Pathology International

WILEY

Pathology International

REVIEW ARTICLE OPEN ACCESS

The Role of Artificial Intelligence in the Evaluation of Prostate Pathology

Lars Egevad[1] | Andrea Camilloni[2] | Brett Delahunt[1,3] | Hemamali Samaratunga[4] | Martin Eklund[2] | Kimmo Kartasalo[5]

[1]Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden | [2]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden | [3]Malaghan Institute of Medical Research, Wellington, New Zealand | [4]Aquesta Pathology and University of Queensland School of Medicine, Brisbane, Queensland, Australia | [5]SciLifeLab, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

Correspondence: Lars Egevad (lars.egevad@ki.se)

Received: 11 December 2024 | **Revised:** 31 January 2025 | **Accepted:** 7 April 2025

Funding: L.E. was funded by grants from The Swedish Cancer Foundation (Grant No. CAN 20 1358 PjFs and 23 2641Pj) and The Stockholm Cancer Society (Grant No. 204043). K.K. received funding from the SciLifeLab & Wallenberg Data Driven Life Science Program (KAW 2024.0159), David and Astrid Hägelen Foundation, Instrumentarium Science Foundation, KAUTE Foundation, Karolinska Institute Research Foundation, Orion Research Foundation, and Oskar Huttunen Foundation.

Keywords: artificial intelligence | diagnosis | grading | pathology | prostate cancer

ABSTRACT

Artificial intelligence (AI) is an emerging tool in diagnostic pathology, including prostate pathology. This review summarizes the possibilities offered by AI and also discusses the challenges and risks. AI has the potential to assist in the diagnosis and grading of prostate cancer. Diagnostic safety can be enhanced by avoiding the accidental underdiagnosis of small lesions. Another possible benefit is a greater degree of standardization of grading. AI for clinical use needs to be trained on large, high-quality data sets that have been assessed by experienced pathologists. A problem with the use of AI in prostate pathology is the plethora of benign mimics of prostate cancer and morphological variants of cancer that are too unusual to allow sufficient training of AI. AI systems need to be able to account for variations in local routines for cutting, staining, and scanning of slides. We also need to be aware of the risk that users will rely too much on the output of an AI system, leading to diagnostic errors and loss of clinical competence. The reporting pathologist must ultimately be responsible for accepting or rejecting the diagnosis proposed by AI.

1 | Background

In recent years, artificial intelligence (AI) has rapidly made its way into many forms of innovative technology, that is, in the process of transforming our everyday life. The practice of medicine is no exception. While there is considerable hope that AI will both facilitate and improve the quality of healthcare, there is also a fear that doctors will lose control over the decision-making process and in the end possibly also lose their employment [1]. The field of pathology is particularly suitable for the use of AI. Pathologists are expected to interpret microscopic images that contain a huge amount of data with enormous complexity. The use of AI has the potential to enhance the speed of this work, limit the risk of diagnostic errors caused by human fatigue or ignorance, and also standardize the interpretation of images. However, it is of fundamental importance that medical practitioners understand the underlying process of AI development, its potential, and also its risks and limitations.

The rapidly spreading digitization of pathology offers easy access to whole-slide images (WSI). The use of AI in prostate

L.E. was invited to the 113th Meeting of The Japanese Society of Pathology in Nagoya, Japan and this manuscript is partly based upon the presentation given at that meeting.

pathology has been evaluated in numerous studies over the last few years [2, 3]. Several AI tools have recently been made commercially available with Food and Drug Administration (FDA) approval and/or Conformité Européenne (CE)/CE-In Vitro Diagnostic Medical Devices (CE-IVD) marking (Table 1).

The use of AI for diagnostic work in histopathology may serve many purposes. It may save time for pathologists by eliminating tedious examination of large numbers of glass slides with few positive findings, it may increase diagnostic safety or it may provide additional prognostic and treatment predictive information. It may also be used for educational purposes or serve as external quality control of laboratories or individual pathologists. As a consequence, in the establishment of an AI system for diagnostic pathology it is essential to decide its main purpose. In this review, we will discuss the prospects of AI in diagnostic pathology of the prostate and how it may affect the work of pathologists. We will also explore also the limitations and hazards of using this novel technology.

