

UCSF

UC San Francisco Previously Published Works

Title

Whole-Slide Telepathology Grading of Prostate Cancer in Biopsies Using ISUP Criteria: Inter- and Intra-Observer Concordance and Implications for Active Surveillance.

Permalink

<https://escholarship.org/uc/item/2mf2j4ts>

Journal

Journal of Urology

ISSN

0021-0005

Authors

Lucia, M Scott
Callis, Sam
Tangen, Catherine
et al.

Publication Date

2025-12-29

DOI

10.1097/ju.0000000000004919

Peer reviewed

Please use this PDF proof to check the layout of your article. If you would like any changes to be made to the layout, you can leave instructions in the online proofing interface. First, return to the online proofing interface by clicking the  button at the top of the page, then insert a Comment in the relevant location. Making your changes directly in the online proofing interface is the quickest, easiest way to correct and submit your proof.

Please note that changes made to the article in the online proofing interface will be added to the article before publication, but are not reflected in this PDF proof.

DATE 1/9/2026

JOB NAME JU

ARTICLE JU-25-1517

QUERIES FOR AUTHORS Lucia et al.

This query form must be returned with all proofs for corrections

For indexing purposes, please confirm that author names have been correctly identified as given names (pink) and surnames (blue). Color in the byline will not appear on the final published version. Names will be indexed in PubMed as they are presented in the author line.

Author Queries

AU1) Please check the edits made to the article title and correct if necessary.

AU2) Please provide academic degree for corresponding author.

AU3) Please confirm funding/support information is correct as presented.

AU4) Please confirm conflict of interest information is correct as presented.

AU5) Please confirm ethics statement is correct as presented.

AU6) Please confirm author contributions are correct as presented.

AU7) Please note that there is no in-text citation for insight figure. Hence, please check the suggested in-text citation and correct the placement if necessary.

AU8) Please provide the accessed date for reference 4.

Whole-Slide Telepathology Grading of Prostate Cancer in Biopsies Using International Society of Urological Pathology Criteria: Interobserver and Intraobserver Concordance and Implications for Active Surveillance

M. Scott Lucia^{ID}, Sam Callis, Catherine Tangen, et al.

Correspondence: M. Scott Lucia (scott.lucia@cuanschutz.edu).

Full-length article available at <https://doi.org/10.1097/JU.0000000000004919>.

Study Need and Importance: Biopsy grading of PCa is critical for determining patient management. A number of refinements have been introduced to PCa grading to improve prognostic accuracy and identify the best candidates for active surveillance. There is also an emerging role for telepathology. The opportunity to assess (1) the ability to grade PCa by telepathology and (2) interobserver and intraobserver variability of grading using contemporary guidelines motivated this study.

What We Found: First, telepathology is satisfactory for grading PCa. Second, there was only fair to moderate grade agreement among our group of 11 genitourinary pathologists (or with themselves when shown an image a second time) for Gleason scores, especially when interpreting GG1 vs GG2 with low amounts ($\leq 10\%$) pattern 4. The heat map shows the range of grade group scorings for 120 cases ordered left to right from lowest to highest grade given for each case, while the table shows the total cases by number of reviewers in agreement. On a second round of grading by the same observers, 83% of images that were upgraded from GG1 to GG2 had $\leq 10\%$ pattern 4.

Limitations: Reviewers had to rely on 1 full-length image without access to additional levels of the tissue; and computer displays were not standardized among observers (Figure).

Interpretation for Patient Care: Variability in Gleason grading among pathologists and when the same pathologist re-reads a biopsy, particularly with low amounts of pattern 4, implies the possibility that an increase in grade from GG1 to GG2 with $\leq 10\%$ pattern 4 during surveillance biopsy may be due to this grading variability rather than a true increase in biologic aggressiveness. Care should be taken when counseling patients on active surveillance of this possibility. Telepathology for expert consultation on challenging cases and for central pathology review is appropriate.

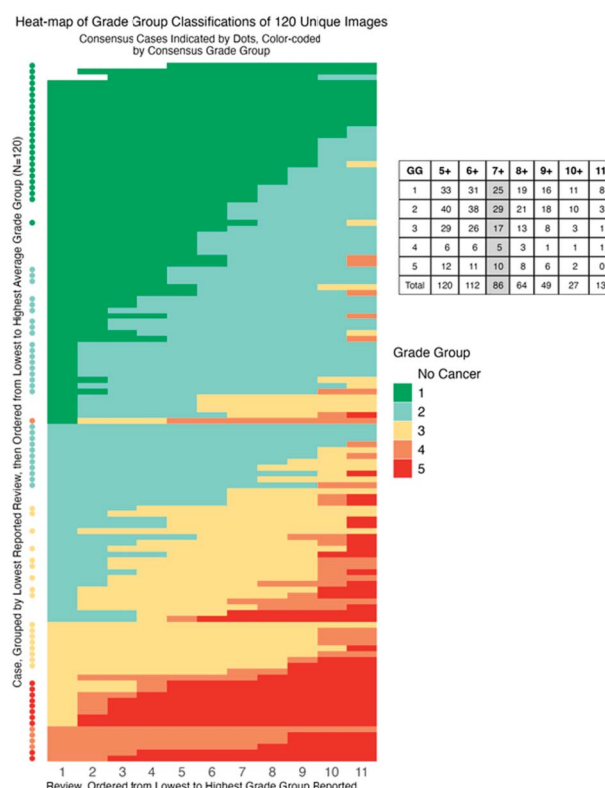


Figure. Heatmap of GG classifications for 120 unique case images. Grade scores are grouped left to right on the x-axis from the lowest GG rating to the highest given for each image irrespective of which reviewer assigned the grade. Cases are grouped from top to bottom on the y-axis by the lowest GG reported for each case and then ordered from lowest to highest mean GG within the groups. Cases that reached consensus (≥ 7 reviewers agreeing on grade) are denoted by the color-coded dots at the left of the figure. Consensus was reached on 72% (86/120) of the cases, with unanimous agreement on 11% (13/120). A table showing the total images by number of reviewers in agreement and GG is included on the right with the consensus definition highlighted in grey. All 120 unique case images had at least 5 reviewers agree on GG, grade group.

