores (mGS)

Table 2 Gleason scores compared with modified Gleason scores

		Modified Gleason score					
		5	6	7	8	9	Totals
Gleason score	5	9	2	5	0	0	16
	6	0	39	49	0	0	88
	7	0	0	119	13	16	148
	8	0	0	0	9	8	17
	9	0	0	0	0	7	7
	Totals	9	41	173	22	31	276

to mGS was two units, a tertiary pattern 5 was more often responsible for the score shift than a tertiary pattern 4 (76% and 24%, respectively), while tertiary patterns 5 were less commonly seen in cases with one unit increment from GS to mGS (29% and 71%, respectively, $P<0.001$). A difference of more than one grade unit between the two patterns included in the score was more common with mGS than GS (6% and 0.7%, respectively, $P<0.001$, Table 3).

An odd mGS (5, 7 or 9) was more often given than an odd GS (77% and 62%, respectively, $P<0.001$). Of 93 changed scores, 57 (61%) were shifted from an even GS to an odd mGS (from 3+3=6 to 3+4=7 or from 4+4=8 to 4+5=9), 15 (16%) from an odd GS to an even mGS (from 2+3=5 to 2+4=6 or from 3+4=7 to 3+5=8) and 21 (23%) from an odd GS to another odd GS (from 3+2=5 to 3+4=7 or from 4+3=7 to 4+5=9) (Table 3).

All responses were compared in pairs, and differences in mGS were calculated. In 218 (53%) of the 414 pairs, agreement was exact. The disagreement between paired responses was one, two, three and four score units in 30%, 14%, 3% and 1%, respectively (Fig. 2). Disagreement of

Table 3 Score changes from Gleason scores to modified Gleason scores

Gleason score	Modified Gleason score	Totals	Peripheral-zone tumours	Transition-zone tumours	Unchanged	Changed
2+3	2+3	1	0	1	1	
2+3	2+4	2	0	2		2
2+3	2+5	0	0	0		0
3+2	3+2	8	0	8	8	
3+2	3+4	5	0	5		5
3+2	3+5	0	0	0		0
3+3	3+3	37	32	5	37	
3+3	3+4	49	36	13		49
3+3	3+5	0	0	0		0
2+4	2+4	1	0	1	1	
4+2	4+2	1	0	1	1	
3+4	3+4	93	78	15	93	
3+4	3+5	13	10	3		13
4+3	4+3	26	25	1	26	
4+3	4+5	16	11	5		16
4+4	4+4	9	9	0	9	
4+4	4+5	8	8	0		8
4+5	4+5	5	5	0	5	
5+4	5+4	2	2	0	2	
		276	216	60	183	93

Table 4 Weighted kappa values for Gleason scores and modified Gleason scores of the four observers

	Weighted kappa values				Totals
Observer no.	1	2	3	4	
Gleason score					
1		0.67	0.627	0.669	
2	0.670		0.507	0.496	
3	0.627	0.507		0.412	
4	0.669	0.496	0.412		
Average	0.655	0.558	0.515	0.526	0.563
Modified Gleason score					
1		0.735	0.668	0.501	
2	0.735		0.586	0.514	
3	0.668	0.586		0.487	
4	0.501	0.514	0.487		
Average	0.635	0.612	0.580	0.501	0.582

more than one score unit was more common with mGS than GS (18% and 4%, respectively, $P<0.001$).

The mean-weighted kappa value for GS and for mGS rendered by each observer and compared with each of the other three observers was 0.52–0.66 (overall mean 0.56) and 0.49–0.74 (overall mean 0.58), respectively (Table 4). The observer concordance was lower for transition-zone tumours than for peripheral-zone tumours, both with GS (weighted kappa 0.43 and 0.54, respectively) and with mGS (weighted kappa 0.52 and 0.59, respectively).

Discussion

The GS is a well-established prognostic parameter for prostate cancer [2, 5, 18]. The value of this grading as a predictor of recurrence after prostatectomy [10, 15] or radiotherapy [13] and as determinant of disease-specific survival in patients on deferred treatment [9] has been shown in several studies. The GS is calculated by adding primary and secondary Gleason grade and patterns occupying less than 5% of the tumour are often not included [6]. Hence, Gleason grading differs from many other grading systems in that the GS does not necessarily include the highest grade present in the tumour. Several authors have demonstrated the prognostic significance of a high-grade prostate cancer component [8, 9, 17, 19, 21]. The Stanford group has shown that the percentage of the tumour that is occupied by Gleason grades 4 or 5 (percentage Gleason grade 4/5) predicts lymph-node metastases [17] and recurrence [21] after radical prostatectomy. In a Swedish study of TURP-detected prostate cancer managed by watchful waiting, percentage Gleason grade 4/5 was a better predictor of disease-specific survival than the GS [8]. Furthermore, men with GS 6 tumours with a focal component of Gleason pattern 4 suffered a significantly higher risk of cancer-specific death than men with pure GS 6 tumours. A mGS calculated by adding the primary and tertiary pattern was also a better predictor of prognosis than the GS. In another study, Pan et al. showed that likelihood of recurrence after radical prostatectomy increased if a tertiary pattern of higher grade was found [19].

