Published in final edited form as:
Prostate. 2012 September 15; 72(13): 1389–1398. doi:10.1002/pros.22484.

Frequency and determinants of disagreement and error in Gleason scores: a population-based study of prostate cancer

Michael Goodman, MD, MPH[1,2,*], **Kevin C. Ward, PhD, MPH**[1,2], **Adeboye O. Osunkoya, MD**[3,4], **Milton W. Datta, MD**[5], **Daniel Luthringer, MD**[6], **Andrew N. Young, MD, PhD**[7], **Katerina Marks**[1], **Vaunita Cohen**[1], **Jan C. Kennedy, MD**[8], **Michael J. Haber, PhD**[1], and **Mahul B. Amin, MD**[6]

[1]Emory University Rollins School of Public Health, Atlanta, GA

[2]Georgia Center for Cancer Statistics, Atlanta, GA

[3]Emory University School of Medicine, Atlanta, GA

[4]Veterans Affairs Medical Center, Atlanta, GA

[5]University of Minnesota Medical School, Minneapolis, MN

[6]Cedars-Sinai Medical Center, Los Angeles, CA

[7]Grady Memorial Hospital

[8]Dekalb Medical Center, Decatur, GA

Abstract

Background—To examine factors that affect accuracy and reliability of prostate cancer grade we compared Gleason scores documented in pathology reports and those assigned by urologic pathologists in a population-based study.

Methods—A stratified random sample of 318 prostate cancer cases diagnosed was selected to ensure representation of whites and African-Americans and to include facilities of various types. The slides borrowed from reporting facilities were scanned and the resulting digital images were re-reviewed by two urologic pathologists. If the two urologic pathologists disagreed, a third urologic pathologist was asked to help arrive at a final "gold standard" result. The agreements between reviewers and between the pathology reports and the "gold standard" were examined by calculating kappa statistics. The determinants of discordance in Gleason scores were evaluated using multivariate models with results expressed as odds ratios (OR) and 95% confidence intervals (CI).

Results—The kappa values (95% CI) reflecting agreement between the pathology reports and the "gold standard," were 0.61 (95% CI: 0.54, 0.68) for biopsies, and 0.37 (0.23, 0.51) for prostatectomies. Sixty three percent of discordant biopsies and 72% of discordant prostatectomies showed only minimal differences. Using free standing laboratories as reference, the likelihood of discordance between pathology reports and expert-assigned biopsy Gleason scores was particularly elevated for small community hospitals (OR=2.98; 95% CI: 1.73, 5.14).

[*]Correspondence to: Michael Goodman MD, MPH, Department of Epidemiology, Emory University Rollins School of Public Health, 1518 Clifton Road, NE, Atlanta GA 30322, mgoodm2@sph.emory.edu.

Conflict of Interest Disclosures
The authors had no conflicts of interest.

Conclusions—The level of agreement between pathology reports and expert review depends on the type of diagnosing facility, but may also depend on the level of expertise and specialization of individual pathologists.

Keywords

prostate cancer; Gleason score; agreement; accuracy

BACKGROUND

The Gleason grade of prostate cancer is a predictor of tumor aggressiveness that plays an important role in determining patient treatment and prognosis (1). The Gleason grading system was developed in the 1960s based on histopathological data from multiple prostate-cancer biopsies and resections (2). Based on the architectural features of the cancer cells, five histologic patterns of decreasing differentiation were developed with a pattern of 1 representing the most differentiation and 5 representing the least differentiation (3). Because prostate adenocarcinoma is a multifocal disease with substantial histologic variability across foci, the most prevalent and second most prevalent patterns were added to obtain the Gleason score (range = 2 to 10). This score is often referred to as the combined Gleason grade. When a tumor included only one pattern, that pattern is counted twice. Thus, for a tumor with a single pattern of 3, the Gleason score is written as 3 + 3 = 6 (4). As indicated in the recent recommendations issued by the College of American Pathologists (5), in needle biopsy specimens where more than 2 patterns are present, and the highest pattern is neither the predominant nor the secondary pattern, the predominant and highest pattern should be chosen to arrive at a score (e.g., 75%, pattern 3; 20%–25%, pattern 4; <5%, pattern 5 is scored as 3 + 5 = 8).

As there is often no clear-cut distinction between Gleason patterns, the grading can be subjective and may depend on the training and experience of the pathologist evaluating the specimen (6,7). For these reasons, several previous studies have sought to examine and quantify the inter-observer agreement among pathologists (8–19). It is important to point out, however, that previous studies usually conducted their assessments of inter-observer agreement in controlled experimental conditions using pre-selected slides on a relatively limited number of cases. By contrast, data on the accuracy and reliability of Gleason scores in the general patient population are lacking.

Additional motivation to evaluate accuracy and reliability of Gleason score was provided by the changes in Gleason grading methods that were codified in the 2005 publication of the International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma (20). The publication of this consensus document completed several years of concerted efforts by leading urologic pathology experts to update practicing general pathologists on the new style of Gleason grading (4,9,21).

In view of these data gaps, the present study seeks to compare Gleason patterns and scores documented in the pathology reports and those assigned by expert review using a large sample of prostate cancer cases reported to the Metro Atlanta and Rural Georgia Surveillance Epidemiology and End Results (SEER) registry between 2004 and 2005. We sought to achieve this goal by answering the following specific research questions. 1) What are the frequency and the extent of disagreement between Gleason scores reported at diagnosis and those assigned by the expert review? 2) Is there disagreement between experts re-evaluating the original diagnostic slides? 3) Does disagreement between diagnostic and expert-assigned Gleason scores differ by type of specimen? 4) Using expert review as a gold standard, what patient-, disease-, and facility-related characteristics are associated with

misclassification of the Gleason scores documented in the pathology reports? 5) What is the accuracy of Gleason scores recorded in the SEER database?