AU1

Whole-Slide Telepathology Grading of Prostate Cancer in Biopsies Using International Society of Urological Pathology Criteria: Interobserver and Intraobserver Concordance and Implications for Active Surveillance

M. Scott Lucia¹, Sam Callis², Catherine Tangen², Fawaz Almutairi³, Anastasios Chatzitoliou⁴, Oleksandr Kravtsov⁵, Sarah Lenon⁶, Varsha Manuchua⁷, Jeffry P. Simko⁸, Cornelia M. Szecei⁴, Yunshin Albert Yeh⁹, Lian Zhang¹ and Kenneth A. Iczkowski¹⁰

¹Department of Pathology, University of Colorado Anschutz Medical Campus, Aurora, Colorado

²SWOG Statistical Center, Fred Hutchinson Cancer Research Center, Seattle, Washington

³Department of Pathology, King Abdulaziz University, Jeddah, Saudi Arabia

⁴Severn Pathology, North Bristol NHS Trust, Bristol, United Kingdom

⁵Department of Pathology, SUNY Upstate Medical University, Syracuse, New York

⁶Department of Pathology, University of Southern California, Los Angeles, California

⁷Department of Pathology, University of Mississippi, Jackson, Mississippi

⁸Departments of Pathology and Laboratory Medicine, University of California—San Francisco, California

⁹Overton Brooks VA Medical Center, Shreveport, Louisiana

¹⁰Department of Pathology and Laboratory Medicine, University of California-Davis Health, Sacramento, California

Purpose: Subjectivity of PCa grading affects biomarker investigations, assessments of imaging, and treatment decisions, especially for active surveillance. We conducted the first telepathology concordance study of PCa needle biopsy grading using contemporary International Society of Urologic Pathology guidelines.

Methods: Whole slide images (n = 120) of PCa biopsy cases across grade groups (GGs) were examined by 11 urologic pathologists. Forty cases were blindly regraded, oversampling cases with lower agreement on first review. Pathologists provided GG, percentage and type of pattern Gleason 4, and tumor length. Consensus for a case was defined as GG agreement by ≥ 7 pathologists. GG agreement among and within pathologists was measured using model-based weighted Kappa statistics (κ_w).

Submitted September 2, 2025; accepted December 24, 2025; published 000.

Funding/Support: None.

Conflict of Interest Disclosures: The Authors have no conflicts of interest to disclose.

Ethics Statement: In lieu of a formal ethics committee, the principles of the Helsinki Declaration were followed.

Author Contributions:

Conception and design: Chatzitoliou, Tangen, Simko, Iczkowski, Lucia, Kravtsov.

Data acquisition: Chatzitoliou, Szecei, Simko, Iczkowski, Zhang, Lucia, Kravtsov, Lenon, Manuchua, Yeh.

Data analysis and interpretation: Chatzitoliou, Tangen, Szecei, Almutairi, Simko, Iczkowski, Lucia, Kravtsov, Callis.

Critical revision of the manuscript for scientific and factual content: Chatzitoliou, Tangen, Szecei, Almutairi, Simko, Iczkowski, Zhang, Lucia, Kravtsov, Callis, Manuchua.

Drafting the manuscript: Chatzitoliou, Tangen, Iczkowski, Lucia, Kravtsov, Callis, Lenon, Yeh.

Supervision: Chatzitoliou, Tangen, Szecei, Almutairi, Simko, Lucia, Kravtsov.

Statistical analysis: Chatzitoliou, Tangen, Callis.

Organizing study and participants, providing data, discussing results, writing, proofreading and editing: Iczkowski.

Producing research data and critically reviewing of the manuscript, including statistical analysis results and figure presentation of the results: Zhang.

Digital image annotation, identifying prostate cancer, assignment of Gleason score on prostate cancer, documenting amount or volume of tumor present: Lenon.

Reviewed all the cases independently: Manuchua.

Review and evaluate the virtual pathology slides and actively participate (including communication and review the initial and revised draft) in the study: Yeh.

Data Availability: The data from this project will be available as supplemental material on the journal website.

Corresponding Author: M. Scott Lucia, Department of Pathology, University of Colorado Anschutz Medical Campus, 12801 E. 17th Ave, Mail Stop 8104, Aurora, CO 80045 (scott.lucia@cuanschutz.edu).

AU3
AU4
AU5
AU6

AU2

Results: Consensus was achieved on 86/120 (72%) cases. Of 34 cases not achieving consensus, most were GG2 vs GG1 (41%) or GG2 vs GG3 (38%). The inter-rater κ_w was 0.34 (95% CI 0.28, 0.39) indicating fair agreement. Forty regraded cases had an intrarater κ_w of 0.49 (95% CI 0.46, 0.53; moderate agreement). Sixty-four percent of the repeated grades agreed. Of the repeated ratings that did not agree, 39% involved GG1 vs GG2. Of the GG1 vs GG2 paired repeated ratings, 77% reported 1-10% pattern 4, and indicated common observation was “poorly formed glands.”

