available at www.sciencedirect.com
journal homepage: www.europeanurology.com/eufocus

Mini Review – Prostate Cancer

Artificial Intelligence for Diagnosis and Gleason Grading of Prostate Cancer in Biopsies—Current Status and Next Steps

Kimmo Kartasalo [a,b], *Wouter Bulten* [c], *Brett Delahunt* [d], *Po-Hsuan Cameron Chen* [e], *Hans Pinckaers* [c], *Henrik Olsson* [a], *Xiaoyi Ji* [a], *Nita Mulliqi* [a], *Hemamali Samaratunga* [f], *Toyonori Tsuzuki* [g], *Johan Lindberg* [a], *Mattias Rantalainen* [a], *Carolina Wählby* [h,i], *Geert Litjens* [c], *Pekka Ruusuvuori* [b,j], *Lars Egevad* [k], *Martin Eklund* [a,*]

[a] *Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden;* [b] *Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland;* [c] *Department of Pathology, Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, The Netherlands;* [d] *Department of Pathology and Molecular Medicine, Wellington School of Medicine and Health Sciences, University of Otago, Wellington, New Zealand;* [e] *Google Health, Palo Alto, CA, USA;* [f] *Aquesta Uropathology and University of Queensland, Brisbane, QLD, Australia;* [g] *Department of Surgical Pathology, School of Medicine, Aichi Medical University, Nagakute, Japan;* [h] *Centre for Image Analysis, Department of Information Technology, Uppsala University, Uppsala, Sweden;* [i] *BioImage Informatics Facility of SciLifeLab, Uppsala, Sweden;* [j] *Institute of Biomedicine, Cancer Research Unit and FICAN West Cancer Centre, University of Turku and Turku University Hospital, Turku, Finland;* [k] *Department of Oncology and Pathology, Karolinska Institutet, Stockholm, Sweden*

Article info

Associate Editor: Derya Tilki

Keywords:
Prostate cancer
Gleason grading
Artificial intelligence
Histopathology
Uropathology

Abstract

Diagnosis and Gleason grading of prostate cancer in biopsies are critical for the clinical management of men with prostate cancer. Despite this, the high grading variability among pathologists leads to the potential for under- and overtreatment. Artificial intelligence (AI) systems have shown promise in assisting pathologists to perform Gleason grading, which could help address this problem. In this mini-review, we highlight studies reporting on the development of AI systems for cancer detection and Gleason grading, and discuss the progress needed for widespread clinical implementation, as well as anticipated future developments.

Patient summary: This mini-review summarizes the evidence relating to the validation of artificial intelligence (AI)-assisted cancer detection and Gleason grading of prostate cancer in biopsies, and highlights the remaining steps required prior to its widespread clinical implementation. We found that, although there is strong evidence to show that AI is able to perform Gleason grading on par with experienced uropathologists, more work is needed to ensure the accuracy of results from AI systems in diverse settings across different patient populations, digitization platforms, and pathology laboratories.

* Corresponding author. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Nobels väg 12, 171 77 Stockholm, Sweden.
E-mail address: martin.eklund@ki.se (M. Eklund).

https://doi.org/10.1016/j.euf.2021.07.002

Diagnosis and Gleason grading of prostate cancer in biopsies are critical to the clinical management of men suspected to have prostate cancer; however, the high level of variability in Gleason grading between pathologists poses clinical challenges. This can lead to both under- and overtreatment, which impacts patient morbidity, mortality, and healthcare costs. The advent of digital pathology permits the development of artificial intelligence (AI) systems for assisting pathologists in the evaluation of prostate biopsies. Such AI systems have shown promise with respect to cancer detection and Gleason grading, and are a potential solution to the problem of high interpathologist variability. In this mini-review, we summarize the current status for AI-based diagnosis and Gleason grading of prostate cancer in biopsies and discuss the potential for future developments.

