Factors Predicting Prostatic Biopsy Gleason Sum Under Grading

Danielle A. Stackhouse, Leon Sun, Florian R. Schroек, Jayakrishnan Jayachandran, Arthur A. Caire, Cyril O. Acholo, Cary N. Robertson, David M. Albala,* Thomas J. Polascik, Craig F. Donatucci, Kelly E. Maloney and Judd W. Moul†‡

From the Division of Urologic Surgery and Duke Prostate Center; Department of Surgery, Duke University Medical Center, Durham, North Carolina

Abbreviations

PSA = prostate specific antigen
TRUS = transrectal ultrasound

Purpose: We determined clinical factors affecting the under grading of biopsy Gleason sum compared with prostatectomy pathology and developed a model predicting the probability of under grading.

Materials and Methods: We analyzed a cohort of 1,701 patients treated for prostate cancer at our institution between 1988 and 2007 with complete biopsy and pathological data available. Patients with a biopsy Gleason sum of 7 or less were included in our analysis. Cases were categorized as under graded or not under graded by comparing biopsy and radical prostatectomy Gleason sums. Logistic regression was used to determine the predictors of under grading based on clinical variables (race, age at diagnosis, body mass index, prostate weight, diagnostic prostate specific antigen, biopsy positive-to-total core ratio, maximal cancer percent in positive cores and time from diagnosis to surgery). A nomogram was developed to calculate the probability of under grading. Results were validated using bootstrapping.

Results: Under grading occurred in 46.6% of our cohort. Significant variables predicting under grading were age at diagnosis, biopsy Gleason sum, diagnostic prostate specific antigen, prostate weight, biopsy positive-to-total core ratio and maximal percent of cancer in cores (p < 0.05). Nomogram predictive accuracy was 72.4%.

Conclusions: The risk of Gleason sum under grading can be predicted to a satisfactory level using our nomogram. Predicting under grading would improve patient consulting and identify those who should consider repeat biopsy, ultimately enhancing the accuracy of prostate cancer diagnosis.

Key Words: prostate, prostatic neoplasms, biopsy, diagnosis, nomograms

Histological grading of prostate biopsy specimens is a main determinant of prostate cancer treatment.1,2 Several retrospective studies demonstrated that concordance rates between biopsy and pathological Gleason sums are inadequate.3,4 The most common grading error is under grading, identified in 38% to 72% of reported case studies.4–6 The potential impact that even a 1 grade error in Gleason sum can have on therapeutic decisions is substantial. Due to the significant risk of under grading it is evident that biopsy Gleason sum alone is not optimal for appropriate risk stratification, outcome prediction and therapeutic decision making in patients with prostate cancer.

Prior studies, including one of our studies, showed that advanced diagnostic age at diagnosis and high PSA are predictive of biopsy Gleason sum under grading.4,5 We identified addi-
tional preoperative clinical factors that could be included in the prediction model to more accurately identify which patients are at higher risk for Gleason sum under grading.

**MATERIALS AND METHODS**

**Study Population**

A total of 4,562 patients diagnosed and treated for prostate cancer at our institution between 1988 and 2008 comprised the study population. We analyzed a cohort of 1,701 patients who underwent radical prostatectomy during this period and had complete records available of biopsy and pathological Gleason sums. Patients without a recorded prostate weight, those younger than 45 years at diagnosis, 171 in whom disease was over graded and those with a Gleason sum of 8 or greater were excluded from analysis, leaving a cohort of 1,530 available for analysis.

Our intent was to analyze a cohort of patients whose potential treatment would be influenced by whether disease was under graded on biopsy. Biopsy was performed based on the usual indications for biopsy, including a palpable abnormality on digital rectal examination or increased PSA. We excluded patients younger than 45 years because we thought that they would receive treatment regardless of whether disease was under graded due to the expected life span. We did not include cases over graded on biopsy because our intent was to analyze the characteristics of cases that were under graded. We did not include patients who had a biopsy Gleason sum of 8 or greater because they were already in a high risk category and up grading patients would not change clinical management.

