HHS Public Access

Author manuscript

Histopathology. Author manuscript; available in PMC 2021 January 27.

Published in final edited form as:

Histopathology. 2018 July ; 73(1): 8–18. doi:10.1111/his.13471.

Utility of Pathology Imagebase for Standardization of Prostate Cancer Grading

Lars Egevad[1], **Brett Delahunt**[2], **Dan M Berney**[3], **David G Bostwick**[4], **John Cheville**[5], **Eva Comperat**[6], **Andrew J Evans**[7], **Samson W Fine**[8], **David J Grignon**[9], **Peter A Humphrey**[10], **Jonas Hörnblad**[1], **Kenneth A Iczkowski**[11], **James G Kench**[12], **Glen Kristiansen**[13], **Katia RM Leite**[14], **Cristina Magi-Galluzzi**[15], **Jesse McKenney**[16], **Jon Oxley**[17], **Chin-Chen Pan**[18], **Hemamali Samaratunga**[19], **John R Srigley**[20], **Hiroyuki Takahashi**[21], **Lawrence D True**[22], **Toyonori Tsuzuki**[23], **Theo van der Kwast**[7], **Murali Varma**[24], **Ming Zhou**[25], **Mark Clements**[26]

[1]Department of Oncology and Pathology, Karolinska Institutet, Stockholm, Sweden [2]Department of Pathology and Molecular Medicine, Wellington School of Medicine and Health sciences, University of Otago, Wellington, New Zealand [3]Barts Cancer Institute, Queen Mary University of London, London, UK [4]Bostwick Laboratories, Orlando, FL, USA [5]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA [6]Hôpital Tenon, HUEP, AP-HP, UPMC Paris VI, Sorbonne Universities, Paris, France [7]University Health Network, Laboratory Medicine Program, Toronto General Hospital, Toronto, ON, Canada [8]Department of Pathology, Memorial Sloan-Kettering Cancer Center, New York, NY, USA [9]Department of Pathology and Molecular Medicine, Indiana University School of Medicine, Indianapolis, IN, USA [10]Department of Pathology, Yale University School of Medicine, New Haven, CT, USA [11]Department of Pathology, Medical College of Wisconsin, Milwaukee, WI, USA [12]Department of Tissue Pathology and Diagnostic Oncology, Royal Prince Alfred Hospital and Central Clinical School, University of Sydney, Sydney, New South Wales, Australia [13]Institute of Pathology, University Hospital Bonn, Bonn, Germany [14]Department of Urology, Laboratory of Medical Research, University of Sao Paulo Medical School, Sao Paulo, Brazil [15]Department of Anatomic Pathology, Cleveland Clinic Lerner College of Medicine, Cleveland Clinic, Cleveland, OH, USA [16]Pathology and Laboratory Medicine Institute, Cleveland Clinic, Cleveland, OH, USA [17]Department of Cellular Pathology, Southmead Hospital, Bristol, UK [18]Department of Pathology, Taipei Veterans General Hospital, Taipei, Taiwan [19]Aquesta Uropathology and University of Queensland, Brisbane, Queensland, Australia [20]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada [21]Department of Pathology, The Jikei University School of Medicine, Tokyo, Japan [22]Department of Pathology, University of Washington Medical Center, Seattle, WA, USA

Address for correspondence: Prof. Lars Egevad, Dept. of Oncology-Pathology, Karolinska Institutet, Radiumhemmet P1:02, Karolinska University Hospital, 171 76 Stockholm, Sweden, Phone: +46-8 5177 5492, lars.egevad@ki.se.

Author Contributions

Lars Egevad: Study design, data analysis, writing of manuscript

Jonas Hörnblad: Computer programming, web design

Mark Clements: Statistical advice

Brett Delahunt: Uploading of images, voting, editing of manuscript

Dan M Berney, David G Bostwick, John Cheville, Eva Comperat, Andrew J Evans, Samson W Fine, David J Grignon, Peter A Humphrey, Kenneth A Iczkowski, James G Kench, Glen Kristiansen, Katia RM Leite, Cristina Magi-Galluzzi, Jesse McKenney, Jon Oxley, Chin-Chen Pan, Hemamali Samaratunga, John R Srigley, Hiroyuki Takahashi, Lawrence D True, Toyonori Tsuzuki, Theo van der Kwast, Murali Varma, Ming Zhou: Uploading of images, voting, discussions on study design

[23]Department of Surgical Pathology, Aichi Medical University, School of Medicine, Nagoya, Japan [24]Department of Cellular Pathology, University Hospital of Wales, Cardiff, UK [25]Department of of Pathology, UT Southwestern Medical Center, Dallas, TX, USA [26]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

Abstract

Aims: Despite efforts to standardize grading of prostate cancer, there is still a considerable variation in grading practices even among experts. In this study we describe the use of Pathology Imagebase, a novel reference image library, for setting an international standard in prostate cancer grading.

Methods and Results: The International Society of Urological Pathology (ISUP) recently launched a reference image database supervised by experts. A panel of 24 international experts in prostate pathology independently reviewed microphotographs of 90 cases of prostate needle biopsies with cancer. A linear weighted kappa of 0.67 (95% confidence interval 0.62 – 0.72) and consensus was reached in 50 cases. The inter-observer weighted kappa varied from 0.48 to 0.89. The highest level of agreement was seen for Gleason score (GS) 3+3=6 (ISUP grade 1), while higher grades and particularly GS 4+3=7 (ISUP grade 3) showed considerable disagreement. Once a 2/3 majority was reached, images were automatically moved into a public database available for all ISUP members at www.isupweb.org. Non-members are able to access a limited number of cases.

Conclusions: It is anticipated that the database will assist pathologists to calibrate their grading and, hence, decrease inter-observer variability. It will also help identify instances where definitions of grades need to be clarified.

Keywords

standardization; database; grading; ISUP; prostate

Introduction

One of the greatest challenges in histopathology is the reproduction of subjective estimates of semiquantitative parameters such as tumor grade.[1] Previous attempts to standardize grading have often been undertaken by small groups of pathologists, while a presumed gold standard has usually been established by one or a few pathologists, sometimes from the same institution.[2–4] This strategy does not take a broader consensus into account and as a consequence it may be difficult to disseminate criteria to the wider pathology community.

