

Interobserver reproducibility of Gleason grading: evaluation using prostate cancer tissue microarrays

M. Burchardt · R. Engers · M. Müller · T. Burchardt ·
R. Willers · J. I. Epstein · R. Ackermann ·
H. E. Gabbert · A. de la Taille · M. A. Rubin

Received: 13 January 2008 / Accepted: 19 March 2008 / Published online: 8 April 2008
© Springer-Verlag 2008

Abstract

Objectives Due to PSA screening and increased awareness, prostate cancer (PCa) is identified earlier resulting in smaller diagnostic samples on prostate needle biopsy. Because Gleason grading plays a critical role in treatment planning, we undertook a controlled study to evaluate interobserver variability among German pathologists to grade small PCas using a series of tissue microarray (TMA) images.

Methods We have previously demonstrated excellent agreement in Gleason grading using TMAs among expert genitourinary pathologists. In the current study, we identified 331 TMA images (95% PCa and 5% benign) to be evaluated by an expert PCa pathologist and subsequently by practicing pathologists throughout Germany. The images were presented using the Bacus Webslide Browser on a CD-ROM. Evaluations were kept anonymous and participant's scoring was compared to the expert's results.

Results A total of 29 German pathologists analysed an average of 278 images. Mean percentage of TMA images which had been assigned the same Gleason score (GS) as done by the expert was 45.7%. GSs differed by no more than one point (± 1) in 83.5% of the TMA samples evaluated. The respondents were able to correctly assign a GS into clinically relevant categories (i.e. <7 , 7 , >7) in 68.3% of cases. A total of 75.9% respondents under-graded the TMA images. Gleason grading agreement with the expert reviewer correlated with the number of biopsies evaluated by the pathologist per week. Years of diagnostic experience, self-description as a urologic pathologist or affiliation with a university hospital did not correlate with the pathologist's performance.

Conclusion The vast majority of participants under-graded the small tumors. Clinically relevant GS categories were correctly assigned in 68% of cases. This raises a potentially significant problem for pathologists, who have not had as much experience evaluating small PCas.

M. Burchardt and R. Engers contributed equally to this work.

M. Burchardt (✉)
Department of Urology, Medizinische Hochschule Hannover,
Carl-Neuberg Str. 1, 30625 Hannover, Germany
e-mail: burchardt.martin@mh-hannover.de

R. Engers · H. E. Gabbert
Department of Pathology, Heinrich-Heine University,
Duesseldorf, Germany

M. Müller · R. Ackermann
Department of Urology, Heinrich-Heine University,
Duesseldorf, Germany

T. Burchardt
Department of Dermatology, Heinrich-Heine University,
Duesseldorf, Germany

R. Willers
Department of Biostatistics, Heinrich-Heine University,
Duesseldorf, Germany

J. I. Epstein
Johns Hopkins School of Medicine,
Johns Hopkins Hospitals, Baltimore, MD, USA

A. de la Taille
CHU Henri Mondor, Creteil, France

M. A. Rubin
Harvard Medical School,
Brigham and Women's Hospital,
Boston, MA, USA

Keywords Prostate cancer · Gleason grading · Interobserver variances · Tissue microarray

Introduction

Multiple studies demonstrate the value of the Gleason grading system for PCa as a predictor of final pathological stage and prognosis (Epstein et al. 1993; Oesterling et al. 1987). Urologists use the Gleason grade for patient counseling, predicting outcome, and treatment decisions (Epstein et al. 1993, 1996; Partin et al. 2001; Rubin et al. 2002). Therefore, accurate histological grading is clinically significant as many treatment decisions will relay on the GS.

Pitfalls of histological grading of prostate carcinoma are inter- and intraobserver reproducibility, under-grading of the original biopsy, and “lumping” of grades. “Interobserver reproducibility” means the repeatability of Gleason grading among different pathologists and presents a significant problem in histopathological grading of prostate tumors. To quote the founder of the current PCa staging system, Donald Gleason wrote: “As part of the art of histopathology, grading undoubtedly will not be performed equally well by all those who attempt it” (Gleason 1992).

While broadly used in the United States, European pathologist have only more recently adopted the Gleason grading system and therefore may have less experience with this system. Previous studies from several countries using different experimental approaches showed interobserver differences in Gleason grading by no more than 1 score point in 10–83% of the cases (Allsbrook Jr et al. 2001a, b; McLean et al. 1997; Ozdamar et al. 1996; Di Loreto et al. 1991; Svanholm and Mygind 1985; De las Morenas et al. 1988).