Discussion: Inter-rater and intrarater variability in GG is significant, particularly discerning GG1 from GG2 with $\leq 10\%$ grade 4. These observations have profound implications for treatment decisions (surveillance vs definitive therapy) and studies of new biomarkers and imaging of prostate cancer.

Key Words: prostate cancer, grading, telepathology

For almost 60 years, the Gleason system has been the standard for PCa grading and is the most important prognostic factor obtained from prostate biopsy. A series of consensus conferences held by the International Society of Urological Pathology (ISUP) in 2005, 2014, and 2019 sequentially refined Gleason grading.¹⁻³ The refinements aimed to better classify the architectural patterns of PCa by restricting the definition of low grade PCa (pattern 3); defining pattern 4 to include fused, poorly formed, cribriform, and glomeruloid patterns; and establishing 5 grade groups (GGs), derived from the Gleason system to improve patient counseling and prognostic accuracy. Specifically, restricting the criteria for Gleason pattern 3 would help identify the best candidates for active surveillance (AS). Although the ISUP guidelines were expected to improve interobserver grade variability, this has not been rigorously tested.

The last 10 years have witnessed remarkable advances in whole-slide imaging of pathology tissue specimens. It is now possible to remotely examine whole tissue sections stained using routine stains that have been imaged in a format and viewed at magnifications controlled by the viewer. While initially introduced as a research and educational tool, improvements in quality and usability have made whole-slide imaging feasible for diagnostic use with commercial systems receiving FDA approval.⁴ In this study, we describe the results of the first multi-institutional PCa grading concordance study performed by telepathology based on contemporary ISUP guidelines.

MATERIALS AND METHODS

Eleven pathologists (hereafter “reviewers”) from the United States and Europe and with academic expertise in GU pathology participated in this study. A stratified sampling approach was used to obtain relatively equal numbers of reviewers who were early in their career (<5 years), midcareer (5-15 years), or senior (>15 years) to provide a representative sample and make results generalizable. Reviewers were trained on the Centriq imaging platform.

Image Collection and Distribution

Hematoxylin and eosin–stained prostate biopsy whole slide images ($n = 120$, 2 cores/image) were obtained from University Medical Center Utrecht (UMCU; Utrecht, Netherlands) and Protean BioDiagnostics (Orlando, FL) through an image library housed at Proscia, Inc. (Philadelphia, PA). Images were deidentified at Proscia and obtained with IRB approval and an appropriate contract from UMCU and Protean. Image selection was made by Proscia to include all Gleason grades and morphologies and copied to a study repository. The study was designed to include 4 batches of 40 slides for review. Reviewers were notified when each batch was available and given access to their own assigned repository. The design called for 40 unique images to be reviewed in Batch 1 and another 40 unique images in Batch 2. At that point, the study team identified 40 of the 80 cases from the first 2 batches to resubmit for repeat grading after a washout period of > 3 months. These included 20 that showed less agreement among the reviewers in terms of Gleason grade plus 20 randomly selected cases. For Batches 3 and 4, 20 unique images were included in each batch and 20 repeat images for a total of 120 unique images of which 40 were reviewed a second time, with 50% cases that had lower agreement on the first review and 50% randomly selected. Overall, each of the 11 reviewers examined 160 images, totaling 1760 individual pathology readings. Reviewers were blinded to the duplicate read aspect of the study.

Data Collection and Workflow

Reviewers examined each image and completed metadata fields in the Centriq platform including Gleason grade, % Gleason patterns, Morphologic pattern, and Tumor length (Supplemental Table 1, <https://www.jurology.com> lists the metadata fields). Metadata fields were reviewed for completeness, and if necessary, follow-up with reviewers was conducted to obtain missing values. After the reviewing period for each batch was complete, the batch repository was unshared so that reviewers no longer had access to completed batches. An end-of-study reviewer survey assessed the experience of the reviewers in using the Centriq system for PCa grading. Reviewers rated their experience on a scale of 1 (lowest) to 5 (highest) for image quality, navigation, metadata reporting, and overall experience.

Statistical Analysis

After all images were completed by the reviewers, the .csv files were sent to the statistical team. Reviewers were

deidentified to everyone in the study and given an ID number by Proscia. The 4 batches were combined for analysis when all reviews were complete.

The primary analysis was based on GGs, as defined by Gleason score.² An ordinal outcome with 5 levels was specified from the Gleason components that were recorded for each image. An ordinal score of 1 was given to a Gleason combination of 3 + 3 (GG1), a score of 2 was given for 3 + 4 (GG2), a score of 3 was given for 4 + 3 (GG3), a score of 4 was given for any of the following combinations: 4 + 4, 3 + 5, or 5 + 3 (GG4), and a score of 5 was given for 4 + 5, 5 + 4, or 5 + 5 (GG5). If a reviewer indicated “no cancer present,” the score was set to missing.

To estimate inter-rater and intrarater variability of GG rating and corresponding 95% CI, a generalized linear mixed model approach was used because it can manage correlated data from many raters, outcomes on an ordinal scale (Gleason score), duplicate ratings, and unbalanced or missing data.⁵⁻⁷ (see supplemental data for additional detail, <https://www.jurology.com>).