In this study we have utilized a novel mechanism for the standardization of pathology reporting, developed under the auspices of The International Society of Urological Pathology (ISUP). [5] This methodology has involved the establishment of a reference image library of graded prostate cancers. Through independent review of microphotographs, members of an international expert panel have built and validated this image database in order that it will serve as a reference library for pathologists.

Material and methods

A web-based image library, named Pathology Imagebase, was established at http://www.isupweb.org. A detailed description of the methodology relating to the construction of the database has recently been reported.[5] Three groups with 24 experts in each were formed to evaluate prostate, bladder and renal pathology. The prostate panel was designed to have broad international representation with participation by pathologists from Asia (3), Australasia (3), Europe (6), North America (11) and South America (1) (Table 1). The experts were selected based on their international reputation and scientific production. A Medline search informed that they had authored an average number of 105 articles on prostate pathology (range 21–321) with an average of 39 first or last author papers (range 5 – 190).

The experts were invited to upload microphotographs in jpg format onto the central Imagebase. The panel was then asked to vote on their preferred grading independent of each other. While voting, the panel members were unable to access the results of the voting by other panel members. The consensus level was defined as at least 2/3 of all possible votes cast for each image. This level was in line with the agreed threshold for consensus decisions at all consensus meetings organized under the auspices of ISUP the past decade. To ensure that the results were not biased by low voting numbers, consensus was considered to have been reached when 16 votes had been cast for a single case. In addition, the voters were asked to determine if the case was typical for a specific grade or if it was considered borderline towards a higher or lower grade. Panel members were also asked to state if the technical quality of the images was acceptable or unacceptable. Once a case achieved 16 votes for a single grading option it was automatically transferred to a public database that is currently available to all ISUP members. Despite the transfer of the images to the public website, the panel members who had not completed voting on all the cases could continue to vote, but until they did they remained blinded to the consensus diagnosis.

The detailed voting results were tabulated in a non-public database. In the public database the comments and voting results of the 24 observers were listed anonymously and labelled 1–24 in random order. The distribution of the replies was displayed both as numerical values and as pie charts. Ten selected consensus cases from the grading project were also displayed in a free version of the Imagebase that was not password protected.

The initial project launched by the prostate panel aimed to standardize prostate cancer grading. Images of 90 prostate cancers diagnosed on core needle biopsy were uploaded by one of the authors (L.E., labelled observer 16) between May and September 2015. All images were obtained from a single biopsy core from men who were recruited to the STHLM3 study, a population-based screening study among men aged 50–69 years.[6] None of the patients had a previous diagnosis of cancer or had received hormonal treatment. The case selection was non-consecutive and it was intended that the images represented a broad range of prostate cancer morphology.

Each case was assigned an identification number when enrolled onto the Pathology Imagebase. All experts were asked to vote by marking an option from a multiple-choice

menu that had five alternatives (ISUP grades 1 – 5). Participants were also permitted to enter other grading combinations where the Gleason scores (GS) could also be assigned (Table 2). In order to distinguish between the various GS that constitute ISUP grade 4, grade 4 was defined as GS 4+4=8 in the multiple-choice list, while GS 3+5=8 and GS 5+3=8 were entered as Other. In the statistical analyses, GS 2+3=5 was included in ISUP grade 1 data and GS 3+5=8 in ISUP grade 4 data. The marginal distribution for grades 1 – 5 was 0.25, 0.34, 0.17, 0.14 and 0.10, respectively.

O'Connell and Dobson estimators were utilized to summarise the kappas for multiple raters and the average agreement for each case was assessed using linear weights.[7] The mean weighted kappas by a pathologist were calculated using Schouten's methodology.[8] To consider agreement for a specific proposed grade, we dichotomised the results and used unweighted kappas. All of the kappas were calculated using the Magree package in R.[9]

Results

All but one panel member graded each of the 90 cases. The exception was a pathologist who graded all non-consensus cases and 17 of those that already had reached consensus. The results from this observer were excluded from the kappa statistics as the results were incomplete, but included in other analyses as the consensus decisions were unaffected. The option to label a case as technically unacceptable was employed by three observers in a total of six cases (0.3%; 6 of 2137 votes). None of these panel members voted for the same case and, as a consequence, no case was removed from the database.

In January 2016 a consensus, as defined in the Materials and Methods, had been reached in 50 of 90 cases (55.6%). The distributions of the assigned grades for each of the cases and for all responses received are shown in Table 2. Among cases uploaded the percentage for which a grading consensus was reached for ISUP grades 1–5 were 80.0% (16/20), 53.1% (17/32), 21.4% (3/14), 45.5% (5/11) and 46.2% (6/13), respectively (Table 3). The consensus grade differed from the initial grading of the case by the submitting pathologist in only three cases (one with higher consensus grade and two with lower grades). Among the seven cases listed as Other grade (Table 2), one case of GS 2+3=5 and six cases of GS 3+5=8 were grouped with ISUP 1 and 4, respectively, for further analysis;[10] however, this did not alter any of the consensus decisions.

The concordance between consensus grades and initial assigned grades were 79.7% - 86.1% with highest agreement on ISUP grade 1 (Table 4). The concordance between grades assigned by the panel and initial assigned grades in the non-consensus cases were 41.7% - 54.2%, again with highest agreement being for ISUP grade 1 (Table 5). Examples of consensus and non-consensus cases are shown in Figs. 1 and 2, respectively.

The number of cases per observer that were discordant with the consensus grade ranged from 2 to 17 (mean 8.3, median 8) among the 23 observers with complete reporting (Table 6). The number of cases that were over-graded and undergraded vs. the consensus grades was 0 to 15 (mean 4.5, median 3) and 0 to 10 (mean 3.8, median 3), respectively.