Tissue microarrays (TMA) have recently been developed to help with the high-throughput evaluation of biomarkers (Allsbrook Jr et al. 2001a). This approach promises to enhance tissue-based molecular research by allowing improved conservation of tissue resources and experimental reagents, improved internal experimental control, and increased sample numbers per experiment (Allsbrook Jr et al. 2001a). Given the limited sample size (0.6 mm diameter), TMA technology may also be an excellent method to compare Gleason grading in small tissue samples of PCa by different pathologists.

The current study used this technique by creating a CD-ROM with 331 high-quality TMA images (95% PCa and 5% benign). The images were evaluated by an expert PCa pathologist (J.I.E.) and then reviewed by 29 practicing pathologists throughout Germany. The goal of this study was to evaluate Gleason grading of German pathologists in comparison to a highly recognized expert in this field and to evaluate the TMA CD-ROM as a potential teaching device for training in Gleason grading.

Material and methods

Tissue microarray construction

A prostate tissue microarray block containing 480 tissue cores (each core 0.6 mm in diameter) was constructed from formalin-fixed, paraffin-embedded radical prostatectomy tissues from men with clinically localized, previously untreated PCa. Specific representative regions of cancer and noncancerous prostate tissue for tissue microarray biopsy were circled on face section of donor blocks from these cases by a pathologist (M.A.R.). Array construction was performed using a manual tissue microarray instrument (Beecher instruments, Silver Spring, MD, USA) as described previously (Kononen et al. 1998). A single 4- μ m section from this array was stained with hematoxylin–eosin. Each array disk of tissue on this tissue array slide is referred to as a tissue microarray spot.

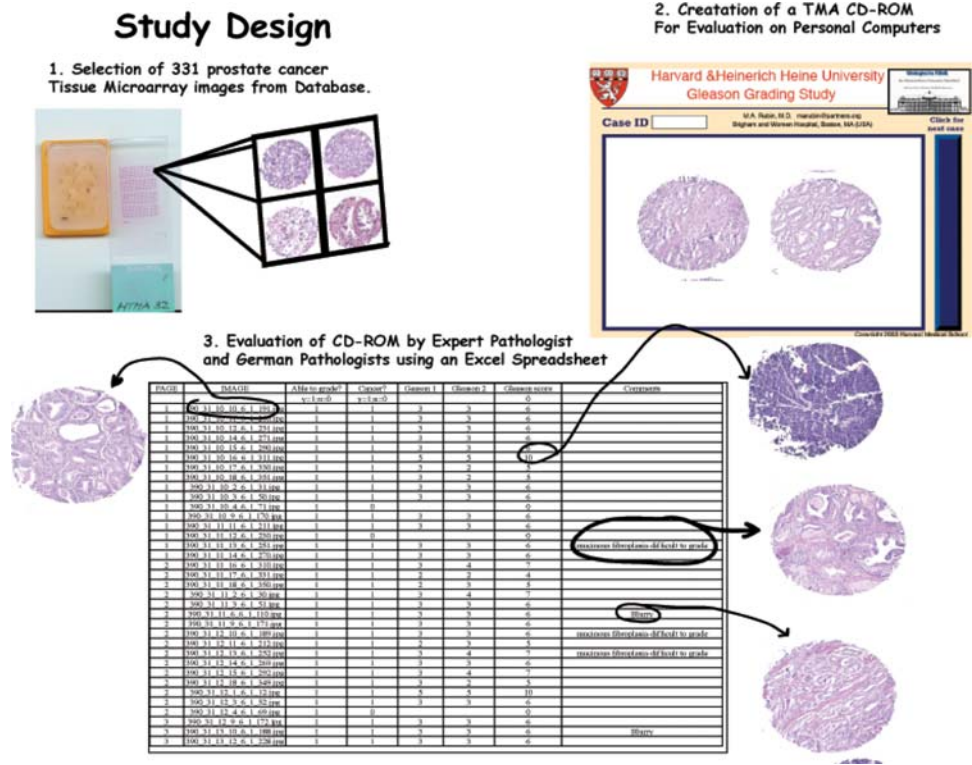
Tissue microarray spot imaging

In brief, digital images were captured using the Bliss Imaging System (Bacus Labs, Chicago, IL, USA). Tissue microarray spot imaging was performed as previously described by the authors (Bova et al. 2001).