**Statistical Analysis**

Biopsies were categorized as under or not under graded by comparing biopsy and pathological Gleason sums. Under graded biopsies were defined as biopsies showing a Gleason sum lower than the Gleason sum in the prostatectomy specimens. Categorical variables included race (black or nonblack), body mass index (less than 25, 25 to 29.9 or 30 kg/m² or greater) and clinical Gleason sum (less than 7 or 7). Continuous variables were patient age at diagnosis, diagnostic PSA, prostate weight, positive-to-total core ratio on biopsy, maximal cancer percent in positive cores and time from diagnosis to surgery. Since PSA and prostate specimen weight did not have a normal distribution, they were entered into the model after logarithmic transformation. Prostate weight was considered based on the assumption that there is a good correlation between TRUS volume and prostate weight. Before the mid 1990s the standard practice at our institution was to perform sextant biopsies, at which point our physicians transitioned to the current standard of 12-core biopsies. We used the positive-to-total core ratio to consider changes from 6 to 12-core biopsies.

Stepwise regression analysis was done to identify significant predictors of under grading. Significant predictors in the multivariate model were used to construct a nomogram to predict the probability of down grading using the Design library functions of R (R Foundation for Statistical Computing, Vienna, Austria). The nomogram was validated and predictive accuracy was calculated using Harrell’s concordance index C on 200 bootstrapped re-samples to decrease overfit bias, which might inflate predictive accuracy. We also generated calibration plots to assess nomogram discrimination capacity. Institutional review board approval was arranged through the Duke Prostate Center.

**RESULTS**

Of the 1,530 cases included in the final analysis 713 (46.6%) were under graded on prostate biopsy. Table 1 lists the overall distribution of patients in

---

Table 1. Patient characteristics

<table>
<thead>
<tr>
<th></th>
<th>Not Under Graded</th>
<th>Under Graded</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median age at diagnosis (IQR)</td>
<td>61.4 (55.9–66.5)</td>
<td>62.2 (57.2–66.9)</td>
<td>0.001</td>
</tr>
<tr>
<td>No. race (%)</td>
<td>102 (12.5)</td>
<td>121 (17.1)</td>
<td>0.013</td>
</tr>
<tr>
<td>Black</td>
<td>714 (87.5)</td>
<td>591 (78.6)</td>
<td></td>
</tr>
<tr>
<td>Nonblack</td>
<td>27.5 (25.3–30.1)</td>
<td>27.6 (25.0–30.7)</td>
<td>0.066</td>
</tr>
<tr>
<td>15–24</td>
<td>138 (21.7)</td>
<td>136 (25.0)</td>
<td></td>
</tr>
<tr>
<td>25–30</td>
<td>339 (53.2)</td>
<td>252 (46.4)</td>
<td></td>
</tr>
<tr>
<td>30 or Greater</td>
<td>160 (25.1)</td>
<td>155 (28.5)</td>
<td></td>
</tr>
<tr>
<td>Median kg/m² body mass index (IQR):</td>
<td>5.5 (4.1–7.3)</td>
<td>6.1 (4.5–8.7)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Less than 7</td>
<td>0.19 (0.13–0.29)</td>
<td>0.19 (0.12–0.29)</td>
<td>0.657</td>
</tr>
<tr>
<td>Median gm prostate wt (IQR)</td>
<td>39.3 (31.3–50.1)</td>
<td>37.8 (30.1–49.8)</td>
<td>0.152</td>
</tr>
<tr>
<td>Less than 7</td>
<td>550 (67.3)</td>
<td>663 (93.0)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>7</td>
<td>267 (32.7)</td>
<td>50 (7.0)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>No. biopsy Gleason sum (%):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than 7</td>
<td>550 (67.3)</td>
<td>113 (15.8)</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>267 (32.7)</td>
<td>469 (66.6)</td>
<td></td>
</tr>
<tr>
<td>Greater than 7</td>
<td>15 (5–40)</td>
<td>15 (3–36)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Median max % core Ca (IQR)</td>
<td>0.33 (0.17–0.5)</td>
<td>0.33 (0.17–0.5)</td>
<td>0.168</td>
</tr>
</tbody>
</table>
the under and nonunder graded groups, and various clinical factors.