Grades were reported as either borderline lower (i.e. bordering onto a lower grade), typical or borderline higher (i.e. bordering onto a higher grade). The 23 observers with complete voting results reported a mean number of 13.6 (15.1%), 55.5 (61.7%) and 20.9 (23.2%) cases, respectively, in these categories. This reporting was highly variable between observers and ranged from 0 – 46 (0% - 51.1%), 30 – 78 (33.3% - 86.7%) and 0 – 25 (0% - 27.8%), respectively.

Among non-consensus cases, four main patterns of disagreement were noted; 1) 14 cases mainly bordering between grades 1 and 2 (Figure 2A-B); 2) 8 cases mainly bordering between grades 2 and 3 (Figure 2C-D); 3) 10 cases mainly bordering between grades 3 and 4; 4) 7 cases spread across grades 4 – 5 and sometimes even grades 3 – 5 or 2 – 5 (Figure 2E-F). In one case the votes were spread across grades 1 – 4 and this result did not fit into any of the patterns of disagreement noted above. Among the 8 cases bordering between grades 1 and 2, the focus suggestive of the higher grade was dominated by poorly formed glands in 11 cases, fused glands in two cases, both poorly formed and fused glands in one case and glomeruloid structures in one case. Cribriform glands were not observed in any of these cases.

Kappa statistics

Results from one of the participating pathologists were excluded from the analysis due to incomplete data. For the remaining 23 pathologists the inter-observer agreement between pairs was calculated using kappa statistics with linear weights. For linear weights using absolute distances, the average O'Connell and Dobson estimator was 0.67 (95% CI: 0.62 – 0.72), range 0.60 – 0.74.

The mean weighted kappa values of individual observers are shown in Table 3 and Fig. 3. In a pairwise comparison between all observers the inter-observer weighted kappa varied from 0.48 to 0.89. There was no obvious geographic trend of reproducibility as experts from different continents were spread among high and low kappa values.

For cases uploaded as ISUP grade 1, the unweighted kappa was 0.59 (moderate agreement) (95% CI: 0.49 – 0.69); for grade 2, kappa=0.44 (moderate agreement) (95% CI: 0.36 – 0.51); for grade 3, kappa=0.32 (fair agreement) (95% CI: 0.24 – 0.41); for grade 4, kappa=0.47 (moderate agreement) 95% CI: 0.32 – 0.62); and for grade 5, kappa=0.58 (moderate agreement) (95% CI: 0.46 – 0.69). This shows strongest agreement for grades 1 and 5, while grade 3 showed the worst agreement.

The weighted kappas of all cases varied from 0.19 to 1.00. All cases with a weighted kappa of 0.40 or lower were ISUP grade 3 or higher. The four cases with the lowest weighted kappa (0.19 – 0.31) were all grade 5. In eight cases the weighted kappa was 1.00: 5 were grade 1, 1 was grade 2 and 2 were grade 4. The highest concordance was achieved among grade 1 cases, where consensus was reached in 80% of cases (Table 3). ISUP grade 3 was particularly problematic with an unweighted kappa of 0.32 and consensus in only 21.4% of cases.

Discussion

Pathologists have struggled with the reproducibility of prostate cancer grading for more than 40 years.[11, 12] It is not surprising that inter-observer variability is problematic, when complex morphological patterns are estimated subjectively and translated into a discrete ordinal scale. Grading of cancer aims to stratify a biological continuum into a limited number of discrete categories. The borders between these categories remain subjective, but improved definitions have the potential to make grading more robust, for both clinical practice and research.

The discipline of pathology is suitable for studies of observer consistency and not surprisingly, an extensive literature has accumulated on the reproducibility of Gleason grading.[1] The inter-observer variability of this grading system has typically been shown to be in the range of moderate (kappa 0.41 – 0.60) to substantial (0.61 – 0.80) with expert groups achieving higher scores than general pathologists.[2–4, 13] This spread of observer performance is problematic in a clinical sense, as the grading of prostate cancer is one of the most critical parameters for determining appropriate patient management.

In recent years criteria relating to the grading of prostate cancer have undergone major changes and this has resulted in the promotion of further confusion amongst pathologists.[14, 15] There is often a delay in the adoption of new grading criteria internationally and the uptake of revised grading systems may show temporal variation amongst countries. Importantly the interpretation of grading guidelines may also vary, even among experts. In these contexts prostate cancer grading has become somewhat of a moving target.[16]

There is evidence that experienced pathologists demonstrate reasonably good intra-observer reproducibility for prostate cancer grading and a weighted kappa as high as 0.91 has been reported.[17] However, in addition to individual grading performances it is also important to strive for a common grading culture. Some efforts have been made to utilize images for the training of the pathology community in prostate cancer grading, although, prior to this study, a platform for international standardization has not been developed.[12, 13, 18, 19]

The ISUP has recently launched Pathology Imagebase, a novel, web-based tool for promoting standardization of histopathology reporting and the education of pathologists.[5] The credibility of the Imagebase system is based upon the involvement of experts across the world. In this system image libraries are continuously built and characterized through an automatized review process. This mechanism has several unique features: 1) images are reviewed by a large panel of leading experts; 2) the experts represent many different geographic regions, which takes into account regional differences in such important aspects as grading traditions; 3) the review and voting processes are independent as the experts are blinded to the opinions of their peers; 4) the system includes an automatic mechanism for transferral of cases into a public domain once they have fulfilled the consensus criteria; 5) the database is a web-based resource that can be easily accessed from around the world; 6) search functions make it easy to triage the database and 7) the library does not exclude borderline cases that are usually not covered by textbook illustrations. Consensus was set at 2/3 agreement amongst the experts according to the accepted practice in international

guideline development in uropathology. With a higher threshold few cancers would have reached consensus and the project would have failed to set a standard for grading except for the most obvious cases.