Prostate cancer TMA images

Prostate Cancer TMA images were selected from a large image bank available at the Brigham and Women’s Hospital (Harvard Medical School, Boston, MA, USA). These TMA images included localized PCa with a range of GSs (<6: 4%; 6: 51%; 7: 24%; >7: 21%). The most common primary Gleason patterns were 3 (69%) and 4 (22%). A total of 5% of the samples were benign. A total of 331 images were used to create a CD-ROM containing a web-browser (Bacus Labs, Chicago, IL, USA). Each image was captured at 200 \times original magnification and contained the entire TMA spot (0.6 mm diameter). These CD-ROMs were returned by 29 of 45 initially addressed German pathologists (12 university-affiliated, 11 community hospital, 6 private practice). Dr. J. Epstein (Department of Pathology, Johns Hopkins School of Medicine, Baltimore, MD, USA) served as the expert prostate pathologist for this study. He reviewed the images using the same CD-ROM and worksheet and determined that a Gleason score could be assigned to 310 images. Figure 1 presents the design used for this study. In brief, the pathologists were asked to evaluate the TMA images. They were requested to determine whether the image contained PCa, whether the image allowed Gleason grading and if so, to provide the GS as well as the respective primary and secondary Gleason patterns. They also had the opportunity to enter a comment.

Fig. 1 Schematic view of the methods used for the current study. Images were selected from a collection of over 400 cases to include a range of Gleason grades and stored on a CD ROM. We intentionally kept sub-pare images to determine if quality of the images would be consistently determined. We also included images of benign prostate tissue (5%). Each participant received a CD-ROM and a work sheet for each image on the CD-ROM for evaluation. Only after all study pathologists had returned their evaluation, they were provided with the expert's results. Analysis compared the results of the 29 participants to the expert pathologist



Participants were unaware of Dr. Epstein's evaluation until all study participants had returned their score sheets.

Questionnaire

All participants had to answer an attached short questionnaire with the following set of questions (translated from the German):

- How many years have you been doing surgical pathology, including your residency?
- How many prostate needle biopsy sets do you analyze per week?
- How many radical prostatectomy specimens do you analyze per week?
- Do you believe that this approach (i.e. using this CD-ROM) may be an appropriate approach to teach residents or other pathologists in Gleason grading?

Statistics

Exact agreement was measured by percentage of agreement and Cohen's Kappa, named κ value in this paper. For the GSs, the weighted κ value was calculated using the scores as weights. For the clinical relevant categories <7, 7, and >7 the group mean was used as weight. A bias in grading (over-/under-grading) was measured as difference to the reference pathologist. All the statistics were computed sep-

arately for each pathologist and pooled over all pathologists. The influence of the factors asked in the questionnaire on the quality of Gleason grading was tested for significance by the Kruskal–Wallis test and Spearman rank correlation, respectively, depending on the scale type of the factors (metric/categorical). The general under-grading was tested for significance by means of the Wilcoxon rank sum test. This was performed for each pathologist and overall for the personal means. A *P* value lower than 0.05 was considered as significant. In case of multiple testing the *P* values were adjusted by the method of Bonferroni–Holm. Unadjusted *P* values lower than 0.3 were reported as trend.

Statistical analyses were performed using SAS Version 9.1.3 (SAS Institute Inc., Cary, NC, USA).

Results

Demographics and questionnaire

Evaluation of the CD-ROM and the questionnaires were returned by 29 pathologists. On average a GS was assigned to 278 images by each participant (range 95–310), resulting in a total number of 8,070 GS-assignments (for 331 TMA images). Complete image data were available from 18 pathologists (62%). The mean number of years as an attending staff pathologist was 10 years (range <1–27). The mean number of prostate needle biopsies evaluated per week was

14 (range 1–100). The mean number of radical prostatectomy specimens evaluated per week was 3 (range 0–15). Six (21%) of 29 participants designated themselves to be specialists in urologic pathology. Twelve (38%) pathologists are working in a university setting, 11 (41%) in community hospitals, and the remaining 6 (21%) in private practice. Of all the factors analyzed only the number of prostate biopsies reviewed by the pathologist per week tend to correlate with a high concordance to the expert pathologist, but due to multiple testing the *P* value obtained (*P* = 0.027) cannot be considered as significant. Other factors such as years of general training as pathologist, self-description as “uropathologist”, the number of weekly graded prostatectomy specimens and the kind of affiliation were of no importance. Twenty-six of 29 participants regarded the CD as a valuable learning tool for residents of pathology.