MATERIALS AND METHODS

The eligibility criteria included: ICD-O code C61.9 (prostate), residence in the metropolitan Atlanta or rural Georgia SEER 15-county catchment area, diagnosis from January 1, 2004 through December 31, 2005; invasive disease (behavior code 3), and availability of both the slides and the pathology report for review. A stratified random sample of 325 eligible patients was selected for scanning and review of pathology slides with oversampling to allow for missing data and/or specimens and to arrive at a final sample size of about 300. The purpose of stratification was to include equal numbers of whites and African Americans. We also oversampled cases from smaller facilities such as rural community hospitals to ensure their representation in the study.

Facilities that established the diagnosis for the study cases were contacted and asked to provide biopsy and prostatectomy slides along with the corresponding pathology reports. Each slide was de-identified and scanned using the ScanScope digital microscopy by Aperio®.

Two urologic pathologists performed the re-review of the biopsies and prostatectomy specimens and assigned their own Gleason scores using the 2005 ISUP consensus recommendations (20). In assigning Gleason scores the urologic pathologists used images for the index (largest) tumor and not the entire case. As one of the study objectives was to assess the accuracy of Gleason scores reported to SEER, the expert review did not assess the tertiary patterns because cancer registries only collect information on primary and secondary tumors.

The use of digital microscopy enabled the two pathologists to read the images and assign Gleason scores from their offices. It is important to emphasize that previous studies have demonstrated excellent agreement between prostate pathology reviews that use glass slides and those that use digital images.(6,22) When the two pathologists assigned exactly the same primary and secondary Gleason pattern the results was considered the “gold standard.” If, however, the two urologic pathologists disagreed on either pattern, a third urologic pathologist (also a person with extensive post-training experience) was asked to review each case in question, and assign his own Gleason score (without knowing the other two reviewers’ scores), which then served as a “tie-breaker.” In a few instances when all three urologic pathologists disagreed, the third reviewer was unmasked with respect to the other two reviewers’ scores, and was asked to reconcile the three opinions to arrive at a final “gold standard” combined Gleason grade.

All pathology reports were abstracted using a data collection instrument developed for the purposes of this study. As a result, each biopsy or prostatectomy (sometimes pertaining to the same patient) had four Gleason score evaluations: one documented in the original pathology report, two assigned by the experts, and the final “gold standard” score using, if necessary, the third reviewer. Additional information on each case included demographic variables such as race, age, and residence; disease characteristics such as stage and prostate specific antigen (PSA) level; facility-related data such as type, size and university affiliation; and sample type (prostatectomy or biopsy).

The Gleason scores reported at diagnosis were compared to those assigned by the final expert review and the agreement between two sets of results for each case was examined by performing weighted kappa calculations where a kappa of 0.0 means that the agreement is no better than that expected by chance alone, and kappa values of 1.00 and −1.00 indicate

perfect agreement or perfect disagreement, respectively. By convention, a kappa between 0.81 and 1.00 is interpreted as indicating excellent agreement. Values of <0.20, 0.21–0.40, 0.41–0.60, and 0.61–0.80 are interpreted as showing poor, fair, moderate, and good agreement, respectively.(23) We also used the same methods to evaluate the agreement between the two urologic pathologists and between biopsy- and prostatectomy-derived results for those patients that had both sets of slides and pathology reports. Each kappa statistic was reported along with the corresponding 95% confidence interval (CI). The agreement was examined for three types of measures: two Gleason's patterns, total Gleason score and a dichotomous outcome of Gleason score ≥8 versus ≤7. The Gleason's patterns were ranked in ascending order based on a total score followed by the primary pattern. For example, 3+4=7 was followed by 4+3=7, followed by 3+5=8, followed by 4+4=8, followed by 5+3=8, and so on.

The associations between misclassification of Gleason scores and various patient-, disease-, and facility-related characteristics were examined using multivariate generalized estimating equation (GEE) models for binary data with a logit link function.(24) The GEE modeling approach allowed us to obtain adjusted odds ratios (OR) while accounting for clustering of observations (because of the same reporting characteristics) within each facility. The independent variables for each model included facility type, patient's age, race and area-based measure of socioeconomic status (SES), and disease stage and serum PSA level as recorded in the SEER data. The facilities that submitted slides and pathology reports represented four categories: freestanding laboratories, university-affiliated hospitals, large community hospitals and small community hospitals. Age was dichotomized: ≥65 years old versus ≤64 years old. All patient addresses were geocoded to the level of census tract and then assigned an area-based measure of SES dichotomized as high versus low based on the percent of individuals in the census tract living below the poverty level.(25) High SES area was defined as including less than 10% of population below poverty level. Stage was categorized as localized versus regional/distant, and serum PSA level was dichotomized as <10 ng/ml versus ≥10 ng/ml. The analyses were carried out using SPSS statistical software (LEAD Technologies, Inc., Chicago, IL), and (for kappa calculations) Computer Programs for Epidemiologic Analyses v. 4.0 (Authors: Abramson, JH and Gahlinger, PM; available at http://sagebrushpress.com)

RESULTS

A total of 1905 slides, pertaining to 268 biopsies and 120 prostatectomies obtained from 318 patients, were retrieved and scanned. For four biopsies the corresponding pathology reports was were not found. As shown in Table 1, 50% of all men that provided biopsy specimens were over the age of 65 years, while prostatectomy specimens tended to come from younger patients. Because our study sample was stratified on race, the numbers of specimens from white and African American patients were roughly equal. A greater proportion of all patients in the study resided in higher SES census tracts (less than 10% of the population living below poverty). Over 80% of all cases were diagnosed with localized disease, and more than half had a PSA of less than 10 ng/ml. With respect to the original diagnosis facility, approximately 35% of biopsy specimens were from freestanding laboratories, 27% from university hospitals, 21% from large community hospitals, and 17% from small community hospitals. University hospitals, large community hospitals, and small community hospitals provided 33%, 49% and 18% of prostatectomy specimens, respectively.