The first attempt to use techniques based on deep neural networks for the detection of cancer on prostate biopsies was reported by Litjens and colleagues [1], who showed early promising results in a study limited by the size of the data. Three years later, in 2019 Campanella and colleagues showed an area under the receiver operating characteristics curve (AUC) of 0.991 for cancer detection on an independent test dataset and 0.943 on external validation data [2]. The study by Campanella et al [2] used a large dataset (n = 12 132 slides) to train the deep learning model using only the reported diagnoses as labels for training, thus circumnavigating the labor-intensive collection of detailed pixel-wise manual annotations. In a recent external validation study of the algorithm developed in Campanella et al [2], Perincheri and colleagues [3] showed that the algorithm achieved 97.7% sensitivity at a specificity level of 99.3%.

Early attempts at applying machine learning and AI to not only detect cancer in prostate biopsies, but also perform Gleason grading showed promise, but were constrained by the limited availability of training data [4□6]. In 2020, three studies independently demonstrated that AI systems can perform both cancer detection and Gleason grading with performance on par with expert uropathologists (Table 1) [7□9]. Specifically, Ström et al [7] used 6953 biopsies mainly from the STHLM3 population□based prostate cancer screening trial to train an AI system that demonstrated a mean pairwise linearly weighted Cohen□s kappa statistic of 0.62. This was within the range of the corresponding values of 23 experienced urological pathologists from the International Society of Urological Pathology□s Imagebase grading reference panel (0.60□0.73). In a second study, Bulten et al [8] showed that a deep learning system trained on 4712 biopsies had higher agreement (quadratic kappa 0.854), with a consensus reference standard defined by three expert urological pathologists, than that of a panel of international pathologists (median kappa 0.819), outperforming ten of the 15 study pathologists [5]. Finally, Nagpal et al [9] demonstrated that an AI system trained on a combination of prostatectomy samples (n = 1226) and biopsies (n = 524) had significantly higher agreement with the diagnosis of subspecialist pathologists than that achieved by general pathologists (71.7% vs 58.0%). When two of these AI systems were used by pathologists in research settings, AI-assisted pathologists were able to achieve higher agreements with panels of urological pathologists than in an unassisted setting [10,11]. Further evidence for AI-based Gleason grading was provided by Pantanowitz et al [12], who demonstrated an AUC of 0.941 for discriminating between Gleason score 6 and Gleason score □7 on external validation data, as well as by Mun et al [13], who recently demonstrated moderate agreement between AI Gleason grading and pathologists on internal and external test data for an AI system trained on 6664 biopsy cores. All these studies also showed accurate results for cancer detection, with AUCs ranging from 0.943 to 0.99 on external validation data (Table 1) [7□9,12,13].

Although the results for Gleason grading are promising, it should be noted that the AI algorithms were validated by the same research groups that developed them, using limited external data. It is clear that, before AI systems for Gleason grading can be considered suitable for introduction into routine clinical practice, there is a need for further validation utilizing independent datasets from diverse and international populations. Development of AI systems that generalize across varying data sources represents one of the central barriers to the clinical adoption of AI algorithms. The questions that arise are as follows: How do we address the current limitations? What is needed for this technology to achieve clinical implementation and bring value to patients?

Fundamental to these questions are the limits of the capabilities of AI systems. It is well known that deep learning systems are susceptible to changes in input data [14]. Biopsies sampled from different patient populations, processing of tissues in different laboratories, and images scanned on different hardware platforms may impact AI performance. Similarly, further development and testing are required to permit the recognition of benign mimics of cancer and deceptively bland morphological variants of prostatic adenocarcinoma. The collection of larger and more diverse datasets for training AI systems will undoubtedly result in improved capacity. Recently, we organized the Prostate cANcer graDe Assessment (PANDA) competition for the development of AI algorithms for Gleason grading of prostate biopsies. By pooling data sources from the studies of Ström et al [7], Bulten et al [8], and Nagpal et al [9], we could demonstrate that top performing algorithms generalize across multisite, international populations and reference standards with high accuracy [15]. The benefits of large and diverse datasets for training and validation are clear, and it is therefore encouraging that initiatives similar to the large genome-wide association studies consortia are emerging also for histopathology images [16,17]. Ultimately, AI systems for the pathological assessment of prostate biopsies need to demonstrate their value in prospective and well-designed clinical trials, as well as in controlled implementation studies.