Interobserver variability

Consensus in the diagnosis of cancer based on TMA images alone was seen in $93.6 \pm 6.8\%$ of the cases (range 72.6–99.4%), giving rise to a mean κ value of 0.58 ± 0.21 (range 0.18–0.91, Table 1). Moreover, there was a high consensus that most TMA cancer samples could be subjected to Gleason grading (85.2%). Exact agreement in Gleason grading between the expert and the study participants was found in $45.7 \pm 14.5\%$ of all the TMA images analyzed (range 17.4–69.5%, Table 1). Gleason scorings differed by no more than one point (± 1) from the expert’s evaluation in $83.5 \pm 12.5\%$ of the TMA samples (range 44.1–99.3%, Table 1). Using the clinically relevant categories <7, 7, and >7, agreement in Gleason grading with the expert was observed in $68.3 \pm 6.3\%$ of the cases (range

Table 1 Level of agreement between each German pathologist and reference pathologist

German pathologist	Gleason score			Cancer diagnosis	
	<i>n</i>	Agreement exact <i>n</i> (%)	Agreement with tolerance ± 1 <i>n</i> (%)	<i>n</i>	Agreement <i>n</i> (%)
1	294	67 (22.8%)	170 (57.8%)	324	315 (97.2%)
2	310	208 (67.1%)	291 (93.9%)	330	324 (98.2%)
3	276	73 (26.4%)	185 (67.0%)	331	294 (88.8%)
4	247	43 (17.4%)	109 (44.1%)	327	279 (85.3%)
5	297	112 (37.7%)	261 (87.9%)	330	315 (95.5%)
6	307	198 (64.5%)	285 (92.8%)	331	327 (98.8%)
7	300	163 (54.3%)	264 (88.0%)	318	316 (99.4%)
8	299	164 (54.8%)	280 (93.6%)	331	322 (97.3%)
9	310	73 (23.5%)	279 (90.0%)	331	315 (95.2%)
10	300	148 (49.3%)	274 (91.3%)	331	325 (98.2%)
11	307	126 (41.0%)	277 (90.2%)	327	318 (97.2%)
12	308	214 (69.5%)	292 (94.8%)	325	321 (98.8%)
13	305	146 (47.9%)	260 (85.2%)	327	321 (98.2%)
14	280	163 (58.2%)	246 (87.9%)	331	320 (96.7%)
15	195	71 (36.4%)	148 (75.9%)	329	239 (72.6%)
16	294	157 (53.4%)	259 (88.1%)	331	312 (94.3%)
17	279	89 (31.9%)	237 (84.9%)	299	292 (97.7%)
18	305	139 (45.6%)	230 (75.4%)	331	320 (96.7%)
19	304	178 (58.6%)	267 (87.8%)	331	319 (96.4%)
20	269	157 (58.4%)	236 (87.7%)	331	289 (87.3%)
21	291	109 (37.5%)	230 (79.0%)	331	311 (94.0%)
22	288	152 (52.8%)	250 (86.8%)	325	309 (95.1%)
23	303	118 (38.9%)	257 (84.8%)	331	325 (98.2%)
24	238	80 (33.6%)	169 (71.0%)	295	267 (90.5%)
25	261	111 (42.5%)	212 (81.2%)	320	294 (91.9%)
26	288	183 (63.5%)	286 (99.3%)	326	307 (94.2%)
27	222	121 (54.5%)	201 (90.5%)	329	246 (74.8%)
28	298	101 (33.9%)	222 (74.5%)	327	316 (96.6%)
29	95	25 (26.3%)	58 (61.1%)	175	149 (85.1%)
Total	8,070	3,689 (45.7%)	6,735 (83.5%)	9,305	8,707 (93.6%)

Table 1 presents the results of the individual participating pathologist by displaying the numbers (percentage) of TMA spots, for which each participating pathologist reached the GS of our reference pathologist either exactly or within a tolerance ± 1 . The last two columns show the level of agreement in diagnosis of cancer based on the TMA spots

55.7–81.4%), yielding a mean weighted κ value of 0.57 ± 0.09 (range 0.40–0.76). More specifically, if 5 GS categories were used (2–4, 5–6, 3 + 4, 4 + 3, 8–10), reference and study pathologists agreed in 56.4% of TMA images. Table 2 displays the comparison between reference and participating pathologists regarding the different grading categories Gleason 2–4, 5–6, 3 + 4, 4 + 3, and 8–10, respectively.

The mean negative deviation of GSs between reference and participating pathologists pooled over all participating pathologists was -0.67 (range 0 to -5), the mean positive deviation was 0.30 (range 0 to $+5$). When focusing on the individual results of each respondent, a significant deviation from the reference pathologist became evident for 25 (86.2%) of 29 participants. Overall this deviation consisted of a highly significant (<0.001) under-grading.