Because of no gold standard for “true” tumor grade in this study, we used the concept of consensus review. The first review of each image by each rater was used, so no repeat reviews by a single pathologist were included in the consensus assessment. For each image, we identified the mode GG (most frequently assigned grade among the 11 reviewers). The mode could theoretically vary from 3 to 11 reviewers agreeing on a GG. Consensus was defined as agreement on GG by 7 or more reviewers (64%) for a given case.

RESULTS

First Review of Images, Inter-rater Evaluation

[F1] Figure 1A shows a heatmap of GG classifications for the 120 unique case images grouped left to right from the lowest to highest GG rating for each image and from top to bottom by the lowest grade given for each image, then grouped by average tumor grade. Cases that achieved a consensus (≥ 7 reviewers agreeing on GG) are denoted by the colored dots to the left of the figure. Of the 120 cases, 86 (72%) reached consensus, with unanimous agreement on 11% (13/120). Of consensus cases, 29% (25/86) were GG1, 34% GG2 (29/86), 20% GG3 (17/86), 6% GG4 (5/86), and 12% GG5 (10/86). The level of agreement was highest for images defined as consensus GG1, with 85% of ratings being GG1, and lowest for GG4 with 75% of ratings being GG4 (Figure 2). Thirty-four cases failed to reach consensus. Of these, for 13 cases (38%), the mode was split between GG2 and GG3, reflecting disagreements about the percentage of pattern 4 in the biopsy, and in 14 cases (154 ratings), the mode was split between GG1 and GG2 with 74 GG1 ratings and 73 GG2. When GG2 was reported in these cases, it most often had $\leq 10\%$ pattern 4 (62/73 ratings, 85%) and either “poorly formed glands” (37/73, 51%) or fused glands (31/73, 42%) (Supplemental Table 2, <https://www.jurology.com>).

Intraobserver Duplicate Reviews

Figure 3 shows the initial grade assignments (Panel **[F3]** A) and repeat assignments (Panel B) for 40 cases, each regraded by 11 reviewers (total of 440 ratings). The largest number of grade changes occurred for tumors initially graded as GG4 with 62% of the grades either going up 1 GG to GG5 or down 1 GG to GG3 (Supplemental Table 3, <https://www.jurology.com>). Only 55% of images initially graded as GG1 remained as GG1 upon regrading, with 42% increasing on second review to GG2. Conversely, 23% of tumors initially graded as GG2 were downgraded to GG1 (Supplemental Table 3, <https://www.jurology.com>). Of the 19 images that failed to reach consensus on first read, 10 again failed to reach consensus on second read illustrating the level of difficulty in assigning a grade category for these cases.

Tumor Characteristics and GG Agreement

To determine whether characteristics of the tumor on biopsy affected re-gradings between GG1 and GG2, the reported linear extent of tumor, percent of pattern 4, and type of pattern 4 were compared between those tumors that changed GG vs those that did not (Table 1). There were 440 total repeat grade ratings for 40 images (40 images $\times$ 11 reviewers). Sixty-nine of these had an original grade of GG1 and 138 as GG2. Of the 69 ratings originally graded as GG1, 55% were also graded as GG1 on re-review, while 42% were upgraded to GG2 and 3% to $\geq$ GG3. For those images upgraded to GG2, the reported percent of grade pattern 4 was 1% to 10% in 82% (1%-5% in 52%) indicating small amounts of the higher-grade component, not recognized on first read, affected the regrade. Of the 138 ratings originally graded as GG2, 64% were also graded as GG2 on re-review with 23% downgraded to GG1 and 13% upgraded to $\geq$ GG3. For those images that were downgraded by a reviewer to GG1, 72% were originally rated as having 1% to 10% grade pattern 4, but were not recognized as having any pattern 4 component on re-review. Where consensus GG varied between 1 and 2, the most frequent types of pattern 4 were reported as “poorly formed glands” (56%) and “fused glands” (31%). By contrast, for those cases that were consistently coded as consensus GG2, poorly formed glands were less frequent at 43%. Linear extent of tumor did not significantly differ between tumors that changed GG after re-review and those that did not (data not shown).

Kappa Statistics

Results from the model-based evaluation of inter-rater and intra-rater agreement, accounting for the correlations in the data, can be found in Supplemental