2 | Diagnosis of Prostate Cancer

Advanced AI commonly employs a mechanism known as deep learning where patterns required for replication of a classification are automatically detected [13]. Deep learning for diagnostic pathology utilizes data sets that have been classified by expert pathologists for the training of an AI model. It is essential that the training data sets are classified by pathologists with acknowledged expertise in the field. WSIs are typically divided into segments designated tiles or patches and AI is instructed as to the accurate diagnosis either at case level, slide level or by indicating areas on the scanned slides. To enable the inclusion of a large number of cases in the training data set, case or slide level labeling is often used, known as weakly supervised training. In deep learning, the algorithms are not instructed to search for specific features, but will attempt to identify patterns that best categorize the cases according to the expert classification. This is both a strength and a weakness of deep learning. Probing features without detailed supervision may enable the detection of patterns that are more predictive of a specific diagnosis than other conventional diagnostic criteria. It is, however, problematic that the deep learning algorithms have a "black box" nature and that it is often challenging to know exactly what the diagnosis is based on. There is a risk, in particular if the training of the AI system is not undertaken with care and caution, that established diagnostic criteria will not be met and the rendered diagnosis may be based on irrelevant features. For example, if the training of the AI system is performed using data from two different laboratories with more high-grade cases from one of them, the AI system will quickly learn to associate images generated from that laboratory with the higher grade diagnoses, resulting in biased assessments from the resulting AI model. A careful approach to AI training is thus essential if the trained model is to be applicable to all clinical situations. AI is able to ignore "noise" consisting of occasional outliers, but it is important that a considerable effort is made to set up a large training data set of high and consistent quality. In a case where deep learning has generated an obviously erroneous diagnosis, it is difficult to know the exact cause of the diagnostic error. Ongoing research on "explainable AI," with the aim of developing algorithms that can explain *why* it makes a certain prediction or decision, is likely to gradually mitigate this problem as our understanding evolves.

Numerous studies have shown that AI is capable of reaching a very high level of diagnostic accuracy in the detection of prostatic adenocarcinoma [4, 5, 7, 14–20]. A measure commonly used for summarizing the sensitivity and specificity of the AI system for the detection of prostatic adenocarcinoma is the area under the curve (AUC) of a receiver operating characteristic (ROC) curve. Campanella and colleagues achieved an AUC for prostatic carcinoma in core needle biopsies of 0.986 [4]. Ström and colleagues determined an AUC for identifying cancer of 0.997 in an independent validation data set and 0.986 in an external data set processed at another laboratory and using a different digital pathology scanner [20]. Furthermore, Bulten and colleagues reported an AUC of 0.990 in an internal data set and 0.98–0.99 when compared against two different reference pathologists using an external data set [14]. For comparisons between AI systems developed by different institutions, the Prostate cANcer graDe Assessment (PANDA) data set was

TABLE 1 | Commercially available AI-driven tools for prostate pathology.

Name	Company (Country)	Marks/approval	Description
Paige Prostate Detect [4–6]	Paige AI (USA)	FDA-approved (USA), CE-IVD marked	Cancer detection
Paige Prostate Grade & Quantify [4–6]	Paige AI (USA)	CE-IVD marked	Cancer detection, grading, and quantification
Galen Prostate [7, 8]	Ibex Medical Analytics (Israel)	CE-IVD marked	Cancer detection and grading
Inify Prostate [9]	Inify Laboratories (Sweden)	CE-IVD marked	Cancer detection
HALO Prostate AI [10]	Indica Labs (USA)	CE-IVD marked	Cancer detection and grading
Aiforia Prostate AI [11]	Aiforia (Finland)	CE-IVD marked	Cancer detection and grading
DeepDx Prostate [12]	Deep Bio (South Korea)	MFDS-approved (Korea), CE marked	Cancer detection, grading, and quantification

Abbreviations: CE = Conformité Européenne, FDA = Food and Drug Administration, IVD = In Vitro Diagnostic Medical Devices, MFDS = Ministry of Food and Drug Safety, USA = United States of America.

made publicly available [21]. It included a total of 12 625 WSIs of prostate biopsies from 6 different institutions from Europe and the United States. The data set was used by 1010 teams of developers from 65 countries in a competition aiming at the development of AI algorithms for Gleason grading. Using this resource, Yang and colleagues achieved an AUC for cancer detection of 0.987 [22].