More specifically, 22 of 29 (75.9%) respondents significantly under-graded the TMA images, whereas 3 respondents exhibited a significant over-grading. The mean deviation in Gleason grading between study pathologists and the expert was -0.41 (range -1.54 to 0.36). In total, 38.9% of the images were under-graded, whereas only 15.4% were over-graded (Fig. 2). Under-grading was most commonly seen for GSs 7 through 9, while over-grading was most commonly observed for GSs 4 and 5 (Table 3). Under-grading resulted mostly from misevaluation of Gleason pattern 3 as Gleason pattern 2 and Gleason pattern 4 as Gleason pattern 3, respectively. In contrast, over-grading was mostly due to misinterpretation of Gleason pattern 3 as Gleason pattern 4 and Gleason pattern 2 as Gleason pattern 3,

Table 2 Level of agreement between participating pathologists and reference pathologist regarding clinically relevant groups (2–4, 5–6, 3 + 4, 4 + 3, 8–10)

Study pathologists	Reference pathologist					Sum	
	2–4	5–6	3 + 4	4 + 3	8–10		
Gleason score	2–4	16	<i>606</i>	<u>69</u>	<u>33</u>	<u>14</u>	738
	5–6	7	2,863	<u>553</u>	<u>262</u>	<u>178</u>	3,863
	3 + 4	0	478	327	<u>196</u>	<u>208</u>	1,209
	4 + 3	0	158	149	168	<u>166</u>	641
	2, 5	0	4	1	3	7	15
	8–10	0	112	117	197	1,178	1,604
	Sum	23	4,221	1,216	859	1,751	8,070

Table 2 presents agreement and disagreement in TMA Gleason grading between German and reference pathologist. For example (third column): For all TMA spots with reference GSs 5–6 a total of 4,221 assignments has been given by the study pathologists. Of these, 2,863 (67.8%) exactly matched the reference Gleason category. In contrast, a GS 2–4 was erroneously assigned in 606 (14.4%) of these cases (*bold*: exact agreement, *italic*: over-grading, *underlined*: under-grading). Agreement regarding these groups was 56.4%. A clear trend towards under-grading is noted in the assignments of the participants when compared to the reference pathologist

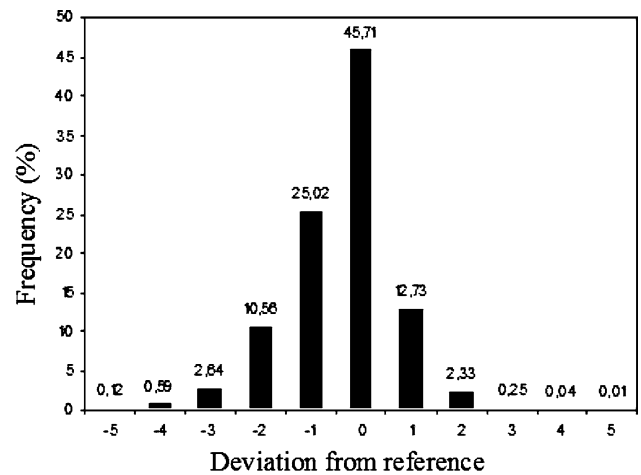


Fig. 2 Comparison of Gleason scores of study to expert pathologist. A bar chart demonstrates the frequency of GS deviation for study versus reference pathologist (J.I.E.). This chart demonstrates that only 45.7% ($n = 3,689/8,070$) of the study pathologists scores matched exactly the expert pathologists scores. Furthermore, the data shows that the majority of images were given a lower grade than that assigned by the expert pathologist

Table 3 Frequency of correct and incorrect interpretation of distinct Gleason Scores by study participants

Reference GS	Undergraded (%)	Correctly graded (%)	Overgraded (%)
4	26.1	43.5	30.4
5	43.4	31.7	24.9
6	33.4	47.9	18.6
7	44.2	40.7	15.1
8	55.4	35.8	8.8
9	45.8	38.3	15.9
10	27.2	72.8	0

Under-grading was most commonly seen for GSs 7 through 9, while over-grading was most commonly observed for GSs 4 and 5

respectively. In some cases, however, the differences between study pathologists and the expert were much more pronounced. Thus, Gleason pattern 5 was misevaluated as Gleason pattern 2 or 3, and Gleason pattern 4 as Gleason pattern 1 or 2, respectively. Conversely, Gleason patterns 2 and 3 were misinterpreted as Gleason patterns 5. Consequently, in terms of the final GS, maximal deviations of ± 5 scoring points were observed. For example GS 7 was substantially under-graded as GS 2 ($n = 3$) and GS 8 as GS 3 ($n = 3$). Conversely, in one case GS 5 was substantially over-graded as GS 10. Most importantly, 49% of all study pathologists under-graded TMA images at least once by 4 or 5 points in GS, respectively, and 10% of all the study pathologists over-graded TMA images at least once by 4 or 5 points in GS, respectively (Fig. 3).