The frequency and magnitude of disagreement between two urologic pathologists are summarized in Table 2. Complete agreement was found in 144 (54%) of 268 biopsies and 76 (63%) of 120 prostatectomies. Most of the disagreements were by one point (category) with only 12% of all biopsy scores and 10% of all prostatectomy scores showing a

discrepancy of two categories or more. Among biopsy scores, the kappa values (95% CIs) were 0.59 (0.52, 0.66) for both Gleason patterns, 0.56 (0.48, 0.63) for the total Gleason score and 0.57 (0.47, 0.66) for the Gleason score of ≤7 versus ≥8. The corresponding results among prostatectomy grades were 0.53 (0.39, 0.67), 0.57 (0.44, 0.70) and 0.61 (0.47, 0.75) for Gleason patterns, Gleason score and Gleason score of ≤7 versus ≥8, respectively (Table 2). The tie-breaker reviewer agreed more often (in almost 60% of cases) with Reviewer I (an experienced urologic pathologist) than with Reviewer II (urologic pathology fellow)

When the results of the original pathology reports were compared to the “gold standard” (final expert review), the kappa statistics reflecting agreement in the two Gleason patterns were 0.61 (95% CI: 0.54, 0.68) for biopsies, and 0.37 (0.23, 0.51) for prostatectomies (Table 3). The corresponding biopsy- and prostatectomy-derived kappa values were 0.60 (0.53, 0.67) and 0.38 (0.26, 0.51) for the total Gleason score, and 0.58 (0.48, 0.67) and 0.43 (0.28, 0.58) for the Gleason score dichotomized at 8. Compared to the “gold standard” the pathology reports tended to over-estimate the Gleason grade. Among biopsy results, complete agreement was seen in 56% of cases, under-grading occurred in 31% of cases and over-grading occurred in 13% of cases. The percentages for complete agreement, under-, and over-grading among prostatectomy specimens were 52%, 36%, and 12%, respectively. Among those specimens that showed any disagreement, 63% of biopsies and 72% of prostatectomies were discordant by only one category.

When we examined the agreement between biopsy and prostatectomy results, we found that pathology reports had kappa values ranging between 0.35 and 0.53, whereas the kappa statistics for urologic pathology reviews ranged from 0.35 to 0.47 (Table 4). Complete agreement between biopsy and prostatectomy results was observed in 66% of pathology reports and 51% of expert assigned reviews. Among pathology reports the over- and under-estimated biopsy Gleason scores were found in 15% and 19% of cases, respectively. Similarly, the proportions of over- and under-estimated results among expert-assigned biopsy scores were 20% and 29%, respectively.

Multivariable GEE analyses that examined the association between Gleason pattern misclassification (comparing pathology reports to “gold standard”) and facility-, disease- or patient-related characteristics are presented in Tables 5 and 6. The results are shown separately for any disagreement in Gleason patterns and for disagreement by two or more categories (using the ranking of Gleason patterns as presented in Table 3). The probability of misclassification of the biopsy grade was higher in older patients with an OR of 2.06 (95% CI: 1.45, 2.92) for any discordance and 2.84 (95% CI 1.27, 6.35) for discordance of at least categories. Patients with more regional or distant disease were significantly more likely to have misclassified Gleason pattern compared to those with localized tumors with ORs of 2.78 (95% CI: 1.54, 5.02) for any misclassification, and 3.92 (95% CI: 1.47, 10.41) for misclassification by two or more categories. Using freestanding laboratories as the reference category, the ORs for any discordance between pathology reports and expert-assigned Gleason patterns were significantly elevated for university hospitals (OR=1.70; 95% CI 1.16, 2.48), large community hospitals (OR=2.08; 95% CI: 1.37, 3.16) and particularly small community hospitals (OR=2.98; 95% CI: 1.73, 5.14). The results for discordance by more than one category were attenuated and not statistically significant. Area-based SES, race and PSA level did not demonstrate a significant association with disagreement in the biopsy-derived Gleason grades (Table 5).

The corresponding GEE models using misclassification in the prostatectomy-derived Gleason score as the dependent variable demonstrated no association with age, race, SES, or disease stage. Higher PSA levels were related to higher probability of disagreement between pathology report and expert review, but the difference was statistically significant only when

the outcome was defined as “any discordance.” In the analyses for disagreement by two or more categories the OR was significantly elevated for small community hospitals (OR=3.61; 95% CI: 1.14, 11.47) compared to university hospitals (Table 6).

DISCUSSION

The data presented here indicate that most Gleason grades fall into a fairly narrow range (total scores of 6 and 7). A substantially smaller proportion of the specimens had a total Gleason score of 8 or higher. Only one biopsy and two prostatectomies were accompanied by pathology reports indicating a Gleason score of less than 6, and none of those was judged to be below 6 according to the “gold standard’ review.