A selected series of core needle biopsies of prostate cancers derived from an earlier screening study was used for the initial testing of the Imagebase system. Rather then enrolling consecutive cases from the screening study, biopsies from the series were chosen based upon a spread of morphology. This approach was undertaken as almost 50% of the cases in the series were reported as GS 3+3=6. We reasoned that many of these cases would deliver few diagnostic challenges, resulting in the skewing of the results. Instead we have attempted to provide a wide spread of morphologies and to include cases that illustrate the potentially grey area of potentially overlapping grades. As it transpired this assumption was confirmed by the study results. The overall weighted kappa for agreement among the experts was 0.67, which is within the range of substantial agreement. Despite this level of agreement it is problematic that as many as 44.4% of cases failed to reach consensus among experts and it is our hope that Pathology Imagebase will provide a tool that will promote grading consensus and consistency.

The level of agreement differed for each of the 5 ISUP grades. Classical GS 3+3=6 (ISUP grade 1) cancers had the highest level of consensus. Five of eight cases with perfect agreement (weighted kappa 1.00) were ISUP grade 1. Among uploaded cases with this previously reported grade as many as 80% of cases achieved a consensus. By contrast, the four cases with the lowest weighted kappa values (0.19 – 0.31) were all ISUP grade 5, which itself is most likely a reflection of the greater morphological complexity in higher grade cancers. As we have previously shown in a screening population, as many as 56% of diagnosed cancers may be ISUP grade 1.[6] Because of this sampling bias, the high level of inter-observer agreement associated with this grade indicates that the overall reproducibility in a consecutive series of needle biopsies will most likely be higher than that reported in the present study. However, inclusion of a large series of consecutive cases in this study would not have been practically feasible. We only included leading international experts in the expert panels and they would most likely not have had time to review several hundreds of cases. Thus, there was an over representation of difficult cases of higher grade among the uploaded images. Yet, the number of consensus cases of ISUP grades 1 and 2 was higher than in the other grades. This both reflects how uncommon high-grade cases are in a screening population and our difficulties in reaching consensus in cancers with a more complex morphology.

We identified several problem areas among the non-consensus cases. In cases bordering between GS 3+3=6 and 3+4=7, the difficulty is usually related to whether a small component of seemingly poorly formed or fused glands may be explained by tangential cutting of a gland. It has been shown that the diagnostic reproducibility for poorly formed glands in Gleason pattern 4 tumors is lower than that for cribriform glands and this was confirmed in the present study. In 11 out of 14 borderline cases in the non-consensus group, poorly formed glands were determined to be the predominant cause of diagnostic difficulties. The ISUP 2014 consensus conference suggested that there should be more than occasional structures of this type for a tumor to qualify as Gleason pattern 4, otherwise they

may represent tangential cuts.[15] This is also in line with a previous reproducibility study on the diagnosis of poorly formed glands, indicating that ≤5 seemingly poorly formed glands would be insufficient for a diagnosis of Gleason pattern 4.[20] It is apparent that, despite these attempts to define the features of poorly formed glands, there is a need to better define a detection threshold for such patterns.[21, 22] Cancers that qualified for GS 3+4=7 with a narrow margin (Figure 1C) may serve to define the minimum level required to recognize Gleason pattern 4. The 2014 ISUP Consensus Conference on grading futhermore advised that in cases with borderline morphology between Gleason patterns 3 and 4 and crush artifacts, the lower grade should be favored.[15]

In cases that were considered to be GS 7 by many of the experts some diagnostic disagreement was expressed by a minority of the participants. In particular cases bordering between ISUP grades 2 and 3 were identified as an area of contention. The reason for this is unclear; however, a likely possibility is that it may be difficult to determine the proportions between Gleason patterns 3 and 4, particularly when they are mixed. In such cases the observer must simultaneously classify patterns as well as determine the area of each pattern present in the section.

The second most common problem encountered in borderline cases was difficulty in distinguishing between ISUP grades 3 and 4. According to current grading criteria it is permissible to overlook minor components of lower grade (<5%); however, it is apparent that drawing the line between GS 4+3=7 and 4+4=8 with a minor lower grade component is subjective. This is illustrated by the considerable difficulties encountered by participants in identifying GS 4+3=7 consensus cases, with only 3 out of 14 uploaded cases being agreed upon. Whether the distinction between GS 4+4=8 and GS 4+3=7 with a small component of pattern 3 is clinically important is questionable, as it is difficult to see how the addition of a minute component of Gleason pattern 3 would mitigate the outcome of a case that is mostly composed of Gleason pattern 4. If reporting of % Gleason pattern 4/5 becomes common practice, this information might be sufficient to inform clinicians that the tumor is predominantly high-grade.[23]

A further problem area involved cancers containing a possible component of Gleason pattern 5, where the observer must determine if single cells and solid strands represent tangential cuts of small glands. Such cases often contain a mixture of elements that can be interpreted as Gleason patterns 3, 4 and 5. Thus, the diagnosis may span ISUP grades 2 – 5 depending on whether or not a Gleason pattern 5 component is identified and how the proportions of the remaining components are judged. It has been noted that most urological pathologists require a cluster of strands, nests or single cells seen at less than 40x lens magnification to diagnose Gleason pattern 5, or more than 10 of these if they are not clustered.[24, 25]

This study demonstrates that for prostate cancer grading there is still a need to improve inter-observer agreement, even among experts. It is our hope that pathologists will utilize the Imagebase catalogue on an on-going basis to facilitate grading calibration. This can be achieved in several ways. The self-testing mode with blinded grades may be used for routine quality control, the range of a certain grade may be reviewed using the filtering function or the database may be consulted for comparison of images, when a challenging case is

encountered in routine clinical practice. The Imagebase mechanism used jpg images rather than digitized slides as it was assumed that busy pathologists will have limited time to consult a large image repository. Speed was, therfore, prioritized over detailed information. A follow-up study involving the same experts after they have used the Imagebase instrument for a longer period may be of interest to confirm the development of consensus-based uniformity of practise. Furthermore, it is our intention to investigate genetic alterations of non-consensus cases and compare these against consensus cases with the aim to develop additional methods to classify cancers where experts do not agree on morphological grounds. There is a need of continued studies of the biology of prostate cancer patterns, as this has the potential to refine grading beyond what can be achieved by a subjective expert consensus. However, until such insights are gained we need to agree on the practical application of grading based on current understanding. The Imagebase platform has the potential to support this process and also provide well-documented material for tissue based studies.