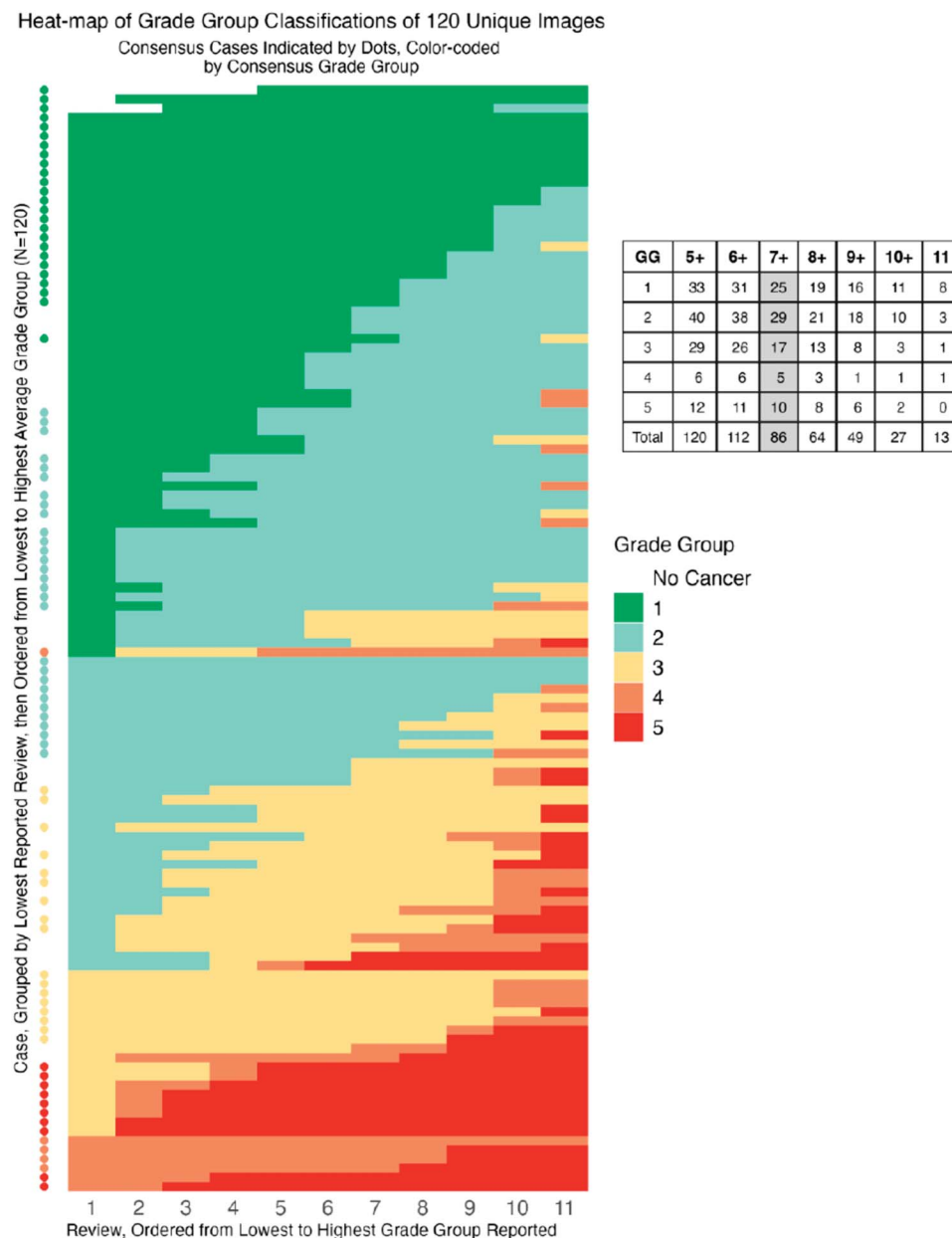


Figure 1. Heatmap of GG classifications for 120 unique case images. Grade scores are grouped left to right on the x-axis from the lowest GG rating to the highest given for each image irrespective of which reviewer assigned the grade. Cases are grouped from top to bottom on the y-axis by the lowest GG reported for each case and then ordered from lowest to highest mean GG within the groups. Cases that reached consensus (≥ 7 reviewers agreeing on grade) are denoted by the color-coded dots at the left of the figure. Consensus was reached on 72% (86/120) of the cases, with unanimous agreement on 11% (13/120). A table showing the total images by number of reviewers in agreement and GG is included on the right with the consensus definition highlighted in grey. All 120 unique case images had at least 5 reviewers agree on GG. GG, grade group.

Table S4 (<https://www.jurology.com>). In brief, inter-rater κ_w for the 5-level ordinal GG is 0.34 and the intra-rater GG κ_w is 0.49. The inter-rater GG κ_w did not improve significantly from the first to the second review of the same images (0.33), suggesting there was no observed learning curve for the telepathology protocol. General guidelines for interpreting kappa values group 0.21 to 0.40 as fair agreement, 0.41 to 0.60 moderate agreement, and 0.61 to 0.80 suggests substantial agreement.⁸

Pathologist Impression of Telepathology Review

Table 2 summarizes the experience of the pathologists using the Concentriq system for this study. Most reviewers were satisfied with image quality (score 2.3, range 1-4) and image navigation (2.3, range 1-5). When compared with microscopic examination of glass slides (1 = better than glass, 5 = worse than glass), the overall experience with telepathology was intermediate (3.2, range 1-4). Most reviewers indicated that reviewing telepathology was slower than glass, but