Based on univariate analysis statistically significant variables were age at diagnosis, race, PSA, biopsy Gleason sum, positive-to-total biopsy core ratio and maximal percent of cancer in biopsy cores ($p < 0.05$). Men with under grading on prostate needle biopsy were significantly older and more likely to be black. Furthermore, they had a higher tumor grade and tumor burden. We also observed that men with under grading had lower prostate weight and were more likely to be obese but neither association attained statistical significance ($p \geq 0.15$). The median number of biopsies during the 20-year period was 3 (IQR 2–6), partly because our institution is a tertiary referral center and many pathologists from elsewhere only send positive specimens for analysis.

Prostate weight and body mass index were marginally significant based on univariate analysis and these variables were included on multivariate analysis. After adjusting for multiple clinical factors, as on univariate analysis age at diagnosis, biopsy Gleason sum, PSA, prostate weight, positive-to-total biopsy core ratio and maximal percent of cancer in cores were significant predictors of biopsy Gleason sum under grading ($p < 0.05$, Table 2).

### Table 2. Independent predictors of Gleason sum under grading

<table>
<thead>
<tr>
<th>Predictive Variables</th>
<th>OR (95% CI)</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biopsy Gleason score 7</td>
<td>0.104 (0.073–0.148)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>PSA at diagnosis (log transformation)</td>
<td>1.946 (1.625–2.33)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Pos/total cores</td>
<td>1.505 (1.074–2.109)</td>
<td>0.019</td>
</tr>
<tr>
<td>Max % biopsy core Ca</td>
<td>1.005 (1.000–1.010)</td>
<td>0.021</td>
</tr>
<tr>
<td>Age at diagnosis</td>
<td>1.025 (1.009–1.042)</td>
<td>0.002</td>
</tr>
<tr>
<td>Prostate wt (log transformation)</td>
<td>0.588 (0.439–0.788)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Significant predictors were used to generate a nomogram that can be used to assess the probability of under grading after prostate needle biopsy (Fig. 1). Figure 2 shows nomogram calibration. The ideal line represents accurate prediction. Our nomogram predicted probability almost overlaps the ideal line with a tendency to under predict in higher risk groups. Predictive accuracy (concordance index C) was 72.4%. Including the maximal percent of positive cores in the model significantly improved the likelihood ratio ($p = 0.018$, data not shown). Furthermore, it improved the predictive ability of the model, increas-

**Figure 1.** Nomogram predicting biopsy Gleason sum under grading. Prob., probability

**Figure 2.** Result validation and predictive accuracy in 1,528 cases (B = 200 repetitions, boot mean absolute error = 0.008).
ing the AUC from 0.7195 to 0.7235. Hence, we included this variable in the final model.

DISCUSSION

In clinical management for prostate cancer decisions are based in part on biopsy Gleason sum and they assume a strong correlation between histological findings in biopsy and prostatectomy specimens. However, biopsy Gleason sum remains a poor predictor of the final pathological outcome with recent series showing a discrepancy rate as high as 58.3%. This is almost inevitable, given that prostate cancer is heterogeneous and multifocal. However, the goal of eliminating or minimizing these discrepancies cannot be ignored, given their clinical significance. Under grading biopsy Gleason sums, which is consistently the most common grading error, causes the greatest unease among physicians. Although over grading may lead to over treatment for clinically insignificant prostate cancer, of greater concern is the under treatment for clinically significant prostate cancer that under grading may create. While definitive surgical treatment can verify and correct these grading errors and appropriately alter the treatment course, patients who undergo radiotherapy may experience a poorer outcome as a result of under grading.