To our knowledge this is the first attempt to systematically build a reference image library of grading with cases selected by global experts through independent voting. It is our hope that this methodology may serve as a model that may be employed in other fields of diagnostic histopathology.

Acknowledgement

This study was funded by grants from Stockholm County, The Swedish Cancer Society and The Stockholm Cancer Society.

References

1. Engers R. Reproducibility and reliability of tumor grading in urological neoplasms. World J Urol 2007;25;595–605. [PubMed: 17828603]
2. Allsbrook WC Jr., Mangold KA, Johnson MH et al. Interobserver reproducibility of Gleason grading of prostatic carcinoma: urologic pathologists. Hum Pathol 2001;32;74–80. [PubMed: 11172298]
3. 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;81–88. [PubMed: 11172299]
4. Glaessgen A, Hamberg H, Pihl CG, Sundelin B, Nilsson B, Egevad L. Interobserver reproducibility of modified Gleason score in radical prostatectomy specimens. Virchows Arch 2004;445;17–21. [PubMed: 15156317]
5. Egevad L, Cheville J, Evans AJ et al. Pathology Imagebase - A reference image database for standardization of pathology. Histopathology 2017;in press.
6. Gronberg H, Adolfsson J, Aly M et al. Prostate cancer screening in men aged 50–69 years (STHLM3): a prospective population-based diagnostic study. Lancet Oncol 2015;16;1667–1676. [PubMed: 26563502]
7. O'Connell DL, Dobson AJ. General observer-agreement measures on individual subjects and groups of subjects. Biometrics 1984;40;973–983.
8. Shouten HJA. Measuring pairwise interobserver agreement when all subjects are judged by the same observers. Statistica Neerlandica 1982;36;45–61.
9. Clements M, O'Connell D. magree: Implements the O'Connell-Dobson-Schouten Estimators of Agreement for Multiple Observers. R package version 1.0., 2016;https://CRAN.R-project.org/package=magree.

10. Harding-Jackson N, Kryvenko ON, Whittington EE et al. Outcome of Gleason 3 + 5 = 8 Prostate Cancer Diagnosed on Needle Biopsy: Prognostic Comparison with Gleason 4 + 4 = 8. J Urol 2016;196;1076–1081. [PubMed: 27265220]
11. Bain GO, Koch M, Hanson J. Feasibility of grading prostate carcinomas. Arch Pathol Lab Med 1982;106;265–267.
12. Egevad L. Reproducibility of Gleason grading of prostate cancer can be improved by the use of reference images. Urology 2001;57;291–295. [PubMed: 11182339]
13. Egevad L, Ahmad AS, Algaba F et al. Standardization of Gleason grading among 337 European pathologists. Histopathology 2013;62;247–256. [PubMed: 23240715]
14. Epstein JI, Allsbrook WC Jr., Amin MB, Egevad L. The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma. Am J Surg Pathol 2005;29;1228–1242. [PubMed: 16096414]
15. Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA. 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;244–252. [PubMed: 26492179]
16. Danneman D, Drevin L, Robinson D, Stattin P, Egevad L. Gleason inflation 1998–2011: a registry study of 97,168 men. BJU Int 2015;115;248–255. [PubMed: 24552193]
17. 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;2006–2010. [PubMed: 12394696]
18. Egevad L, Algaba F, Berney DM et al. Interactive digital slides with heat maps: a novel method to improve the reproducibility of Gleason grading. Virchows Arch 2011;459;175–182. [PubMed: 21698392]
19. 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;1818–1823. [PubMed: 11042578]
20. Zhou M, Li J, Cheng L et al. Diagnosis of "Poorly Formed Glands" Gleason Pattern 4 Prostatic Adenocarcinoma on Needle Biopsy: An Interobserver Reproducibility Study Among Urologic Pathologists With Recommendations. Am J Surg Pathol 2015;39;1331–1339. [PubMed: 26099009]
21. Dong F, Wang C, Farris AB et al. Impact on the clinical outcome of prostate cancer by the 2005 international society of urological pathology modified Gleason grading system. Am J Surg Pathol 2012;36;838–843. [PubMed: 22592143]
22. Kweldam CF, Nieboer D, Algaba F et al. Gleason grade 4 prostate adenocarcinoma patterns: an interobserver agreement study among genitourinary pathologists. Histopathology 2016;69;441–449. [PubMed: 27028587]
23. Sauter G, Steurer S, Clauditz TS et al. Clinical Utility of Quantitative Gleason Grading in Prostate Biopsies and Prostatectomy Specimens. Eur Urol 2016;69;592–598. [PubMed: 26542947]
24. Egevad L, Allsbrook WC Jr., Epstein JI. Current practice of Gleason grading among genitourinary pathologists. Hum Pathol 2005;36;5–9. [PubMed: 15712175]
25. 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;1242–1249. [PubMed: 25929349]