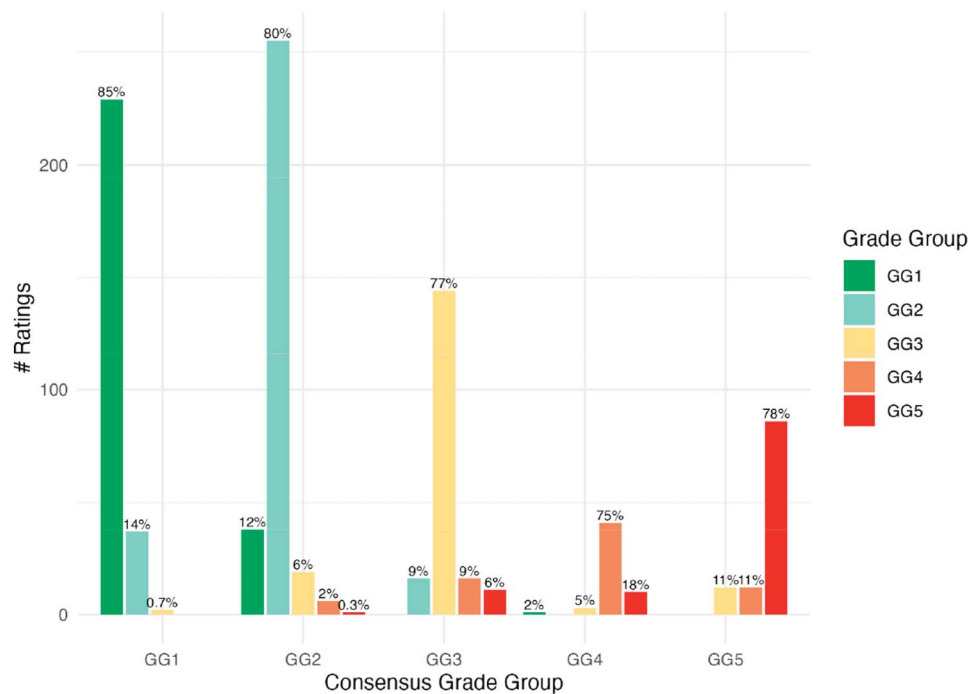


Figure 2. Variability of grade scores for each consensus GG for 86 cases achieving consensus. A total of 946 individual grade ratings were recorded from 11 pathologists. The agreement of the total image scores is indicated above each consensus bar. GG4 consensus tumors showed the lowest agreement at 75%, indicating that 25% of ratings were assigned another GG. The highest level of agreement was for images defined as consensus GG1, with 85% of ratings being GG1. GG, grade group.

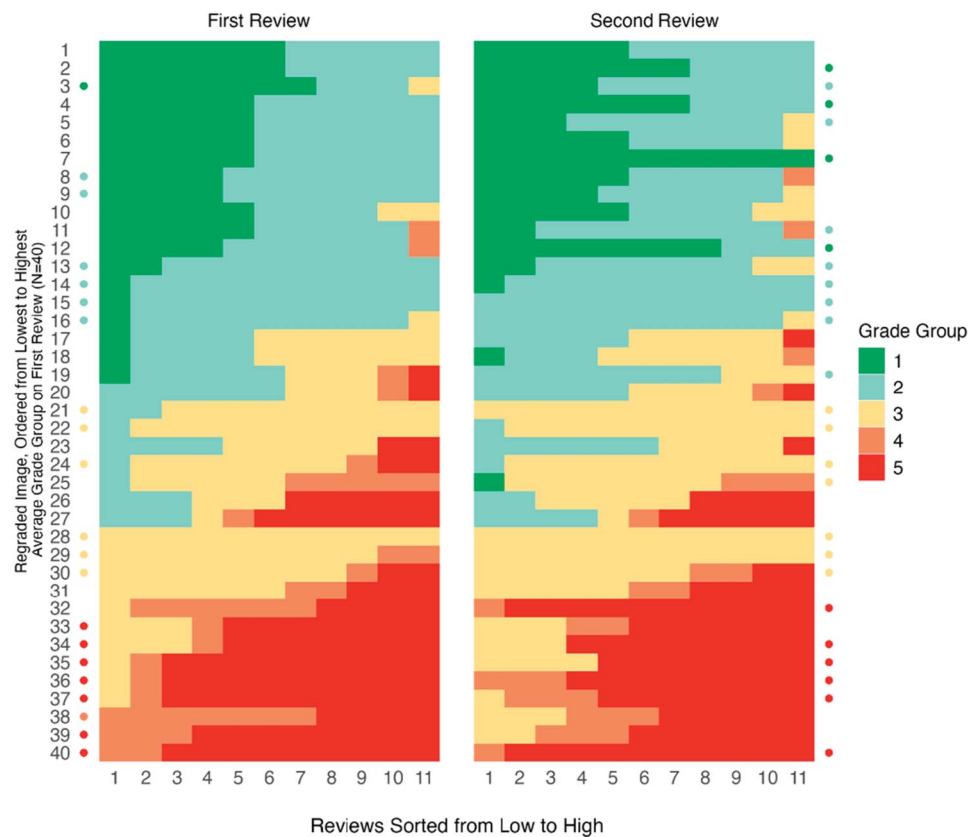


Figure 3. Heatmap of GG classification for 40 cases that were regraded after a washout period of > 3 months comparing the initial review grade with the repeat grade. Cases that reached consensus are denoted by dots the left or right of the panels. Grade changes were most prevalent for cases originally scored as GG1 or GG2. GG, grade group.

Table 1. Reported Characteristics of 207 Ratings Originally Submitted by Eleven Reviewers as GG1 or GG2 on First Review That Were Regraded on Repeat Review

Initial grade	Regrade	Count (%)	N (%) with 1%-5% pattern 4	N (%) with 1%-10% pattern 4	Pattern of grade 4			
					Poorly formed glands	Fused glands	Cribriform	Unspecified
GG1 (N = 69)	No change	38 (55%)	-	-	-	-	-	-
	GG2	29 (42%)	15 (52%)	24 (83%)	16 (55%)	9 (31%)	1 (3%)	3 (10%)
	GG3+	2 (3%)	-	-	1 (50%)	-	1 (50%)	-
GG2 (N = 138)	No change	88 (64%)	33 (38%)	57 (65%)	38 (43%)	34 (44%)	13 (15%)	3 (3%)
	GG1	32 (23%)	18 (56%)	23 (72%)	18 (56%)	10 (31%)	4 (13%)	-
	GG3+	18 (13%)	3 (17%)	8 (45%)	10 (56%)	4 (13%)	3 (3%)	5 (4%)

GG, grade group.

Numbers indicate individual image assessments made that comprised each grade group by 11 reviewers. The percent and pattern of grade 4 is given for those tumors that had grade 4 reported as part of the grade group.

that speed and comfort improved after experience with the Concentriq system.