The inter-observer kappa estimates in this study were consistent with those reported for urologic pathologists, and higher than those found among general pathologists in other studies.(8–19) We found only weak-to-moderate agreement between biopsy-and prostatectomy derived Gleason scores; an observation that is also in concurrence with earlier studies.(26–34) As expected based on previous reviews,(35,36) examination of biopsies in our study tended to under-estimate the prostate cancer grade; although the extent of underestimation was greater in the expert reviews compared to pathology reports. The finding that disagreement between biopsy- and prostatectomy-derived Gleason scores was more pronounced in the expert assessment compared to the original pathology reports is not surprising. The pathologists performing the original prostatectomy evaluations were most likely aware of the biopsy-derived Gleason scores, whereas the expert review was carried out blindly. The disagreement between original pathology and expert “gold standard review” was also modest.

One of the stated objectives of this study was to examine the patient-, disease- and facility-related characteristics that may predict misclassification of the Gleason score at diagnosis. In the analyses of biopsy specimens we found that the most important patient- and disease-related predictors of discordance between the original pathology report and the expert re-review results were patient age and disease stage. While older patients (≥65 years of age) in general were more likely to have discordant Gleason scores, among misclassified biopsies, the proportion of under-graded specimens was higher among younger (78%) than among (64%) older men. The difference in the frequency of under-graded biopsy specimens between the localized and advanced prostate cancer cases (66% versus 70%) was less pronounced. Another important finding that warrants further discussion is the better performance (at least with respect to agreement with expert review) of freestanding laboratories. One reason for this observation may be the use of more than one signing pathologist in each case at freestanding laboratories. In these situations the final diagnostic report reflects the diagnostic consensus of two pathologists, and may be more likely to eliminate any outlier opinions. It is also possible that freestanding laboratories are staffed with pathologists that are more likely to be specialized in a particular area.

A noteworthy feature of this study is that it examined prostate cancer cases that were diagnosed right around the time the ISUP was making its push for updating the Gleason scoring recommendations. It is important to emphasize that although ISUP published the consensus report in 2005, efforts to introduce the new recommendations into routine pathology practice had been on-going for a number of years. Urologic pathology experts (including two co-authors of the current paper) were part of a rather widespread campaign that used local, national and international conferences to update practicing general pathologists on the new style of Gleason grading. The success of these efforts is evidenced in studies reporting that concordance between practicing pathologists and experts improved over time (37,38). It is reasonable to assume that the new recommendations were adopted by

practicing pathologists at different times and varied by facility. This variable degree of awareness of the new recommendations is another likely explanation for the observation that the agreement between experts and diagnosing pathologists differed greatly by type of facility.

In contrast to our biopsy analyses, there were few significant determinants of discordance in Gleason scores assigned to prostatectomy specimens. It does appear that small community hospitals are more likely to misclassify the Gleason score by two or more categories, but this finding was not consistent across the analyses. There appears to be a consensus that prostatectomy-derived grade is a more accurate predictor of disease recurrence and prognosis than the grade assigned during the biopsy evaluation.(36,39) On the other hand, from the patient care point of view the grade of a biopsy specimen is more important because it serves as the basis for critical initial prostate cancer treatment decisions.(40)

There was little evidence that race or SES was associated with a discernable increase or decrease in the agreement between pathology reports and expert reviews. Although these results indicate lack of appreciable sociodemographic disparity, one needs to keep in mind the limitations of our analyses, such as reliance on area-based as opposed to individual measures of SES and lack of information regarding health insurance. In general, the interpretation of our results requires understanding of the strengths and limitations of the registry-based (in this case SEER-based) data. As previously noted elsewhere, the large sample size enables SEER-based studies to have sufficient power of detecting relatively moderate associations and permits a variety of stratified and multivariate analyses.(41) The population-based, as opposed to institution-based, selection of cases increases the external validity and generalizability of findings. While institutional studies often have more detailed information about each patient, those studies usually are confined to major referral centers and may not be representative of the cases treated in the community (42).

CONCLUSIONS

In summary, we found that a substantial proportion of biopsy and prostatectomy specimens have different Gleason scores assigned at diagnosis compared to those assigned by expert review; however the magnitude of disagreement is rather modest. The highest level of disagreement was present in small community hospitals. This finding requires confirmation and, if confirmed, further exploration. Our study also demonstrates the feasibility of linking registry data with digitized pathology slides, and perhaps other clinical images.

Acknowledgments

This study was funded by the grant N01-PC- 32187 (PI Michael Goodman) from the National Cancer Institute's SEER program

References

1. Eble, JN.; Sauter, G.; Epstein, JI.; Sesterhenn, IA. Pathology and Genetics of Tumours of the Urinary System and Male Genital Organs. Lyon: IARC Press; 2004. World Health Organization Classification of Tumours.
2. Gleason DF. Classification of prostatic carcinomas. Cancer Chemother Rep. 1966; 50(3):125–128. [PubMed: 5948714]
3. Harnden P, Shelley MD, Coles B, Staffurth J, Mason MD. Should the Gleason grading system for prostate cancer be modified to account for high-grade tertiary components? A systematic review and meta-analysis. Lancet Oncol. 2007; 8(5):411–419. [PubMed: 17466898]