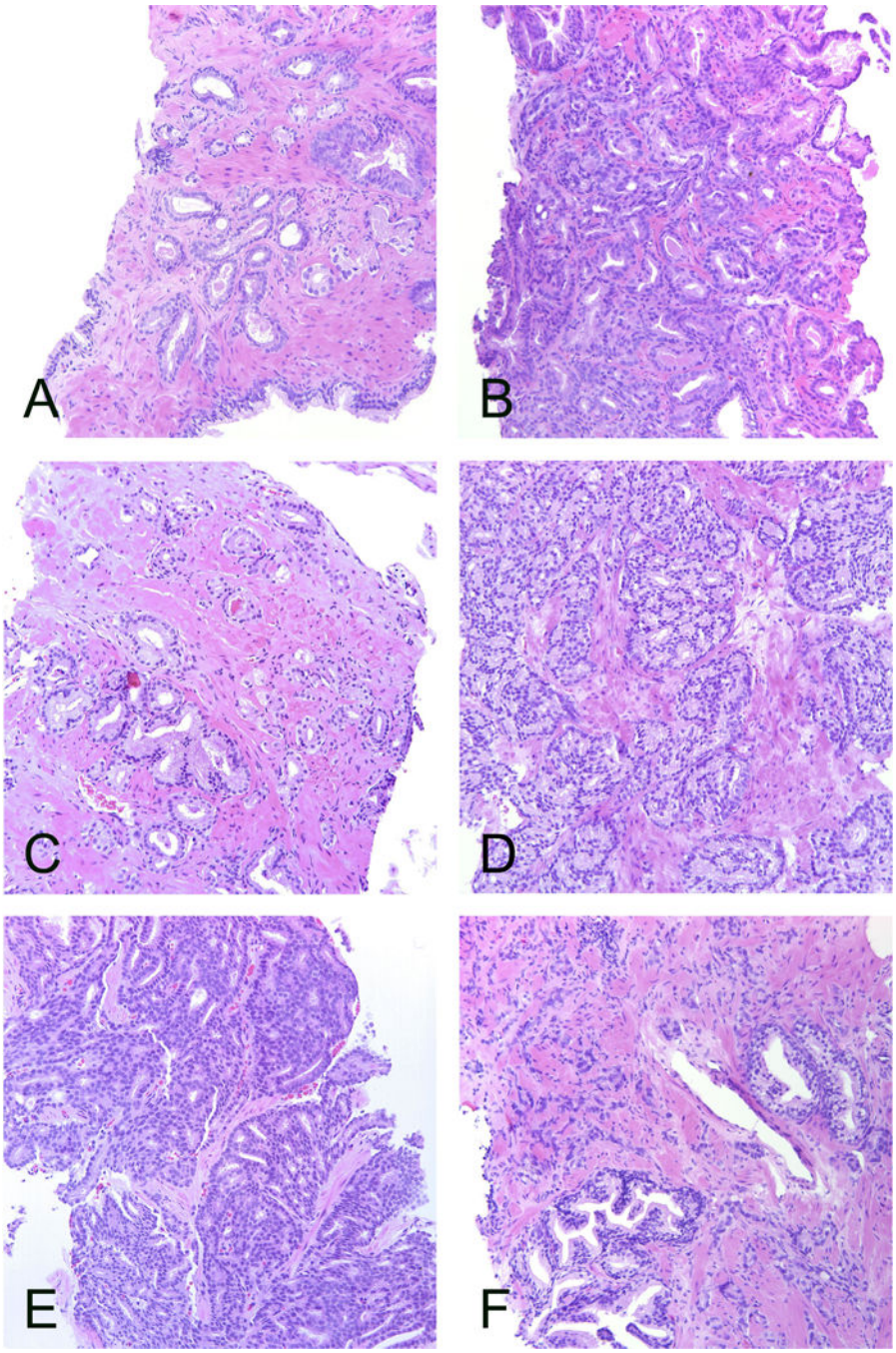

Figure 1.

Cases that reached grading consensus. **A.** Consensus ISUP grade 1 supported by 100%. Well-formed infiltrating glands. **B.** Consensus ISUP grade 2 supported by 83%. Glandular fusion. **C.** Consensus ISUP grade 2 supported by 67%. Poorly formed glands. This case illustrates the minimum number of poorly formed glands required for ISUP grade 2. **D.** Consensus ISUP grade 3 supported by 96%. Large cribriform sheets and fused glands but also some individual glands. **E.** Consensus ISUP grade 4 supported by 100%. Cribriform pattern. This pattern is easily identified. **F.** Consensus ISUP grade 5 supported by 87%.

Strands of epithelium and single cells scattered in the stroma. All microphotographs haematoxylin and eosin, 20x lens magnification.