Although AUC is an appropriate and commonly used measure for assessing discriminatory performance of a predictive algorithm, it should be noted that a high AUC is not sufficient for a well-performing algorithm. It is also important that an appropriately chosen cutpoint, to classify slides or cases as "positive" versus "negative," is utilized. This cutoff balances the tradeoff between sensitivity and specificity, and needs to generalize to unseen, fully external data (i.e., tissue samples processed in a different laboratory and digitized on a different digitalization platform). While a high AUC may be retained on fully external data, this does not imply that the chosen cutoff value generalizes and strikes an appropriate balance between sensitivity and specificity.

3 | Grading of Prostate Cancer

Histopathological grading of prostatic carcinoma is currently the most important tissue-based prognostic biomarker. Prostate cancer is graded according to the Gleason grading system. This system is based on architectural patterns [23]. This grading is of critical importance for the treatment decision. Although grading is a powerful predictor of prognosis for prostate cancer, a well-known problem is its lack of reproducibility. Several studies have shown that even pathologists with an expertise in prostate pathology have an interobserver reproducibility that is in the range of moderate to substantial with a linearly weighted κ value of 0.48–0.67 [24–26]. Pathologists with a lower level of expertise generally have an even lower reproducibility, only reaching a moderate level with weighted κ values of 0.41–0.43 [27, 28]. Despite numerous attempts to standardize grading [24, 29, 30], even the experts disagree [24, 27]. A problem that has been noted in recent years is that there has been a general Gleason inflation over the past decades [31]. This is problematic since the prognostic impact of a certain grade changes over time. This Gleason inflation also limits the utility of historical data. More recently, the Gleason scores have increased further by the transition from systematic to multiparametric magnetic resonance imaging (mpMRI)-guided fusion biopsies that target areas of higher grade.

One of the more systematic attempts to standardize grading is the Imagebase image repository, organized under the auspices of the International Society of Urological Pathology [24, 32]. A total of 24 leading international experts in prostate pathology from 5 continents were asked to upload images of prostate cancers into a nonpublic database. Each of the experts then assigned a Gleason score to the cases without knowing the preferences of the other experts. Once a case had reached a 2/3 consensus level (corresponding to a minimum of 16 votes in favor of a certain grade) the case was automatically transferred to a public database, which is an expert-vetted reference image library. This library may be utilized for education and testing purposes, as well as the standardization and calibration of pathologists, and is available to pathologists worldwide. Despite this, the Imagebase project also demonstrates how even leading experts may come to different diagnostic conclusions in borderline cases [24, 33]. In a study focusing on the nonconsensus cases, sources of disagreement were found to include interpretations relating to tangential sectioning, crush artifacts, and mixed high-grade patterns [33].

Using AI for grading has the potential to decrease interobserver variability by using a standardized approach to the diagnosis of borderline morphologies. In numerous studies, a greater reproducibility of grading has been reached with AI than with pathologists [10, 12, 14, 18, 34, 35]. To achieve this with deep learning, the AI system needs to be instructed by a very large training data set that has been graded by expert pathologists. In a study by Ström and colleagues, 6682 slides of prostate biopsies from the STHLM3 screening trial were used for training purposes [20]. As part of the evaluation the performance of AI relating to grading, the Imagebase data set was compared against the diagnoses of Imagebase expert pathologists. The linearly weighted κ of AI was 0.62, which was within the range of these leading experts (0.60–0.73).