4. Epstein JI, Potter SR. The pathological interpretation and significance of prostate needle biopsy findings: implications and current controversies. J Urol. 2001; 166(2):402–410. [PubMed: 11458037]
5. Srigley JR, Humphrey PA, Amin MB, Chang SS, Egevad L, Epstein JI, Grignon DJ, McKiernan JM, Montironi R, Renshaw AA, Reuter VE, Wheeler TM. Protocol for the examination of specimens from patients with carcinoma of the prostate gland. Arch Pathol Lab Med. 2009; 133(10):1568–1576. [PubMed: 19792046]
6. Helin H, Lundin M, Lundin J, Martikainen P, Tammela T, Helin H, van der Kwast T, Isola J. Web-based virtual microscopy in teaching and standardizing Gleason grading. Hum Pathol. 2005; 36(4): 381–386. [PubMed: 15891999]
7. Iczkowski KA, Bostwick DG. The pathologist as optimist: cancer grade deflation in prostatic needle biopsies. Am J Surg Pathol. 1998; 22(10):1169–1170. [PubMed: 9777978]
8. 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. Interobserver reproducibility of Gleason grading of prostatic carcinoma: urologic pathologists. Hum Pathol. 2001; 32(1):74–80. [PubMed: 11172298]
9. 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(1):81–88. [PubMed: 11172299]
10. Burchardt M, Engers R, Muller M, Burchardt T, Willers R, Epstein JI, Ackermann R, Gabbert HE, de la Taille A, Rubin MA. Interobserver reproducibility of Gleason grading: evaluation using prostate cancer tissue microarrays. J Cancer Res Clin Oncol. 2008; 134(10):1071–1078. [PubMed: 18392850]
11. Coard K, Freeman V. Gleason grading of prostate cancer: level of concordance between pathologists at the University Hospital of the West Indies. Am J Clin Pathol. 2004; 122(3):373–376. [PubMed: 15362366]
12. 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. Evaluation of the interobserver reproducibility of Gleason grading of prostatic adenocarcinoma using tissue microarrays. Hum Pathol. 2003; 34(5):444–449. [PubMed: 12792917]
13. Evans AJ, Henry PC, Van der Kwast TH, Tkachuk DC, Watson K, Lockwood GA, Fleshner NE, Cheung C, Belanger EC, Amin MB, Boccon-Gibod L, Bostwick DG, Egevad L, Epstein JI, Grignon DJ, Jones EC, Montironi R, Moussa M, Sweet JM, Trpkov K, Wheeler TM, Srigley JR. Interobserver variability between expert urologic pathologists for extraprostatic extension and surgical margin status in radical prostatectomy specimens. Am J Surg Pathol. 2008; 32(10):1503–1512. [PubMed: 18708939]
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(5):2006–2010. [PubMed: 12394696]
15. Griffiths DF, Melia J, McWilliam LJ, Ball RY, Grigor K, Harnden P, Jarmulowicz M, Montironi R, Moseley R, Waller M, Moss S, Parkinson MC. A study of Gleason score interpretation in different groups of UK pathologists; techniques for improving reproducibility. Histopathology. 2006; 48(6):655–662. [PubMed: 16681680]
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. Observer variability in the histopathological reporting of needle biopsy specimens of the prostate. Hum Pathol. 1997; 28(6): 646–649. [PubMed: 9190997]
17. McLean M, Srigley J, Banerjee D, Warde P, Hao Y. Interobserver variation in prostate cancer Gleason scoring: are there implications for the design of clinical trials and treatment strategies? Clin Oncol (R Coll Radiol). 1997; 9(4):222–225. [PubMed: 9315395]
18. Melia J, Moseley R, Ball RY, Griffiths DF, Grigor K, Harnden P, Jarmulowicz M, McWilliam LJ, Montironi R, Waller M, Moss S, Parkinson MC. A UK-based investigation of inter- and intra-observer reproducibility of Gleason grading of prostatic biopsies. Histopathology. 2006; 48(6): 644–654. [PubMed: 16681679]