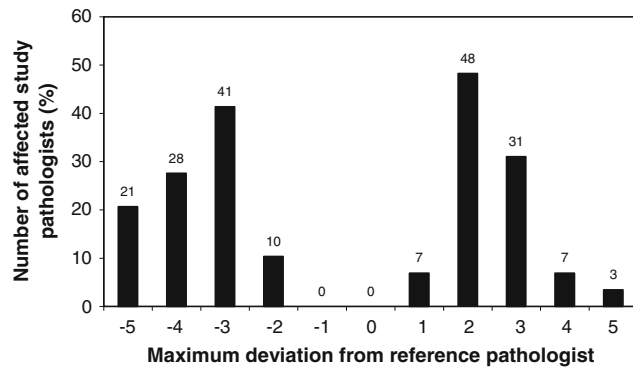


Fig. 3 Maximum deviations of Gleason scores between study pathologists and the expert. A total of 49% of all study pathologists undergraded TMA images at least once by 4 or 5 points in GS, respectively, and 10% of all study pathologists overgraded TMA images at least once by 4 or 5 points in GS, respectively

Discussion

The present study used a set of 331 TMA images to evaluate Gleason grading of German pathologists in comparison to a highly recognized expert in this field (J.I.E). In an earlier study, this expert assigned exactly the same GS in 179 of 221 TMA PCa images (81%) and agreed in 100% within a range of 1 when compared to another expert genitourinary pathologist (T. Wheeler) (Bova et al. 2001). This performance rate is similar if not better than that previously reported for direct microscopic Gleason grading. Therefore, this study and others indicate that TMA images are a useful and reliable tool to test the quality of Gleason grading (Bova et al. 2001; De la Taille et al. 2003; Egevad 2001). In addition, the 0.6 mm diameter is consistent with the extent of up to 20% of tumors diagnosed by prostate needle biopsy in a PSA screened population (personal observation MAR).

In contrast to earlier studies and the previous study with ten French uropathologists (De la Taille et al. 2003), we almost tripled the amount of participating pathologists (10 vs. 29). The number of participants and the high number of graded histologies ($n = 310$) enhanced the statistical value of the work. On average, 278 judgments were given for a total of 331 samples. This displays that most participants did not provide complete data sets. However, only pathologists who submitted consecutive data beginning with sample one participated to prevent a selection bias. Furthermore, the study protocol required exact GSs in this setup instead of grouping into Gleason categories, as many earlier studies have pointed out the importance of exact distinction between each score (Epstein et al 1993, 1996; Partin et al. 2001). Another important addition to previous studies is the fact that GSs of all participants were compared to GSs of a highly recognized genitourinary pathologist (J.I.E).

The question of how to interpret our main results is debatable—is an exact agreement in GSs in 45.7% of the cases and an agreement within the range of ± 1 in 83.5% of the cases excellent, acceptable, or inadequate? Should the correct assignment of GSs into clinically relevant categories in 68% ($<7, 7, >7$) of cases be satisfying?

Gleason wrote, that “histologic grading, no matter how well defined, is not a verifiable scalar measurement” (Gleason 1992). On the other hand, given the fact that agreement among expert uropathologists is so much better (Bova et al. 2001) and given the strong impact of GSs on treatment decisions, the interobserver variability observed in the present study is unsatisfying. Nevertheless, in comparison to studies in other countries, the results obtained in Germany are neither better nor worse. In these studies with different experimental setups the levels of agreement between the different pathologists were low, the exact agreement ranging from 18.8 to 82.5%, and the agreement in GS ± 1 ranging from 10 to 83% (Allsbrook Jr et al. 2001a, b; McLean et al. 1997; Ozdamar et al. 1996; Di Loreto et al. 1991, 1988; Svanholm and Mygind 1985). There are diverse explanations for these low levels of agreement. First, experience with the Gleason grading system is important. Allsbrook et al. (2001b) found that pathologists, who learned the system during residency tended to apply this grading system better than those pathologists, who learned on the job. Second, many of the studies examine grading between pathologists at different institutions. Allsbrook also suggests that pathologists from the same department tend to have similar grading patterns (Allsbrook Jr et al. 2001b). Third, the variability in interobserver agreement might be due to the small sample size of the TMA spots (De la Taille et al. 2003). However, we have previously tested intraobserver and interobserver reproducibility amongst two expert genitourinary pathologists and found results consistent with standard slides (Ozdamar et al. 1996). In contrast to studies with standard slides, the present study on TMA images did not provide the option to zoom into the image. This, however, may have been helpful in some cases to determine whether tumor glands are fused together or not, which is important in distinguishing Gleason pattern 4 from 3 and consequently would influence the final GS. Nevertheless, we have previously shown that experts evaluating prostate cancer on TMAs can reliably grade these samples with levels of agreement similar to those observed in biopsy-based studies (Ozdamar et al. 1996).