To minimize the risk of biochemical or clinical failure in patients who elect nonsurgical or nonactive prostate cancer treatment a physician should consider the risk of biopsy under grading. An error of only 1 grade may have serious implications on nonsurgical therapy, including the type and extent of radiotherapy, the potential need for hormonal therapy and the decision for or against active surveillance. According to the algorithm that Klotz developed to identify appropriate active surveillance candidates a key eligibility requirement for consideration is a Gleason sum of 6 or less. In a man presenting with Gleason sum 6 prostate cancer according to prostate biopsy a 1-grade increase could have a profound effect on the initial treatment course by excluding him from an active surveillance track. Also, consider a patient with a Gleason sum of 6 who is considered for brachytherapy or radiation therapy. Brachytherapy as monotherapy is typically reserved for patients with Gleason sum 6 or less disease, while in patients with a Gleason sum of 7 external beam radiation therapy is recommended in conjunction with brachytherapy.

Active surveillance is an appropriate option in some men. However, concern about under treating clinically significant cancer warrants that only men at low risk for progression undergo this treatment course. In general older men are usually recommended to undergo active surveillance, although they are also at higher risk for under grading. Given the impact of Gleason sum for identifying appropriate candidates for active surveillance, having a clinical tool that can correct for the error inherent in Gleason sum grading is essential.

To appropriately care for patients it is imperative to recognize the limitations of the diagnostic methods used. TRUS biopsy is by nature a random and partial sampling process, and we are increasingly finding that prostate cancer histological grading is by no means an exact process. The purpose of this nomogram is to correct for unavoidable errors based on the nature of the biopsy process and allow patients and physicians to make more informed decisions with regard to prostate cancer treatment. In the same way that a man with increased PSA but negative biopsy is counseled with regard to the risk of cancer missed by biopsy, physicians must also counsel patients with regard to the risk of a misgraded Gleason sum. Using this clinical tool we may enhance our ability to distinguish more appropriate candidates for active surveillance and other definitive therapies based on the risk of Gleason sum under grading.

A limitation of this study is interobserver variability since all specimens were not assessed by the same pathologist. Some groups suggested that eliminating this constraint does not introduce significant improvement. Another potential concern with regard to pathological findings is that there have been changes in grading with time. Today pathologists are more likely to assign worse Gleason scores than in the past. In our trial of reevaluating pathological Gleason sums in 204 men we found that pathologists today are more aggressive when assigning pathological Gleason sums than in 1995 to 1998. Historical Gleason score has a better role for predicting PSA recurrence compared to the contemporary Gleason score, as reevaluated in 2008 (data not shown). Since the patients in our study were collected during 20 years this could have had an impact on patient stratification into different Gleason sum categories. However, median time from diagnosis to surgery in our cohort was 2.3 months in each group. Since biopsy Gleason sums were read during the same period as pathological Gleason sums, this factor would not have impacted whether our cases were under graded.

Because of the extent to which the Gleason sum dictates treatment options, significant inaccuracies cannot be taken lightly. Therefore, the goal should be to find a way to eliminate errors in grading biopsy specimens. Histological grading is by nature a subjective process and while educational efforts have proven somewhat fruitful in improving interobserver reliability, there will always be room for interpretation in Gleason grading.
some evidence that increasing the number of biopsy cores may improve grading accuracy, there are also other data that refute this observation. In the absence of eliminating inaccuracies the ideal approach to improve oncological practice is to identify patients who are most at risk for biopsy Gleason sum under grading. In this way patients can be counseled appropriately with the knowledge that the Gleason sum may in fact be wrong and the prostate cancer risk may be significantly higher than initially intimated by TRUS guided prostate biopsy.

Kattan et al reported that there is usefulness for predictive nomograms in urology, specifically prostate cancer, and various nomograms have now been generated with much success in diagnostic and prognostic purposes. Therefore, we developed a nomogram as a means of identifying patients at highest risk for under grading to improve treatment outcomes. Chun et al identified diagnostic PSA, age at diagnosis, biopsy Gleason sum and clinical stage as factors predictive of Gleason sum under grading. Their nomogram used a cohort from Europe, where screening practices are not as widespread. Those patients have more advanced clinical stage and tumors are higher grade. Therefore, our nomogram is more generalizable to the American population due to current screening practices. Our analysis study incorporated prostate weight, the positive-to-total biopsy core ratio and the maximal percent of cancer in biopsy cores, which were also significant factors for predicting Gleason sum under grading.