19. Veloso SG, Lima MF, Salles PG, Berenstein CK, Scalon JD, Bambirra EA. Interobserver agreement of Gleason score and modified Gleason score in needle biopsy and in surgical specimen of prostate cancer. Int Braz J Urol. 2007; 33(5):639–646. discussion 647–651. [PubMed: 17980061]
20. Epstein JI, Allsbrook WC Jr, Amin MB, Egevad LL. The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma. Am J Surg Pathol. 2005; 29(9):1228–1242. [PubMed: 16096414]
21. Epstein JI. Gleason score 2–4 adenocarcinoma of the prostate on needle biopsy: a diagnosis that should not be made. Am J Surg Pathol. 2000; 24(4):477–478. [PubMed: 10757394]
22. Kronz JD, Silberman MA, Allsbrook WC, Epstein JI. 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. 2000; 89(8):1818–1823. [PubMed: 11042578]
23. Fleiss, J. Statistical methods for rates and proportions. New York, NY: John Wiley and Sims; 1981.
24. Kleinbaum, DG.; Klein, M. Logistic regression: A self-learning text. New York, NY: Springer-Verlag; 2002.
25. Krieger N, Chen JT, Waterman PD, Soobader MJ, Subramanian SV, Carson R. Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter?: the Public Health Disparities Geocoding Project. Am J Epidemiol. 2002; 156(5):471–482. [PubMed: 12196317]
26. Cam K, Yucel S, Turkeri L, Akdas A. Accuracy of transrectal ultrasound guided prostate biopsy: histopathological correlation to matched prostatectomy specimens. Int J Urol. 2002; 9(5):257–260. [PubMed: 12060438]
27. Danziger M, Shevchuk M, Antonescu C, Matthews G, Fracchia J. Predictive accuracy of transrectal ultrasound-guided prostate biopsy: correlations to matched prostatectomy specimens. Urology. 1997; 49(6):863–867. [PubMed: 9187692]
28. Divrik RT, Eroglu A, Sahin A, Zorlu F, Ozen H. Increasing the number of biopsies increases the concordance of Gleason scores of needle biopsies and prostatectomy specimens. Urol Oncol. 2007; 25(5):376–382. [PubMed: 17826653]
29. Emiliozzi P, Maymone S, Paterno A, Scarpone P, Amini M, Proietti G, Cordahi M, Pansadoro V. Increased accuracy of biopsy Gleason score obtained by extended needle biopsy. J Urol. 2004; 172(6 Pt 1):2224–2226. [PubMed: 15538236]
30. Fukagai T, Namiki T, Namiki H, Carlile RG, Shimada M, Yoshida H. Discrepancies between Gleason scores of needle biopsy and radical prostatectomy specimens. Pathol Int. 2001; 51(5): 364–370. [PubMed: 11422794]
31. Gregori A, Vieweg J, Dahm P, Paulson DF. Comparison of ultrasound-guided biopsies and prostatectomy specimens: predictive accuracy of Gleason score and tumor site. Urol Int. 2001; 66(2):66–71. [PubMed: 11223746]
32. Noguchi M, Stamey TA, McNeal JE, Yemoto CM. Relationship between systematic biopsies and histological features of 222 radical prostatectomy specimens: lack of prediction of tumor significance for men with nonpalpable prostate cancer. J Urol. 2001; 166(1):104–109. discussion 109–110. [PubMed: 11435833]
33. Shen BY, Tsui KH, Chang PL, Chuang CK, Hsieh ML, Huang ST, Wang TM, Lee SH, Huang HC, Huang SC. Correlation between the Gleason scores of needle biopsies and radical prostatectomy specimens. Chang Gung Med J. 2003; 26(12):919–924. [PubMed: 15008327]
34. Stav K, Judith S, Merald H, Leibovici D, Lindner A, Zisman A. Does prostate biopsy Gleason score accurately express the biologic features of prostate cancer? Urol Oncol. 2007; 25(5):383–386. [PubMed: 17826654]
35. Gleason DF. Histologic grading of prostate cancer: a perspective. Hum Pathol. 1992; 23(3):273–279. [PubMed: 1555838]
36. Humphrey PA. Gleason grading and prognostic factors in carcinoma of the prostate. Mod Pathol. 2004; 17(3):292–306. [PubMed: 14976540]

37. Mikami Y, Manabe T, Epstein JI, Shiraishi T, Furusato M, Tsuzuki T, Matsuno Y, Sasano H. Accuracy of Gleason grading by practicing pathologists and the impact of education on improving agreement. Hum Pathol. 2003; 34(7):658–665. [PubMed: 12874761]

38. Renshaw AA, Schultz D, Cote K, Loffredo M, Ziemba DE, D'Amico AV. Accurate Gleason grading of prostatic adenocarcinoma in prostate needle biopsies by general pathologists. Arch Pathol Lab Med. 2003; 127(8):1007–1008. [PubMed: 12873176]

39. Narain V, Bianco FJ Jr, Grignon DJ, Sakr WA, Pontes JE, Wood DP Jr. How accurately does prostate biopsy Gleason score predict pathologic findings and disease free survival? Prostate. 2001; 49(3):185–190. [PubMed: 11746263]

40. Gleason D. Undergrading of prostate cancer biopsies: a paradox inherent in all biologic bivariate distributions. Urology. 1996; 47(3):289–291. [PubMed: 8633390]

41. Ludwig MS, Goodman M, Miller DL, Johnstone PA. Postoperative survival and the number of lymph nodes sampled during resection of node-negative non-small cell lung cancer. Chest. 2005; 128(3):1545–1550. [PubMed: 16162756]

42. Esiashvili N, Goodman M, Marcus RB Jr. Changes in incidence and survival of Ewing sarcoma patients over the past 3 decades: Surveillance Epidemiology and End Results data. J Pediatr Hematol Oncol. 2008; 30(6):425–430. [PubMed: 18525458]

Table 1

Characteristics of the study population

Study Variables	All specimens (n=388)	Biopsies (n=268)		Prostatectomies (n=120)	
		Number	Percent	Number	Percent
Diagnosis age					
<65	232	134	50.0%	98	81.7%
≥65	156	134	50.0%	22	18.3%
Race					
White	190	130	48.5%	60	50.0%
Black	198	138	51.5%	60	50.0%
Area-based measure of SES *					
Low	166	120	44.8%	46	38.3%
High	222	148	55.2%	74	61.7%
Disease stage					
Localized	328	229	85.4%	99	82.5%
Regional/Distant	60	39	14.6%	21	17.5%
PSA at diagnosis					
<10 ng/ml	214	138	51.5%	76	63.4%
≥10 ng/ml	118	96	35.8%	22	18.3%
Not reported/missing	56	34	12.7%	22	18.3%
Facility type					
Freestanding laboratories	95	95	35.4%	0	0.0%
University hospitals	111	71	26.5%	40	33.3%
Large community hospitals	115	56	20.9%	59	49.2%
Small community hospitals	67	46	17.2%	21	17.5%

* Defined based on percent population in a census tract living below poverty level: low SES ≥ 10%, high SES <10%

Table 2

Agreement of Gleason patterns and scores assigned by two expert reviewers

A. BIOPSY SPECIMENS (n=268)

		Reviewer 2							
		3+3=6	**3+4=7**	**4+3=7**	**3+5=8**	**4+4=8**	**4+5=9**	**5+4=9**	**5+5=10**
Reviewer 1	3+3=6	**83**	14	4	0	0	0	0	0
	3+4=7	34	**37**	14	0	2	0	0	0
	4+3=7	3	12	**16**	0	8	0	0	0
	3+5=8	0	1	2	**0**	0	0	0	0
	4+4=8	1	1	3	0	**3**	1	0	0
	4+5=9	1	1	7	0	9	**3**	1	0
	5+4=9	0	0	0	0	1	3	**1**	0
	5+5=10	0	0	0	0	0	0	1	**1**