A well-known problem of Gleason grading is a tendency of under-grading and this is supported by the present study. We found that 38.9% of the images were under-graded, whereas only 15.4% were over-graded. Previous studies showed that the most common and important misinterpretation in Gleason grading is under-grading GSs 5, 6, and 7 (King 2000; Egevad et al. 2001; Kronz et al. 2000; Altay

et al. 2001; Gregori et al. 2001; Smith et al. 2002; Mikami et al. 2003). Thus, Allsbrook et al. (2001a) found 47% under-grading of score 5–6 tumors, 43% under-grading of score 7 tumors, and 25% under-grading of score 8–10 tumors. Kronz et al. (2000) observed 30.6% under-grading of score 5–6 tumors and 33.2% of score 7 tumors. In contrast, in our study under-grading was most commonly observed in tumors of score 7–9 (frequency rates 44.2–55.4%, Table 3) but a high rate of under-grading (43.4%) was also seen in score 5 tumors. Under-grading of score 7–9 tumors mostly resulted from mistaking Gleason pattern 4 for pattern 3, whereas under-grading of score 5 tumors mostly resulted from mistaking pattern 3 for pattern 2. This is in accordance with a study of Allsbrook et al. (2001a) and suggests that training in Gleason grading should put special emphasis on the characteristics of these three Gleason patterns. Nevertheless, other Gleason patterns have been misinterpreted in our study as well giving rise to differences in GSs between the expert and study pathologists of up to ± 4 and ± 5 score points, respectively. Although such strong differences in GSs were rather rare (<1%), almost half of all the study participants (49%) showed at least once an under-grading of 4 or 5 score points and about 10% of all the study pathologists showed at least once an over-grading by 4 or 5 score points. Given the high relevance of GSs on treatment planning, differences like these strongly argue in favor of a systematic training in Gleason grading for all pathologists evaluating prostate tissue samples. In this context, the CD-ROM used in the present study might be a valuable training device to improve agreement, as judged by most of the study participants, who obtained the reference pathologist's evaluation at the end of the study.

In search for demographic factors, which could possibly reflect the quality in Gleason grading of our study, only the number of prostate biopsies reviewed by the pathologist per week correlated significantly with a high concordance to the expert pathologist. In contrast, years of general training as pathologist, self-description as "uropathologist", the number of weekly graded prostatectomy specimens and the kind of affiliation were of no importance.

Conclusions

The present study demonstrates that Gleason grading of small PCa areas by German pathologists is characterized by a strong tendency of under-grading when compared to a highly recognized expert pathologist (J.I.E.) in this field. Exact agreement in GS between German pathologists and the expert was only found in 46% of the cases and major differences of up to ± 5 score points became evident. The quality of Gleason grading correlated positively with the

number of prostate biopsies evaluated per week. Given the high impact of Gleason grading on treatment planning, systematic training is required and the CD-ROM used in this study might be a valuable teaching device in this respect. Re-review of the images by the respondents at the close of this study should provide a more precise explanation as to why they differed with the expert.

Acknowledgments We would like to thank all participating pathologists, since this study could not have been possible without their extensive work. As we promised anonymity to the study participants their names and affiliations are not mentioned here.