Irrespective of the quality of AI systems for Gleason grading, there is always a risk that unusual cases cannot reliably be assigned a grade. We therefore believe that a key component for the clinical implementation of AI algorithms is the development of anomaly detection systems that can quantify the confidence in predictions and serve as quality

Table 1 – Publications describing algorithms for detection and grading of prostate cancer in core needle biopsies[a]

Article	Dataset size for training and tuning	Dataset size for internal testing	Dataset size for external validation	Outcomes	Summary of main findings
Campanella et al (2019) [2]	10 348 slides from 711 cases	1784 slides from 125 cases	12 727 slides from 6323 cases	Cancer detection	AUC of 0.991 for discriminating between slides with cancer present and slides with only benign tissue on internal test data; AUC of 0.932 on external validation sets. Clinical application would allow pathologists to exclude 65□75% of slides while retaining 100% sensitivity
Ström et al (2020) [7]	6953 biopsy cores from 1069 cases	1718 biopsy cores from 333 cases	330 biopsy cores from 73 cases	Cancer detection, Gleason score, cancer length (mm)	AUC of 0.997 for discriminating between biopsy cores with cancer present and cores with only benign tissue on internal test data; AUC of 0.986 on external validation sets. Gleason grading and estimates of cancer length (mm) on internal test data demonstrated to be on par with uropathologists
Bulten et al (2020) [8]	5209 biopsy cores from 1033 cases	550 biopsy cores from 210 cases	245 tissue microarrays from 245 cases	Cancer detection, Gleason score	AUC of 0.99 for discriminating between biopsy cores with cancer present and cores with only benign tissue on internal test data; AUC of 0.98□0.99 on external validation sets. Gleason grading on internal test data demonstrated to have higher concordance with uropathologist consensus reference standard than with general pathologists
Nagpal et al (2020) [9]	1557 slides from prostatectomies of 758 cases plus 524 biopsy cores form 360 cases	430 biopsy cores from 430 cases	331 biopsy cores from 331 cases	Cancer detection, Gleason score	Accuracy of 94.3% for distinguishing biopsy cores containing cancer from those without. AI Gleason grading showed significantly higher agreement with expert uropathologists than with general pathologists
Pantanowitz et al (2020) [12]	549 slides	2501 slides from 213 cases	1627 slides from 100 cases organized into 389 parts [b]	Cancer detection, Gleason score 6 vs □7; detection of atypical small acinar proliferation (ASAP) and perineural invasion (PNI)	AUC of 0.997 for discriminating between slides with cancer present and slides with only benign tissue on internal test data; AUC of 0.991 on external validation sets for discrimination between parts with or without cancer. AUC of 0.941 for discriminating between parts with Gleason score 6 or ASAP and Gleason score □7. AUC of 0.957 for discriminating between parts with cancer with PNI and those without
Perincheri et al (2021) [3]	NA [b]	NA [c]	1876 biopsy cores from 118 cases	Cancer detection, atypia, high-grade prostatic intraepithelial neoplasia	Sensitivity of 97.7% and positive predictive value of 97.9%, and specificity of 99.3% and a negative predictive value of 99.2% in identifying core biopsies with cancer
Mun et al (2021) [13]	6664 biopsy cores from 689 cases	936 biopsy cores from 99 cases	244 tissue microarrays from 244 cases	Cancer detection, Gleason score	AUC of 0.983 for discriminating between biopsy cores with cancer present and cores with only benign tissue on internal test data; AUC of 0.943 on external validation sets. Moderate agreement between AI Gleason grading and pathologists on internal and external test data