Weighted kappa for Gleason pattern: 0.59 (95% CI: 0.52, 0.66)

Weighted kappa for total Gleason score: 0.56 (95% CI: 0.48, 0.63)

Kappa for Gleason score ≤7 versus ≥8: 0.57 (95% CI: 0.47, 0.66)

B. PROSTATECTOMY SPECIMENS (n=120)

		Reviewer 2								
		3+2=5	**3+3=6**	**3+4=7**	**4+3=7**	**3+5=8**	**4+4=8**	**5+3=8**	**4=5=9**	**5+4=9**
Reviewer 1	3+2=5	**0**	1	0	0	0	0	0	0	0
	3+3=6	0	**41**	6	1	0	0	0	0	0
	3+4=7	0	14	**26**	9	0	2	0	0	0
	4+3=7	0	0	2	**6**	0	0	1	0	0
	3+5=8	0	1	2	0	**0**	0	0	0	1
	4+4=8	0	1	0	1	0	**0**	0	0	0
	5+3=8	0	0	0	1	0	0	**0**	0	0
	4+5=9	0	0	1	0	0	0	0	**3**	0

B. PROSTATECTOMY SPECIMENS (n=120)

					Reviewer 2				
	3+2=5	**3+3=6**	**3+4=7**	**4+3=7**	**3+5=8**	**4+4=8**	**5+3=8**	**4=5=9**	**5+4=9**
5+4=9	0	0	0	0	0	0	0	0	**0**

Weighted kappa for Gleason patterns: 0.53 (95% CI: 0.39, 0.67)

Weighted kappa for total Gleason score: 0.57 (95% CI: 0.44, 0.70)

Kappa for Gleason score ≤7 versus ≥8: 0.61 (95% CI: 0.47, 0.75)

Prostate. Author manuscript; available in PMC 2013 September 15.

Table 3

Agreement between Gleason patterns and scores assigned by final expert review and those documented in pathology reports

A. BIOPSY SPECIMENS (n=264)

Final Expert Review \ Pathology Report	1+2=3	3+3=6	3+4=7	4+3=7	3+5=8	4+4=8	5+3=8	4+5=9	5+4=9	5+5=10
1+2=3	**0**	0	0	0	0	0	0	0	0	0
3+3=6	0	**87**	5	0	0	0	0	0	0	0
3+4=7	0	43	**40**	8	1	1	0	0	1	0
4+3=7	1	7	14	**4**	0	8	0	1	1	0
3+5=8	0	0	0	0	**0**	0	0	1	0	0
4+4=8	0	1	4	2	0	**9**	0	3	0	0
5+3=8	0	0	0	0	0	0	**0**	0	0	0
4+5=9	0	0	2	1	0	5	1	**8**	1	3
5+4=9	0	0	0	0	0	0	0	0	**0**	1
5+5=10	0	0	0	0	0	0	0	0	0	**0**

Weighted kappa for Gleason patterns: 0.61 (95% CI: 0.54, 0.68)

Weighted kappa for total Gleason score: 0.60 (95% CI: 0.53, 0.67)

Kappa for Gleason score ≤7 versus ≥8: 0.58 (95% CI: 0.48, 0.67)

B. PROSTATECTOMY SPECIMENS (n=120)

Final Expert Review \ Pathology Report	3+2=5	2+4=6	3+3=6	3+4=7	4+3=7	3+5=8	4+4=8	4+5=9	5+4=9	5+5=10
3+2=5	**0**	0	0	0	0	0	0	0	0	0
2+4=6	0	**0**	0	0	0	0	0	0	0	0
3+3=6	1	0	**36**	6	0	0	0	1	0	0
3+4=7	1	1	24	**24**	0	2	1	1	0	1
4+3=7	0	0	2	9	**1**	0	1	1	0	0
3+5=8	0	0	0	2	0	**0**	0	0	0	0

B. PROSTATECTOMY SPECIMENS (n=120)

	Pathology Report									
	3+2=5	**2+4=6**	**3+3=6**	**3+4=7**	**4+3=7**	**3+5=8**	**4+4=8**	**4+5=9**	**5+4=9**	**5+5=10**
4+4=8	0	0	0	0	0	0	**1**	0	0	0
4+5=9	0	0	0	1	0	0	2	**0**	1	0
5+4=9	0	0	0	0	0	0	0	0	**0**	0
5+5=10	0	0	0	0	0	0	0	0	0	**0**

Weighted kappa for Gleason patterns: 0.37 (95% CI: 0.23, 0.51)

Weighted kappa for total Gleason score: 0.38 (95% CI: 0.26, 0.51)

Kappa for Gleason score ≤ 7 versus ≥8: 0.43 (95% CI: 0.28, 0.58)