DISCUSSION

PCa grading remains the most important prognostic factor determined on biopsy, providing an estimate of the tumor's biologic aggressiveness. In the clinic, it guides management decisions while in research, it is used to evaluate biomarkers, imaging, and treatment outcomes. Refinements to the Gleason system and the establishment of ISUP grade groupings (GG1-GG5) provide criteria for grading and improved stratification of outcomes.¹⁻³ One important effect of the refinements introduced at the 2005 and 2014 ISUP Consensus Conferences was to define low grade (GG1) tumors more stringently so that patients and their providers "could be reassured that they have...the lowest grade tumor possible, which in most cases can be followed with active surveillance."² Nonetheless, key to achieving this goal is uniformity and reproducibility of Gleason grading on biopsy.

We examined the level of concordance for Gleason grading using ISUP recommendations from the consensus conferences of 2005, 2014 and 2019 on 120

unique images using the Concentriq telepathology system. Telepathology provided several advantages for this study. Reviewers from around the world participated without the need for travel or round-robin shipping of slides. Reviewers graded images in isolation without the possibility of being biased by discussions often arising in a group setting. Importantly, the protocol facilitated re-reads of images in a "masked" fashion after a washout period to assess intra-rater variability. Finally, it supported the utility of whole-slide telepathology for the grading of PCa using ISUP guidelines. The Concentriq platform was approved for telepathology by the FDA in February 2024.⁴

Our telepathology protocol also introduced limitations to the study that may have affected the results. Reviewers had to rely on 1 full-length image of the biopsy for grading without the ability to ask for additional levels in which tissue blocks are resectioned to provide additional slides for examination. Comparison of multiple levels may resolve questions about true gland fusion vs tangential sectioning. Computer hardware was not standardized for monitor size or resolution and original glass slides were not circulated for comparison. Most reviewers indicated that grading cases by telepathology was slower than their experience with glass slides, but they gained comfort with grading as they accrued experience with the system.

Although the inter-rater weighted kappa of 0.34 was similar to other concordance studies in the literature,⁹⁻¹² the data indicated that interobserver variability in Gleason grading of PCa remains problematic on biopsies, particularly for recognition and quantification of pattern 4. An important aspect of the study was the ability to assess intra-rater variability on a subset of images reggraded after a washout period from the original review. Regrading data provided insight into the elements that often lead to grading discrepancies. Because each reviewer was reggrading the same image as previously graded, grade changes are more likely to reflect features of the tumor than

Table 2. Experience of the Pathologists Using the Concentriq Telepathology System for the Review of Prostate Biopsy Images

Concentriq for Research (1 = lowest, 5 = highest) Mean (IQR)	Rate efficiency of navigating the image 2.45 (1.5-3.5)
	Rate ease of using metadata panel 2 (1-3)
	Rate experience with measuring tool 2.64 (1.5-4)
	Rate the quality of images 2.36 (1.5-3.5)
Did your ability to use Concentriq for research improve over time?	Yes (9) No (2)

GG, grade group.

Pathologists rated the system for viewing and navigating the images and the metadata reporting tool on a scale of 1 (lowest) to 5 (highest).

limitations of telepathology. Images were upgraded from GG1 to GG2 42% of the time upon repeat grading, and 83% of the images were reported as 1-10% Gleason pattern 4 when upgraded. Reviewers reported the pattern 4 component as “poorly formed glands” in 55% of these regrades. These same observations held true for the images that failed to achieve consensus: in images in which reviews were split between GG1 and GG2, 85% of GG2 scores were reported to have 1-10% pattern 4 and 51% reported “poorly formed glands.” Thus, small amounts of higher grade patterns blur the distinction between GG1 and GG2, particularly when interpreting whether glands are the “poorly formed” pattern of Gleason 4. Intra-rater agreement in this study is likely an underestimate because half of the images selected for repeat review had been considered “difficult” because of lack of agreement by the raters in their first review.

Similarly, tumors originally classified as GG4 were often reclassified upon regrading as either GG3 or GG5 when reviewers identified small amounts of Gleason patterns 3 or 5 on regrading that were not initially appreciated. While a diagnosis of a GG3, GG4 or GG5 tumor often affects prognosis and choice of therapy, few such tumors would be managed with AS.

More concerning is the discrimination of GG2 from GG1 or GG3. Most patients with GG1 tumors are candidates for AS. While some providers may offer AS for GG2 tumors, particularly with $\leq 10\%$ Gleason pattern 4, others may not. AS is seldom warranted for GG3. Therefore, grading variability between GG2 and GG1 or GG3 is likely to alter treatment recommendations. The data also suggest that AS monitoring “progression” of PCa grade on periodic surveillance biopsies by the same or different pathologist may prompt definitive treatment while not truly reflecting a change in the biology of the tumor.

Studies have demonstrated that GG2 tumors with $\leq 10\%$ Gleason pattern 4 on biopsy are associated with favorable low-risk pathology on subsequent prostatectomy and outcomes similar to GG1 cancer.^{13,14} In support of this contention, our data show a high rate of interobserver and intraobserver

variability for distinguishing GG2 with $\leq 10\%$ pattern 4 from GG1 and subjectivity of interpretation of “poorly-formed glands.” Our data also indicate that published outcomes for GG1 are likely contaminated by GG2 with low amounts of pattern 4. This also aligns with recent literature emphasizing that the length of Gleason ≥ 4 cancer on biopsy,^{15,16} or absolute volume of Gleason ≥ 4 on prostatectomy¹⁷ outperform the ISUP GG grading system itself, for accurate treatment choice and outcome prediction. Ultimately, the pathological examination of biopsies is a restricted to using a tiered grading system to categorize tumor biology that exists in a continuum and is subject to the vagaries of human interpretation. Whole-slide imaging does present the opportunity for on-going efforts to employ artificial intelligence approaches to better predict tumor behavior.