To be practically useful, a grading system must have satisfactory reproducibility. Numerous studies on interobserver reproducibility of the GS have been published in recent decades [3, 4, 7, 16]. Allsbrook et al. provided a detailed review [4]. The results of reproducibility studies are often difficult to compare, because the number of observers as well as number and type of specimens have not been the same. Some of the studies have been preceded by a tutorial, and others include selected sets of specimens rather than consecutive series. In the present study, we used the same consecutive series of prostatectomy slides as in a recent study on reproducibility of percentage Gleason grade 4/5 [11]. The same group of specialists in genitourinary pathology reviewed the slides. This gave us the opportunity to compare our results from three different grading systems, i.e. GS, percentage Gleason grade 4/5 and mGS.

We have previously shown that interobserver reproducibility of percentage Gleason grade 4/5 in prostatectomy specimens and needle biopsies is at least as good as that of the GS [11, 12]. In the present study, according to kappa analysis, interobserver reproducibility of mGS in prostatectomy specimens was at least as good as that of the GS. However, severe disagreement was more common with mGS than with GS, possibly because tertiary patterns are not sufficiently defined. The minimum amount of a pattern that is required for being identified at all is not clear. Occasional incomplete, fused or branching glands may be interpreted either as cutting artefacts or as a focal pattern 4. Similarly, occasional small solid epithelial structures may be interpreted either as cutting artefacts or as a focal pattern 5. It remains unclear whether pattern 5 requires clusters of individual cells, solid strands or solid nests seen at lower than ×40 magnification or whether this pattern also may be diagnosed when such structures are found interspersed among cancerous glands. Thus, to improve reproducibility of mGS, the definition of focal high-grade components has to be clarified.

Although weighted kappa for mGS was lower for transition-zone tumours than for peripheral-zone tumours (0.519 and 0.594, respectively), the zonal difference was less than that previously reported for GS (0.433 and 0.540, respectively) [11]. In tumours with three or more Gleason patterns, the proportion of the tumour occupied by the patterns is decisive for the GS. In tumours of transition-zone origin, Gleason patterns 2, 3 and 4 are often simultaneously present in the same tumour, resulting in GSs 5, 6 or 7, depending on the amount of each grade, and this may explain the lower weighted kappa for GS of transition-zone tumours compared with peripheral-zone tumours [11]. With mGS, this source of interobserver variability is to some extent avoided because a tertiary pattern is included in the score regardless of its proportion of the tumour.

The score distribution of GS and mGS differed significantly. The overall mean difference between mGS and

GS was 0.41, and in only 66% of the responses were GS and mGS identical. We believe that it must be clearly stated which of these scores is used. Redefining GS as the sum of primary and tertiary patterns is potentially confusing, and comparison with results from previous studies using the old definition of GS will turn out to be difficult. It is preferable that a new term is introduced to designate the sum of primary and tertiary patterns, tentatively mGS, a term that also was used by Pan et al. [19].

In this series, an odd mGS was more often given than an odd GS, leading to an increased clustering in a few scores. Although the Gleason system was originally designed as a nine-tier system (score 2–10), few of these scores are actually used. A reason to ignore minor components is that relatively few cases would otherwise be assigned an even score (such as GS 3+3=6), especially in radical prostatectomy specimens [1, 20]. A disadvantage of the mGS is that its value as predictor of aggressive behaviour may be questioned when a majority of cases are lumped into a single score.

For several reasons, we believe that it is too early to abandon the conventional GS. Before recommending a general use of mGS, previous results on its prognostic value need to be corroborated. The limited number of studies that have addressed the prognostic value of a mGS [8, 19, 21, 22] are as yet insufficient to support the introduction of a new grading system. Furthermore, the ability of preoperative core biopsies to predict mGS in prostatectomy specimens needs to be investigated. Theoretically, a minimal component of tertiary pattern would easily go undetected by needle biopsies, decreasing the practical utility of a mGS. Additionally, reproducibility of a mGS has to be improved, and consensus must be reached on the definition of tertiary Gleason patterns. No systematic training preceded the present study, but Kronz et al. [14] and Egevad [7] described that the use of reference images or web-based tutorials improves reproducibility of Gleason grading. Similar techniques would possibly improve interobserver reproducibility of mGS.