AI = artificial intelligence; AUC = area under the receiver operating characteristics curve; NA = not available.
[a] The table summarized large studies with reporting of results on external validation data.
[b] A □part□ is one of three biopsy regions (upper, mid, or base) in one of the prostate lobes.
[c] Perincheri et al [3] describe the results of an external validation of the AI algorithm developed by Campanella et al [2] and thus do not concern the training of the AI model.

control. Such systems should be able to detect occasions when an AI algorithm is presented with data that are outside its applicability limits, necessitating human intervention. This would ensure that only valid input data are used and that only valid predictions are made. So far, little work relating to this has been published, although methods for constructing quality control systems, such as conformal prediction and multihead convolutional neural networks, exist [18,19].

Beyond reaching a mature enough state for clinical implementation, what developments can we expect to see within this field in the coming years? Ultimately, the goal of AI systems is not to reproduce assessments of pathologists, but to improve upon them to provide more accurate prognostication. One way to achieve this is to directly train AI algorithms against long-term follow-up data, such as time to development of metastases or time to death in large cohort studies. There are, however, major challenges in achieving this. A major issue is the size of the training data: millions of data points can be easily obtained for training AI systems to perform Gleason grading by using all ten to 12 biopsy cores obtained from a single patient and by extracting thousands of smaller images□patches□from each biopsy image. In contrast, a man develops metastases