Table 4

Agreement between biopsy and prostatectomy Gleason patterns and scores

A. PATHOLOGY REPORTS (n=67)

		Prostatectomy specimens								
		3+2=5	**2+4=6**	**3+3=6**	**3+4=7**	**4+3=7**	**3+5=8**	**4+4=8**	**4+5=9**	**5+4=9**
Biopsy specimens	3+2=5	**0**	0	1	0	0	0	0	0	0
	2+4=6	0	**0**	1	0	0	0	0	0	0
	3+3=6	0	0	**27**	4	0	0	0	0	0
	3+4=7	0	0	8	**16**	2	1	0	1	0
	4+3=7	0	0	0	0	**0**	0	0	0	0
	3+5=8	0	0	1	0	0	**0**	0	0	0
	4+4=8	0	0	0	1	0	0	**1**	0	0
	4+5=9	0	0	0	0	2	0	0	**0**	0
	5+4=9	0	0	0	0	0	0	0	1	**0**

Weighted kappa for Gleason patterns: 0.50 (95% CI: 0.33, 0.67)

Weighted kappa for total Gleason score: 0.53 (95% CI: 0.35, 0.70)

Kappa for Gleason score ≤7 versus ≥8: 0.35 (95% CI: −0.05, 0.88)

B. FINAL EXPERT REVIEW (n=70)

		Prostatectomy specimens					
		3+3=6	**3+4=7**	**4+3=7**	**3+5=8**	**4+4=8**	**4+5=9**
Biopsy specimens	3+3=6	**15**	7	2	0	0	0
	3+4=7	13	**18**	3	0	0	1
	4+3=7	0	4	**2**	0	0	0
	3+5=8	0	1	0	**0**	0	1
	4+4=8	0	0	1	0	**0**	0
	4+5=9	0	1	0	0	0	**1**

Weighted kappa for Gleason patterns: 0.35 (0.16–0.55)

Weighted kappa for total Gleason score: 0.36 (95% CI: 0.15, 0.56)

Prostate. Author manuscript; available in PMC 2013 September 15.

B. FINAL EXPERT REVIEW (n=70)

	Prostatectomy specimens					
	3+3=6	3+4=7	4+3=7	3+5=8	4+4=8	4+5=9

Kappa for Gleason score ≤7 versus ≥8: 0.47 (95% CI: 0.03, 0.91)

Table 5

Multivariate analyses evaluating determinants of discordance between biopsy-derived Gleason patterns assigned by final (gold standard") expert review and those documented in pathology reports

Determinants of disagreement	Any discordance	Discordance by two or more categories*
Diagnosis age:	OR (95% CI)	OR (95% CI)
<65 years	1.0 (reference)	1.0 (reference)
≥65 years	2.06 (1.45, 2.92)	2.84 (1.27, 6.35)
Race:		
White	1.0 (reference)	1.0 (reference)
Black	1.23 (0.59, 2.55)	0.89 (0.36, 2.25)
Area, based measure of SES**		
Low	1.0 (reference)	1.0 (reference)
High	1.15 (0.83, 1.60)	0.75 (0.38, 1.45)
Disease stage		
Localized	1.0 (reference)	1.0 (reference)
Regional/Distant	2.78 (1.54, 5.02)	3.92 (1.47, 10.41)
PSA at diagnosis		
<10 ng/ml	1.0 (reference)	1.0 (reference)
≥10 ng/ml	1.37 (0.67, 2.82)	2.41 (0.77, 7.53)
Not reported/missing	1.69 (0.78, 3.65)	2.81 (1.00, 7.95)
Facility type:		
Freestanding laboratories	1.0 (reference)	1.0 (reference)
University hospitals	1.70 (1.16, 2.48)	1.17 (0.62, 2.23)
Large community hospitals	2.08 (1.37, 3.16)	1.63 (0.77, 3.44)
Small community hospitals	2.98 (1.73, 5.14)	2.54 (0.92, 7.02)

*Gleason's patterns were ranked in ascending order based on a total score followed by a primary pattern as shown in Tables 2–4. For example, 3+4=7 was followed by 4+3=7, followed by 3+5=8, followed by 4+4=8, tec.

**Defined based on percent population in a census tract living below poverty level: low SES ≥10%, high SES <10%

Table 6

Multivariate analyses evaluating determinants of discordance between prostatectomy-derived Gleason patterns assigned by final (gold standard") expert review and those documented in pathology reports

Determinants of disagreement	Any discordance	Discordance by two or more categories*
Diagnosis age	OR (95% CI)	OR (95% CI)
<65 years	1.0 (reference)	1.0 (reference)
≥65 years	1.05 (0.38, 2.92)	1.38 (0.36, 5.27)
Race		
White	1.0 (reference)	1.0 (reference)
Black	1.67 (0.78, 3.53)	2.18 (0.73, 6.52)
Area, based measure of SES**		
Low	1.0 (reference)	1.0 (reference)
High	1.39 (0.57, 3.36)	0.84 (0.33, 2.15)
Disease stage		
Localized	1.0 (reference)	1.0 (reference)
Regional/Distant	1.26 (0.56, 2.87)	1.16 (0.33, 4.05)
PSA at diagnosis		
<10 ng/ml	1.0 (reference)	1.0 (reference)
≥10 ng/ml	2.63 (1.19, 5.84)	2.51 (0.68, 9.31)
Not reported/missing	0.73 (0.32, 1.68)	1.40 (0.32, 6.22)
Facility type		
University hospitals	1.0 (reference)	1.0 (reference)
Large community hospitals	1.41 (0.59, 3.38)	1.31 (0.33, 5.18)
Small community hospitals	1.14 (0.55, 2.35)	3.61 (1.14, 11.47)

*Gleason's patterns were ranked in ascending order based on a total score followed by a primary score as shown in Tables 2–4. For example, 3+4=7 was followed by 4+3=7, followed by 3+5=8, followed by 4+4=8, etc.

**Defined based on percent population in a census tract living below poverty level: low SES ≥10%, high SES <10%