References

- Allsbrook WC Jr, Mangold KA, Johnson MH, Lane RB, Lane CG, Epstein JI (2001a) Interobserver reproducibility of Gleason grading of prostatic carcinoma: general pathologist. *Hum Pathol* 32(1):81–88
- 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 (2001b) Interobserver reproducibility of Gleason grading of prostatic carcinoma: urologic pathologists. *Hum Pathol* 32(1):74–80
- Altay B, Kefi A, Nazli O, Killi R, Semerci B, Akar I (2001) Comparison of Gleason scores from sextant prostate biopsies and radical prostatectomy specimens. *Urol Int* 67:14–18
- Bova GS, Parmigiani G, Epstein JI, Wheeler T, Mucci NR, Rubin MA (2001) Web-based tissue microarray image data analysis: initial validation testing through prostate cancer Gleason grading. *Hum Pathol* 32:417–427
- De la Taille A, Viellefond A, Berger N, Boucher E, de Fromont M, Fondimare A, Molinie V, Piron D, Sibony M, Staroz F, Triller M, Peltier E, Thiounn N, Rubin MA (2003) Evaluation of the interobserver reproducibility of Gleason grading of prostatic adenocarcinoma using tissue microarrays. *Hum Pathol* 34(5):444–449
- De las Morenas A, Siroky MB, Merriam J, Stilmant MM (1988) Prostatic adenocarcinoma: reproducibility and correlation with clinical stages of four grading systems. *Hum Pathol* 19(5):595–597
- Di Loreto C, Fitzpatrick B, Underhill S, Kim DH, Dytch HE, Galera-Davidson H, Bibbo M (1991) Correlation between visual clues, objective architectural features, and interobserver agreement in prostate cancer. *Am J Clin Pathol* 96(1):70–75
- Egevad L (2001) Reproducibility of Gleason grading of prostate cancer can be improved by the use of reference images. *Urology* 57(2):291–295
- Epstein JI, Partin AW, Sauvageot J, Walsh PC (1996) Predication of progression following radical prostatectomy: a multivariate analysis of 721 men with long-term follow-up. *Am J Surg Pathol* 20:286–292
- Egevad L, Norlen BJ, Norberg M (2001) The value of multiple core biopsies for predicting the Gleason score of prostate cancer. *BJU Int* 88:716–721
- Epstein JI, Pizov G, Walsh PC (1993) Correlation of pathologic findings with progression after radical retropubic prostatectomy. *Cancer* 71:3582–3593
- Gleason DF (1992) Histologic grading of prostate cancer: a perspective. *Hum Pathol* 23:273–279
- Gregori A, Vieweg J, Dahm P, Paulson DF (2001) Comparison of ultrasound-guided biopsies and prostatectomy specimens: predictive accuracy of Gleason score and tumor site. *Urol Int* 66:66–71

- King CR (2000) Patterns of prostate cancer biopsy grading: trends and clinical implications. *Int J Cancer* 90:305–311
- Kononen J, Bubendorf L, Kallioniemi A, Barlund M, Schraml P, Leighton S, Torhorst J, Mihatsch MJ, Sauter G, Kallioniemi OP (1998) Tissue microarray for high throughput molecular profiling of tumor specimens. *Nat Med* 4:844–847
- Kronz JD, Silberman MA, Allsbrook WC (2000) A web-based tutorial improves practicing pathologists' Gleason grading of images of prostate carcinoma specimens obtained by needle biopsy: validation of a new medical education paradigm. *Cancer* 89:1818–1823
- McLean M, Srigley J, Banerjee D, Warde P, Hao Y (1997) Interobserver variation in prostate cancer Gleason scoring: are there implications for the design of clinical trials and treatment strategies? *Clin Oncol (R Coll Radiol)* 9(4):222–225
- Mikami Y, Manabe T, Epstein JI, Shiraishi T, Furusato M, Tsuzuki T, Matsuno Y, Sasano H (2003) Accuracy of Gleason grading by practicing pathologists and the impact of education on improving agreement. *Hum Pathol* 34(7):658–665
- Oesterling JE, Brendler CB, Epstein JI, Kimball AW Jr, Walsh PC (1987) Correlation of clinical stage, serum prostatic acid phosphatase and preoperative Gleason grade with final pathological stage in 275 patients with clinically localized adenocarcinoma of the prostate. *J Urol* 38(1):92–98
- Ozdamar SO, Sarikaya S, Yildiz L, Atilla MK, Kandemir B, Yildiz S (1996) Intraobserver and interobserver reproducibility of WHO and Gleason histologic grading systems in prostatic adenocarcinomas. *Int Urol Nephrol* 28(1):73–77
- Partin AW, Mangold LA, Lamm DM, Walsh PC, Epstein JI, Pearson JD (2001) Contemporary update of prostate cancer staging nomograms (Partin Tables) for the new millennium. *Urology* 58(6):843–848
- Rubin MA, Dunn R, Strawderman M, Pienta KJ (2002) Tissue microarray sampling strategy for prostate cancer biomarker analysis. *Am J Surg Pathol* 26(3):312–319
- Smith EB, Frierson HF Jr, Mills SE, Boyd JC, Theodorescu D (2002) Gleason scores of prostate biopsy and radical prostatectomy specimens over the past 10 years: is there evidence for systematic upgrading? *Cancer* 94:2282–2287
- Svanholm H, Mygind H (1985) Prostatic carcinoma reproducibility of histologic grading. *Acta Pathol Microbiol Immunol Scand [A]* 93(2):67–71