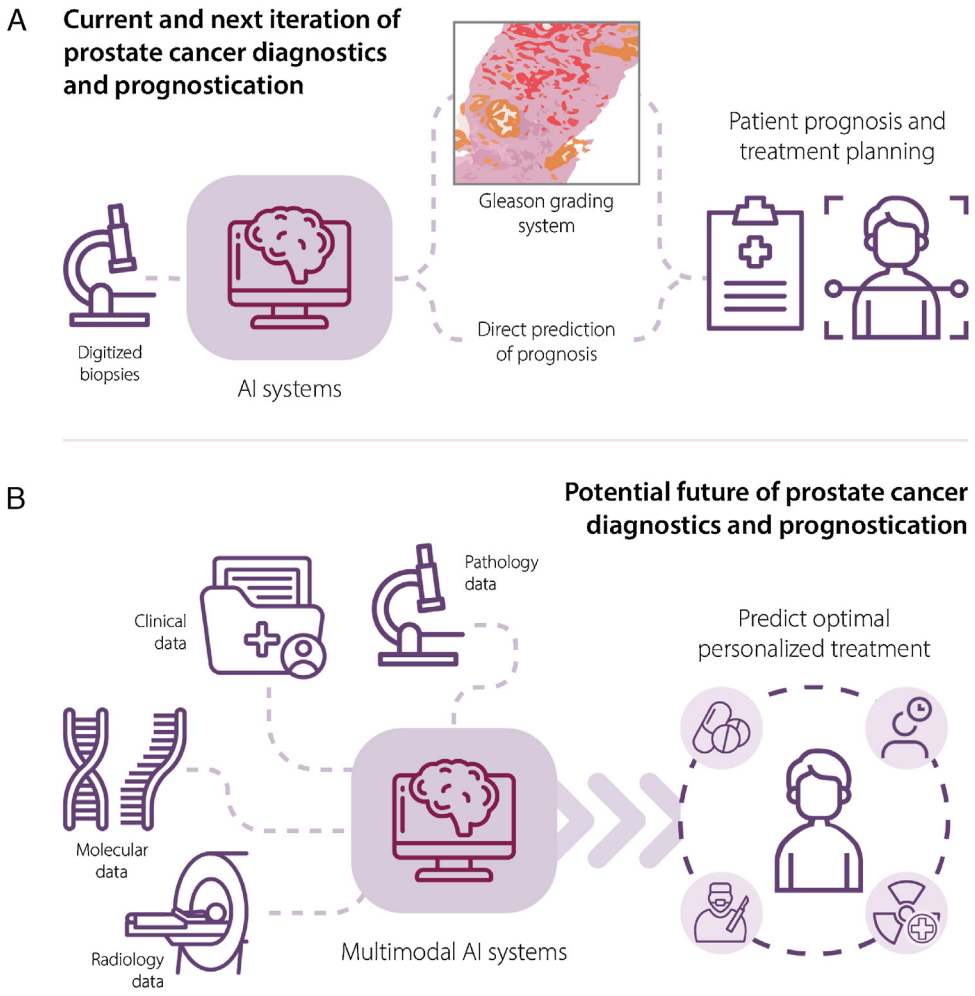

Fig. 1 – Future AI-assisted prostate cancer diagnosis and prognosis. (A) Current AI systems have shown that they can perform Gleason grading on par with expert uropathologists and that they can assist pathologists to achieve higher agreement with consensus grading. Current development also indicates that AI systems may be trained using long-term outcomes to achieve higher agreement with prognosis. (B) AI has the potential to leverage and integrate the increasingly complex data collected during the prostate cancer diagnostic workup process, in order to improve prognostication and treatment selection. Data from initial risk stratification based on modern prediction models can be integrated with magnetic resonance imaging scans and histological imaging data within AI systems. In addition, tumor profiling using sequencing of tissue or circulating tumor DNA can be integrated to predict prognosis and optimize treatment. AI = artificial intelligence.

or dies only once, meaning that there is only a single event for each man that can be used for model training. The limitations of retrospective study design also present challenges, both with respect to historical cohorts not being representative in a contemporary treatment setting and for model evaluation. In historical cases, the original treatment choice was based on the Gleason score assigned by the pathologist. The statistical consequence of this is that the strength of the association between the pathologists□ Gleason score and longer-term outcomes can be severely biased toward the null. In fact, even if Gleason grading by pathologists had a stronger association with longer-term outcomes than the AI model, this bias could cause the results to be reversed [20,21]. Despite these challenges, development of AI systems to improve prognostication is clearly a path that we should explore, as its value to patients could be substantial. This also opens the potential for developing AI systems that combine histopathology information with other data sources such as information from magnetic resonance imaging, genomic information, and biomarkers in multimodal AI systems. This is an exciting vision for harnessing the increasingly data-rich prostate cancer diagnostic pipeline to better predict prognosis and optimize treatment (Fig. 1). Further useful advances would be development of AI algorithms for cancer detection in frozen section analysis. Such systems could assist with the interpretation of surgical margins from the prostate specimen after radical prostatectomy, to improve surgical outcomes.

Author contributions: Martin Eklund had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study concept and design: Kartasalo, Bulten, Pinckaers, Olsson, Mulliqi, Ji, Eklund.

Acquisition of data: Delahunt, Chen, Samaratunga, Tsuzuki, Egevad, Wählby, Ruusuvuori, Lindberg, Rantalainen, Litjens, Eklund.
Analysis and interpretation of data: Kartasalo, Bulten, Pinckaers, Olsson, Mulliqi, Ji, Egevad, Eklund.
Drafting of the manuscript: Kartasalo, Eklund.
Critical revision of the manuscript for important intellectual content: All authors.
Statistical analysis: Kartasalo, Bulten, Eklund.
Obtaining funding: Eklund, Rantalainen, Lindberg, Egevad, Litjens, Ruusuvuori, Wählby.
Administrative, technical, or material support: Kartasalo, Mulliqi, Olsson, Ji.
Supervision: Eklund, Egevad.
Other: None.