References

1. Aihara M, Wheeler TM, Ohori M, Scardino PT (1994) Heterogeneity of prostate cancer in radical prostatectomy specimens. Urology 43:60–66
2. Allsbrook WC Jr, Mangold KA, Yang X, Epstein JI (1999) The Gleason grading system: an overview. J Urol Pathol 10:141–157
3. Allsbrook WC Jr, Mangold KA, Johnson MH, Lane RB, Lane CG, Amin MB, Bostwick DG, Humphrey PA, Jones EC, Reuter VE, Sakr W, Sesterhenn IA, Troncoso P, Wheeler TM, Epstein JI (2001) Interobserver reproducibility of Gleason grading of prostatic carcinoma: urologic pathologists. Hum Pathol 32:74–80
4. Allsbrook WC Jr, Mangold KA, Johnson MH, Lane RB, Lane CG, Epstein JI (2001) Interobserver reproducibility of Gleason grading of prostatic carcinoma: general pathologists. Hum Pathol 32:81–88
5. Bostwick DG, Grignon DJ, Hammond ME, Amin MB, Cohen M, Crawford D, Gospadarowicz M, Kaplan RS, Miller DS, Montironi R, Pajak TF, Pollack A, Srigley JR, Yarbro JW (2000) Prognostic factors in prostate cancer. College of American Pathologists Consensus Statement 1999. Arch Pathol Lab Med 124:995–1000
6. Deshmukh N, Foster CS (1998) Grading prostate cancer. In Foster CS, Bostwick DG (eds) Pathology of the prostate. Saunders, Philadelphia, pp 191–227
7. Egevad L (2001) Reproducibility of Gleason grading of prostate cancer can be improved by the use of reference images. Urology 57:291–295
8. Egevad L, Granfors T, Karlberg L, Bergh A, Stattin P (2002) Percent Gleason grade 4/5 as prognostic factor in prostate cancer diagnosed at transurethral resection. J Urol 168:509–513
9. Egevad L, Granfors T, Karlberg L, Bergh A, Stattin P (2002) Prognostic value of the Gleason score in prostate cancer. BJU Int 89:538–542
10. Epstein JI, Partin AW, Sauvageot J, Walsh PC (1996) Prediction of progression following radical prostatectomy. A multivariate analysis of 721 men with long-term follow-up. Am J Surg Pathol 20:286–292
11. Glaessgen A, Hamberg H, Pihl CG, Sundelin B, Nilsson B, Egevad L (2002) Interobserver reproducibility of percent Gleason grade 4/5 in total prostatectomy specimens. J Urol 168:2006–2010
12. Glaessgen A, Hamberg H, Pihl CG, Sundelin B, Nilsson B, Egevad L (2003) Interobserver reproducibility of percent Gleason grade 4/5 in prostate biopsies. J Urol 171:664–667
13. Green GA, Hanlon AL, Al-Saleem T, Hanks GE (1998) A Gleason score of 7 predicts a worse outcome for prostate carcinoma patients treated with radiotherapy. Cancer 83:971–976
14. Kronz JD, Silberman MA, Allsbrook WC, Epstein JI (2000) A web-based tutorial improves practising pathologists' Gleason grading of images of prostate carcinoma specimens obtained by needle biopsy: validation of a new medical education paradigm. Cancer 89:1818–1823
15. Lerner SE, Blute ML, Bergstralh EJ, Bostwick DG, Eickholt JT, Zincke H (1996) Analysis of risk factors for progression in patients with pathologically confined prostate cancers after radical retropubic prostatectomy. J Urol 156:137–143
16. Lessells AM, Burnett RA, Howatson SR, Lang S, Lee FD, McLaren KM, Nairn ER, Ogston SA, Robertson AJ, Simpson JG, Smith GD, Tavadia HB, Walker F (1997) Observer variability in the histopathological reporting of needle biopsy specimens of the prostate. Hum Pathol 28:646–649
17. McNeal JE, Villers AA, Redwine EA, Freiha FS, Stamey TA (1990) Histologic differentiation, cancer volume, and pelvic lymph node metastasis in adenocarcinoma of the prostate. Cancer 66:1225–1233
18. Montironi R, Mazzucchelli R, Kwast T (2003) Morphological assessment of radical prostatectomy specimens. A protocol with clinical relevance. Virchows Arch 442:211–217
19. Pan CC, Potter SR, Partin AW, Epstein JI (2000) 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 24:563–569
20. Ruijter ET, van de Kaa CA, Schalken JA, Debruyne FM, Ruiter DJ (1996) Histological grade heterogeneity in multifocal prostate cancer. Biological and clinical implications. J Pathol 180:295–299
21. Stamey TA, McNeal JE, Yemoto CM, Sigal BM, Johnstone IM (1999) Biological determinants of cancer progression in men with prostate cancer. JAMA 281:1395–1400
22. Stamey TA, Yemoto CM, McNeal JE, Sigal BM, Johnstone IM (2000) Prostate cancer is highly predictable: a prognostic equation based on all morphological variables in radical prostatectomy specimens. J Urol 163:1155–1160