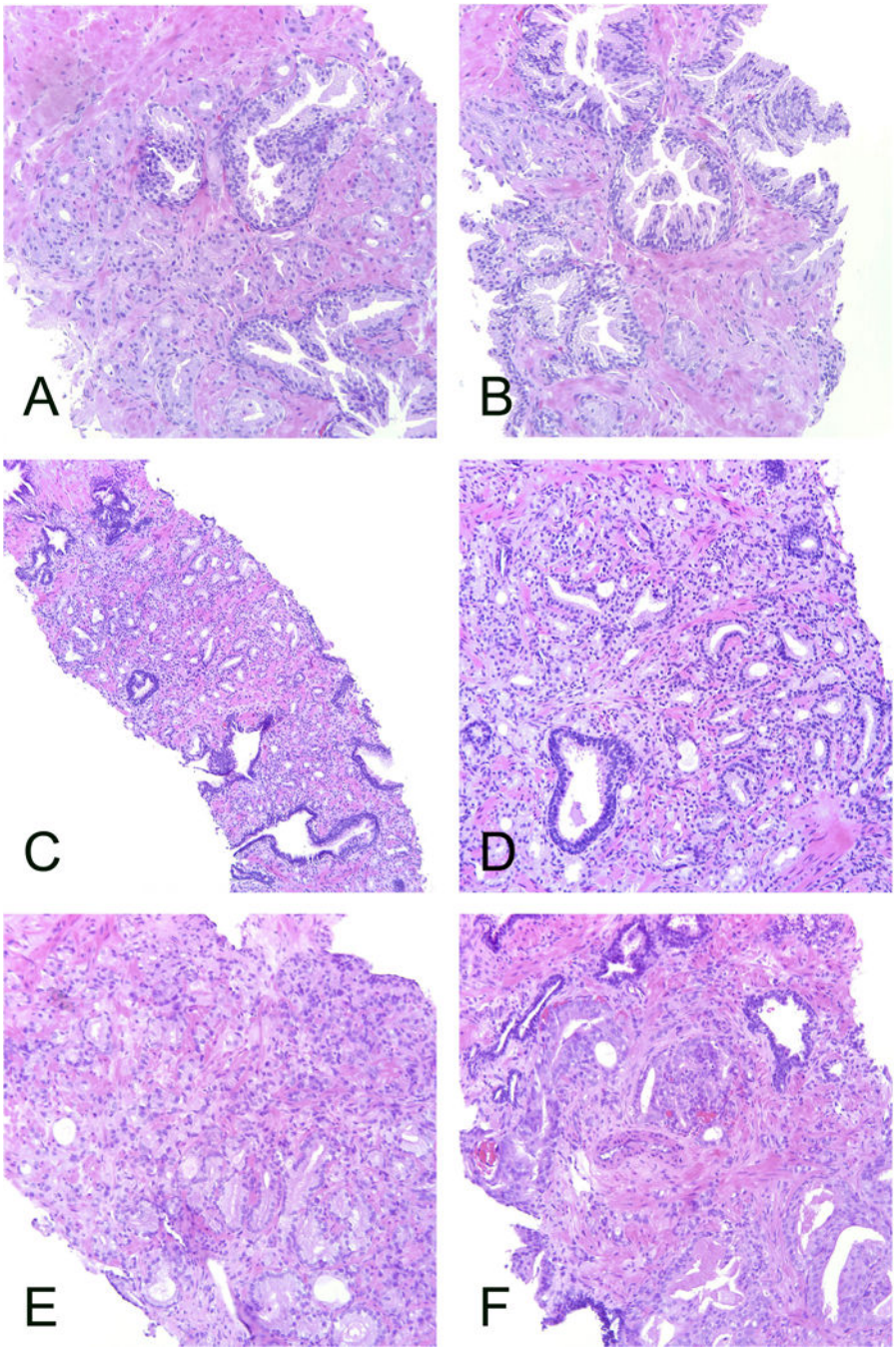

Figure 2.

Cases that did not reach grading consensus. **A-B.** The distribution of ISUP grades was 1 – 3 in 33%, 63% and 4%, respectively. This case would have reached the 2/3 consensus level if there had not been a vote for grade 4. **C-D.** The distribution of ISUP grades was 2 – 4 in 42%, 50%, 8%, respectively. All acknowledged the presence of poorly formed or fused glands but disagreement as to which component predominate. The patterns are mixed throughout the tumor which makes it more difficult to estimate the proportions. **E-F.** The distribution of ISUP grades was 2 – 5 in 4%, 42%, 8% and 46%, respectively. A wide scatter

of the grades that illustrates the problem when there are both borderline morphology between patterns and a mixture of several patterns. All microphotographs haematoxylin and eosin. A-B, D-F: 20x lens magnification, C: 10x.

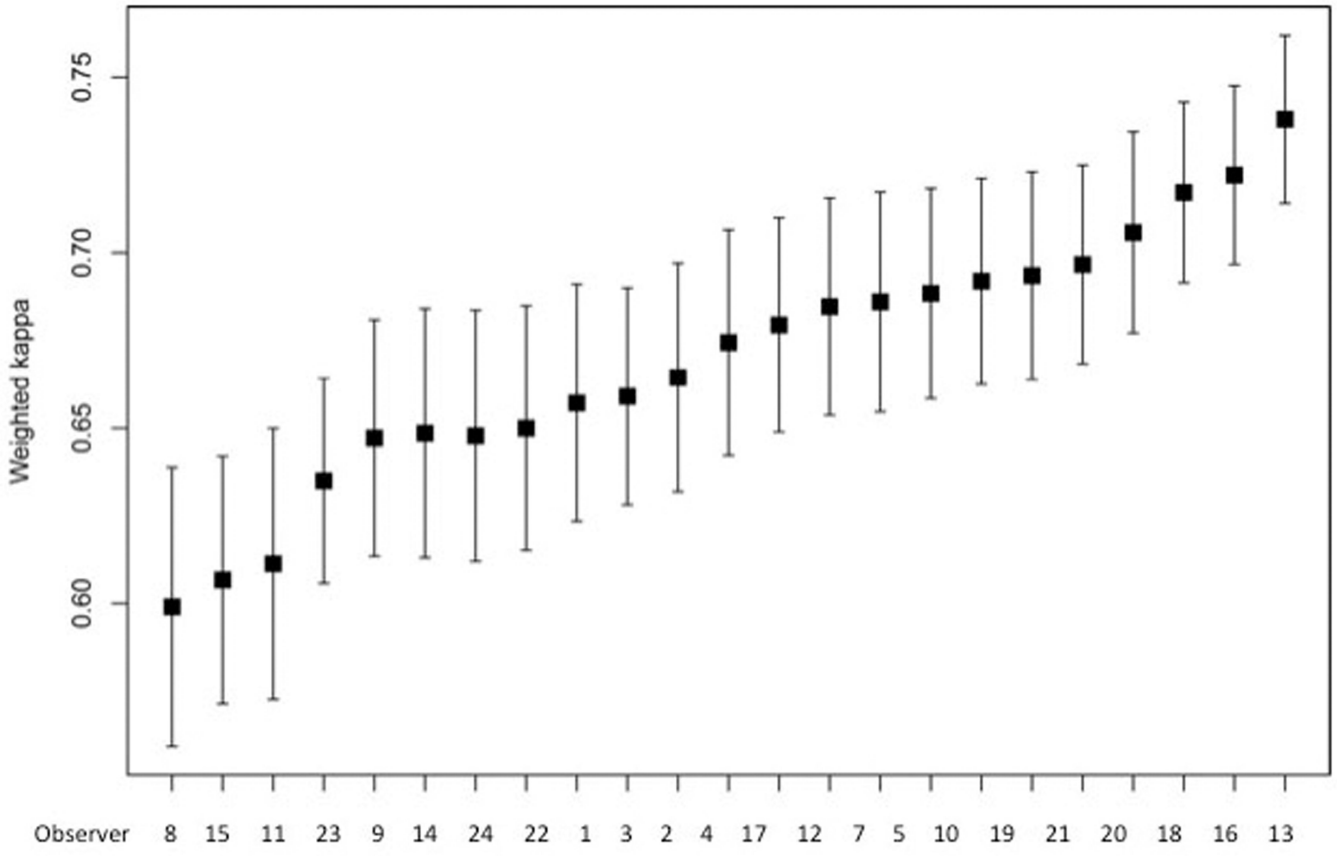

Figure 3.