Financial disclosures: Martin Eklund certifies that all conflicts of interest, including specific financial interests and relationships and affiliations relevant to the subject matter or materials discussed in the manuscript (eg, employment/affiliation, grants or funding, consultancies, honoraria, stock ownership or options, expert testimony, royalties, or patents filed, received, or pending), are the following: None. P.-H.C.C. is an employee of Google LLC and owns Alphabet stock.

Funding/Support and role of the sponsor: This work was supported by the Swedish Research Council (2019-01466,2020-00692 and 2018-03056), the Swedish Cancer Society (CAN 2018/741 and 200906PjF01H), Academy of Finland grants #341967, #334782 and #335976, and Cancer Foundation Finland. The funders have no role in the collection, analysis, interpretation, manuscript writing, or decision to submit the manuscript.

References

[1] Litjens G, Sánchez CI, Timofeeva N, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep 2016;6:26286.
[2] Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019;25:1301–9.
[3] Perincheri S, Levi AW, Celli R, et al. An independent assessment of an artificial intelligence system for prostate cancer detection shows strong diagnostic accuracy. Mod Pathol 2021;34:1588–95. http://dx.doi.org/10.1038/s41379-021-00794-x.
[4] Gummeson A, Arvidsson I, Ohlsson M, et al. Automatic Gleason grading of H&E stained microscopic prostate images using deep convolutional neural networksMedical imaging 2017: digital pathology. SPIE; 2017. p. 101400S.
[5] Jafari-Khouzani K, Soltanian-Zadeh H. Multiwavelet grading of pathological images of prostate. IEEE Trans Biomed Eng 2003;50:697–704.
[6] Källén H, Molin J, Heyden A, et al. Towards grading Gleason score using generically trained deep convolutional neural networks. 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI) 2016;1163–7.
[7] Ström P, Kartasalo K, Olsson H, et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic ssue>222232.
[8] Bulten W, Pinckaers H, van Boven H, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. Lancet Oncol 2020;21:233–41.
[9] Nagpal K, Foote D, Tan F, et al. Development and validation of a deep learning algorithm for Gleason grading of prostate cancer from biopsy specimens. JAMA Oncol 2020;6:1372–80.
[10] Bulten W, Balkenhol M, Belinga J-JA, et al. Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. Mod Pathol 2021;34:660–71.
[11] Steiner DF, Nagpal K, Sayres R, et al. Evaluation of the use of combined artificial intelligence and pathologist assessment to review and grade prostate biopsies. JAMA Netw Open 2020;3: e2023267.
[12] Pantanowitz L, Quiroga-Garza GM, Bien L, et al. An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. Lancet Digit Health 2020;2, e407□16.
[13] Mun Y, Paik I, Shin S-J, Kwak TY, Chang H. Yet Another Automated Gleason Grading System (YAAGGS) by weakly supervised deep learning. NPJ Digit Med 2021;4:99.
[14] Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. Science 2019;363:1287–9.
[15] Grand Challenge. The PANDA challenge. grand-challenge.org, https://panda.grand-challenge.org/.
[16] Moulin P, Grünberg K, Barale-Thomas E, der Laak JV. IMI-Bigpicture: a central repository for digital pathology. Toxicol Pathol 2021;49:711–3.
[17] Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 2013;26:1045–57.
[18] Messoudi S, Rousseau S, Destercke S. Deep conformal prediction for robust models. Inform Process Manage Uncertainty Knowl Based Syst 2020;1237:528–40.
[19] Linmans J, Laak J, Litjens G. Efficient out-of-distribution detection in digital pathology using multi-head convolutional neural networks. In: Medical imaging with deep learning. PMLR. p. 465□78.
[20] Eklund M, Kartasalo K, Olsson H, Ström P. The importance of study design in the application of artificial intelligence methods in medicine. NPJ Digit Med 2019;2:101.
[21] Nagpal K, Liu Y, Chen P-HC, Stumpe MC, Mermel CH. Reply: □The importance of study design in the application of artificial intelligence methods in medicine□. NPJ Digit Med 2019;2:100.