CONCLUSIONS

This study highlights additional key predictors of Gleason sum under grading other than those already identified, including age at diagnosis, diagnostic PSA, clinical Gleason sum, prostate weight, the positive-to-total biopsy core ratio and the maximal percent of cancer in biopsy cores. The nomogram provides a clinical tool to estimate the probability of cancer under grading based on these factors. Its use in urological practice would aid physicians and patients in determining the course of treatment to identify appropriate candidates for active surveillance and other definitive therapies based on the risk of biopsy Gleason sum under grading.

REFERENCES


2. Epstein JI, Pizzo G and Walsh PC: Correlation of pathological findings with progression after radical retropubic prostatectomy. Cancer 1993; 71: 3582.


EDITORIAL COMMENTS

This study demonstrates that biopsy Gleason score remains an inaccurate predictor of final pathological Gleason sum even in the era of the extended pattern biopsy. Risk assessment tools such as nomograms that predict under grading can better inform physicians and patients of downstream prostate cancer risks and help delineate appropriate treatment plans.

However, nomograms are limited by input variables. Efforts should be made to include previously identified significant variables in subsequent nomograms with the goal of building on prior work. Examples of variables associated with misclassification of biopsy Gleason scores that the authors did not include in their nomogram are year of biopsy (reference 15 in article) and the characteristics of pathologists reading the biopsies.1 Incorporating these factors may have improved the predictive accuracy of their tool. Although variables such as these may not be available in retrospective analyses, well designed prospective studies with a priori variable selection would overcome this barrier.

Girish Kulkarni
Division of Urology
Department of Surgery
University of Toronto
Toronto, Ontario
Canada

REFERENCE

Given the wide range of therapeutic options for clinically localized prostate cancer, underestimating biopsy Gleason grade may have a critical impact on treatment decisions and may also influence followup strategies after definitive therapy. Thus, several statistical models based on classification and regression tree analysis1 or regression coefficient based nomograms (reference 20 in article) were developed to predict the risk of underestimating Gleason variants at biopsy.

These authors used a cohort of 1,701 patients. Readily available clinical data and quantitative biopsy information were combined in a multivariate risk model with 72.4% predictive accuracy. While most tools for the risk prediction of prostate cancer up grading show excellent performance after internal validation, none of the models has been confirmed by an external validation procedure. Given the known interinstitutional variability of Gleason grading, quantification of cancer or prostate volume and variation among several PSA assays, an external validation study including a head-to-head comparison between published prediction models would be of greatest interest for physicians to help select the most robust clinical tool for decision making in the future.

Thomas Steuber
Martini-Clinic, Prostate Cancer Center
University Hospital Hamburg-Eppendorf
Hamburg, Germany

REFERENCE

REPLY BY AUTHORS

In our previous study of 2,963 patients treated with radical prostatectomy 55.8% of diagnostic Gleason sums differed from those on final surgical pathology (58.6% in the 1988 to 1999 and 49.3% in the 2000 to 2006 groups, reference 5 in article). Diagnostic Gleason sums were under graded in 41.2% of cases and over graded in 12.8%. These results indicate that significant discrepancies between diagnostic and pathological Gleason sums, and severe diagnostic Gleason sum underestimation remain even in the contemporary era. Underestimated biopsy Gleason sum may negatively impact optimal decision making and prognosis, and widespread PSA screening is associated with a significant over diagnosis.1 However, in how many of these so-called over diagnosed cases is the bi-
opsy Gleason sum underestimated? What is the solution?

We used a population based statistical method to develop an algorithm to predict biopsy Gleason sum underestimation for clinicians and to assist patients decide whether to undergo therapy or repeat biopsy. Because a majority of the biopsies were evaluated by a single pathologist at our institution we did not consider that factor as a variable in the modeling process. The algorithm was internally validated using the bootstrap method. For future model development, biopsy protocol should be standardized, the year of biopsy and the pathologist or urologist who conducted the biopsy should be considered in the data modeling process, and the algorithms should be validated internally and externally.

REFERENCE