CONCLUSIONS

Our study adds to the accumulating data demonstrating the feasibility of telepathology as an aid for pathologists for diagnosis and grading of prostate biopsies and a useful tool for consultations, second opinions, mentoring, and central review of clinical trials. It also demonstrates that Gleason grading does not have high reliability, either among pathologists or when the same pathologist re-reads the slide, particularly when separating GG1 (Gleason 3 + 3) from GG2 (3 + 4) with small amounts ($\leq 10\%$) of grade pattern 4 and when the poorly formed pattern is present. Care should be taken to ensure that physicians and patients are aware that a GG2 tumor with minimal amounts of grade pattern 4 may have a biologic potential more akin to a GG1 tumor because of this variability of interpretation. This limitation of the Gleason grading system should inform treatment decision-making.

ACKNOWLEDGMENTS

We thank the people at Proscia, Inc. for access to their image library and training on the Concentriq system. The authors wish to thank Ian M. Thompson, Jr, MD, for encouraging the conduct of this study.

REFERENCES

- Epstein JI, Allsbrook WC, Amin MB, Egevad LL; ISUP Grading Committee. The 2005 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma. *Am J Surg Pathol*. 2005;29(9):1228-1242. doi:10.1097/01.pas.0000173646.99337.b1
- Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA; Grading Committee. The 2014 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma: definition of grading patterns and proposal for a new grading system. *Am J Surg Pathol*. 2016;40(2):244-252. doi:10.1097/PAS.0000000000000530
- van Leenders GJLH, van der Kwast TH, Grignon DJ, et al., ISUP Grading Workshop Panel Members. The 2019 International Society of Urological Pathology (ISUP) consensus conference on grading of prostatic carcinoma. *Am J Surg Pathol*. 2020;44(8):e87-e99. doi:10.1097/PAS.0000000000001497
- 21 CFR 864.3700-510(k) substantial equivalence determination decision summary, K230839, US Food and Drug Administration; 2024. https://www.accessdata.fda.gov/cdrh_docs/reviews/K230839.pdf

5. Nelson KP, Edwards D. Measures of agreement between many raters for ordinal classifications. *Stat Med*. 2015;34(23):3116-3132. doi:10.1002/sim.6546
6. Zhou TJ, Raza S, Nelson KP. Methods of assessing categorical agreement between correlated screening tests in clinical studies. *J Appl Stat*. 2021;48(10):1861-1881. doi:10.1080/02664763.2020.1777394
7. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20(1):37-46. doi:10.1177/001316446002000104
8. Sim J, Wright CC. The Kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther*. 2005;85(3):257-268.
9. Melia J, Moseley R, Ball RY, et al. A UK-based investigation of inter- and intra-observer reproducibility of Gleason grading of prostatic biopsies. *Histopathology*. 2006;48(6):644-654. doi:10.1111/j.1365-2559.2006.02393.x
10. Al-Maghrabi JA, Bakshi NA, Farsi HMA. Gleason grading of prostate cancer in needle core biopsies: a comparison of general and urologic pathologists. *Ann Saudi Med*. 2013;33(1):40-44. doi:10.5144/0256-4947.2013.40
11. Shah RB, Li J, Cheng L, et al. Diagnosis of Gleason pattern 5 prostate adenocarcinoma on core needle biopsy: an interobserver reproducibility study among urologic pathologists. *Am J Surg Pathol*. 2015;39(9):1242-1249. doi:10.1097/PAS.0000000000000442
12. Ozkan TA, Erucar AT, Cebeci OO, Memik O, Ozcan L, Kusonmaz I. Interobserver variability in Gleason histological grading of prostate cancer. *Scand J Urol*. 2016;50(6):420-424. doi:10.1080/21681805.2016.1206619
13. Klotz L. Active surveillance in intermediate-risk prostate cancer. *BJU Int*. 2020;125(3):346-354. doi:10.1111/bju.14935
14. Sato S, Kimura T, Yorozu T, et al. Cases having a Gleason score 3+4=7 with <5% of Gleason pattern 4 in prostate needle biopsy show similar failure-free survival and adverse pathology prevalence to Gleason score 6 cases in a radical prostatectomy cohort. *Am J Surg Pathol*. 2019;43(11):1560-1565. doi:10.1097/PAS.0000000000001345
15. Dean LW, Assel M, Sjoberg DD, et al. Clinical usefulness of total length of gleason pattern 4 on biopsy in men with grade group 2 prostate cancer. *J Urol*. 2019;201(1):77-82. doi:10.1016/j.juro.2018.07.062
16. Perera M, Assel MJ, Benfante NE, et al. Oncologic outcomes of total length gleason pattern 4 on biopsy in men with grade group 2 prostate cancer. *J Urol*. 2022;208:309-316. doi:10.1097/JU.0000000000002685
17. Scuderi S, Tin AL, Klug J, et al. The absolute volume of Gleason pattern 4 on radical prostatectomy is more strongly associated with advanced stage and biochemical recurrence than gleason grade groups. *J Urol*. 2025;213(6):722-729. doi:10.1097/JU.0000000000004484