Mean weighted kappas for ISUP grades 1 – 5 of 23 observers with complete data submission. 95% confidence intervals.

Table 1.

Members of the prostate panel. Panel leader is marked with *.

Prostate panel	City, state	Country
Lars Egevad *	Stockholm	Sweden
Dan Berney	London	United Kingdom
David Bostwick	Orlando, FL	Canada
Eva Comperat	Paris	France
Brett Delahunt	Wellington	New Zealand
Andrew Evans	Toronto, ON	Canada
Samson Fine	New York, NY	Canada
David Grignon	Indianapolis, IN	USA
Peter Humphrey	New Haven, CT	USA
Kenneth Iczkowski	Milwaukee, WI	USA
James Kench	Sydney	Australia
Glen Kristiansen	Bonn	Germany
Katia Leite	Sao Paolo	Brazil
Cristina Magi-Galluzzi	Cleveland, OH	USA
Jesse McKenney	Cleveland, OH	USA
Jon Oxley	Bristol	United Kingdom
Chin-Chen Pan	Taipei	Taiwan
Hemamali Samaratunga	Brisbane	Australia
John Srigley	Hamilton, ON	Canada
Hiroyuki Takahashi	Tokyo	Japan
Toyonori Tsuzuki	Nagoya	Japan
Theo van der Kwast	Toronto, ON	Canada
Murali Varma	Cardiff	United Kingdom
Ming Zhou	Dallas, TX	USA

Table 2.

Distributions of initial grades and grades reported in all responses.

Grade	Initial grade % (n)	All responses % (n)
GS 3+3=6 (ISUP 1)	22.2% (20)	25.2% (536)
GS 3+4=7 (ISUP 2)	35.6% (32)	34.1% (727)
GS 4+3=7 (ISUP 3)	15.6% (14)	17.3% (368)
GS 4+4=8 (ISUP 4)	12.2% (11)	13.5% (288)
GS 9–10 (ISUP 5)	14.4% (13)	9.9% (211)
Other	0% (0)	0.3% (7)
	100% (90)	100% (2137)

Table 3.

Reproducibility by proposed ISUP grade among all cases.

Proposed grade	Consensus n	Total n	% consensus Proposed grade	% consensus Any grade
1	16*	20	80.0	85.0
2	17	32	53.1	53.1
3	3**	14	21.4	35.7
4	5	11	45.5	45.5
5	6	13	46.2	46.2

*An additional case reached consensus for ISUP grade 2.

**Two additional cases reached consensus for ISUP grade 2.

Table 4.

Number and % of consensus cases that were assigned an ISUP grade 1 – 5.

	Assigned grades % (mean n)				
Consensus grade	**1**	**2**	**3**	**4**	**5**
1	86.1% (19.8)	13.3% (3.1)	0.5% (0.1)	0% (0.0)	0% (0.0)
2	9.8% (2.3)	80.4% (18.5)	8.9% (2.1)	0.9% (0.2)	0% (0.0)
3	0% (0.0)	11.6% (2.7)	79.7% (18.3)	7.2% (1.7)	1.4% (0.3)
4	0% (0.0)	0% (0.0)	7.0% (1.6)	92.2% (21.2)	0.9% (0.2)
5	0% (0.0)	0% (0.0)	5.1% (1.2)	14.5% (3.3)	80.4% (18.5)

Table 5.

Number and % of non-consensus cases that were assigned an ISUP grade 1 – 5.

	Assigned grades % (mean n)				
Proposed grade	**1**	**2**	**3**	**4**	**5**
1	54.2% (13.0)	45.8% (11.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
2	33.6% (8.1)	53.1% (12.7)	11.4% (2.7)	1.9% (0.5)	0.0% (0.0)
3	0.9% (0.2)	27.3% (6.6)	50.5% (12.1)	19.0% (4.6)	2.3% (0.6)
4	0.0% (0.0)	3.5% (0.8)	37.5% (9.0)	44.4% (10.7)	14.6% (3.5)
5	0.0% (0.0)	2.4% (0.6)	29.2% (7.0)	26.8% (6.4)	41.7% (10.0)

Table 6.

Number of consensus cases (n = 50) that were assigned a grade different from the consensus grade, over-graded and under-graded by the observers. Weighted kappa statistics and confidence intervals based on all cases (n = 90). Weighted kappa is missing for observer 6 because of incomplete data.

Observer	Discordant with consensus	Over-grading	Under-grading	Weighted kappa	C.I.
8	11	3	8	0.599	(0.559 – 0.639)
15	17	15	2	0.607	(0.571 – 0.642)
11	14	14	0	0.611	(0.573 – 0.650)
23	17	7	10	0.635	(0.606 – 0.664)
9	10	9	1	0.647	(0.613 – 0.681)
14	11	2	9	0.648	(0.612 – 0.684)
24	10	6	4	0.649	(0.613 – 0.684)
22	9	0	9	0.650	(0.615 – 0.685)
1	7	5	2	0.657	(0.623 – 0.691)
3	10	1	9	0.659	(0.628 – 0.690)
2	9	6	3	0.664	(0.632 – 0.697)
4	9	3	6	0.674	(0.642 – 0.707)
17	7	3	4	0.679	(0.649 – 0.710)
12	5	2	3	0.685	(0.654 – 0.716)
7	7	2	5	0.686	(0.655 – 0.717)
5	8	8	0	0.688	(0.659 – 0.718)
10	8	6	2	0.692	(0.663 – 0.721)
19	6	1	5	0.693	(0.664 – 0.723)
21	2	0	2	0.697	(0.668 – 0.725)
20	5	5	0	0.706	(0.677 – 0.734)
18	4	1	3	0.717	(0.691 – 0.743)
16	3	2	1	0.722	(0.697 – 0.748)
13	2	2	0	0.738	(0.714 – 0.762)
6	5	0	5	NA	NA
Mean	**8.2**	**4.3**	**3.9**	**0.670**	**(0.619 – 0.718)**