When comparing grading results of AI-based studies against previous studies with human pathologists, it is important to take into account the different methodologies used for the calculation of weighted κ statistics. Most historical studies of interobserver reproducibility among pathologists have used linear weighting of κ [10, 25, 27, 28]. However, in AI studies, quadratic weighting is commonly used [12, 14, 36, 37]. κ values obtained by quadratic weighting are not comparable against linearly weighted κ values. Quadratic weighting penalizes greater differences to a larger extent than linear weighting, while at the same time placing less emphasis on small deviations. Since Gleason score discrepancies are often only within ± 1 score [24, 28], the κ values obtained by quadratic weighting tend to be markedly higher than those obtained by linear weighting. For example, Bulten and colleagues asked a panel of 14 pathologists to grade 160 prostate biopsies [34]. When using AI, the agreement with an expert panel grading increased from a quadratically weighted κ of 0.799–0.872. Yang and colleagues reached a quadratically weighted κ of 0.860 for ISUP grades in the PANDA data set [22]. Interestingly, the extremes of grading (ISUP Grades 1 and 5) had the best agreement while the mid grades were more challenging for AI.

4 | Interaction Between Pathologists and AI

A key to the understanding of the utility of AI in the reporting of prostate biopsies is the analysis of its performance in a clinical setting. For the evaluation of the impact of AI on the diagnosis, there is a need to define a ground truth that AI evaluation can be compared against. This has been done either by review by an expert panel or by trusting the original diagnosis supported by immunohistochemistry (Table 2). In several studies, prostate biopsies have been read by pathologists without and with AI assistance [15, 19, 40].

4.1 | Accuracy in Cancer Detection

Raciti and colleagues tasked 3 pathologists to classify 304 core needle specimens as benign or suspicious for cancer [5].

TABLE 2 | Studies validating the performance of AI-driven prostate pathology in needle biopsies compared to panels of multiple human pathologists.

References	Year	Company	Number of pathologists	Validation cases	Tested parameters
Raciti et al. [5]	2020	Paige AI	3	304	Diagnosis
Pantanowitz et al. [7]	2020	Ibex Medical Analytics	2–3	100	Diagnosis, grade, PNI
Dov et al. [16]	2020	—	4	100	Diagnosis
Steiner et al. [35]	2020	—	20	240	Diagnosis, grade, time
Ström et al. [20]	2020	—	1 + 23[a]	73 + 87[b]	Diagnosis, grade
Bulten et al. [14]	2020	—	13 + 2[c]	100	Diagnosis, grade
da Silva et al. [15]	2021	Paige AI	3	100	Diagnosis
Bulten et al. [34]	2021	—	11 + 3[d]	160	Grade
Huang et al. [18]	2021	—	3	162	Diagnosis, grade, time
Marginean et al. [38]	2021	—	2	21	Diagnosis, grade
Jung et al. [12]	2022	DeepDx Prostate	3	594	Diagnosis, grade, time
Kartasalo et al. [39]	2022	—	4	286	PNI
Raciti et al. [19]	2023	Paige AI	16	610	Diagnosis
Eloy et al. [40]	2023	Paige AI	4	41	Diagnosis, grade, PNI, cribriform cancer
Vazzano et al. [9]	2023	Inify Laboratories	2/slide	30	Diagnosis
Tolkach et al. [10]	2023	Indica Labs	11	423	Diagnosis, grade
Santa-Rosario et al. [8]	2024	Ibex Medical Analytics	4	101	Diagnosis, grade

Abbreviations: AUC = area under the curve, PNI = perineural invasion.
[a]1 pathologist for diagnosis and 23 for grading.
[b]73 men for diagnosis and 87 biopsies for grading.
[c]13 pathologists and 2 pathologists in training.
[d]11 pathologists and 3 pathologists in training.

The original diagnoses rendered by pathologists specialized in urological pathology were used as ground truth. After a washout period of 4 weeks they reread the biopsies again, assisted by the Paige Prostate Alpha system. The sensitivity versus ground truth was 74% in the first reading and this increased to 90% with AI assistance, without any significant change in specificity. The low sensitivity level is surprising since the classification grouped suspicious and carcinoma into a single category. The lack of available immunohistochemistry may make it difficult in some cases to render a definitive diagnosis of cancer, but at the very least, a suspicion of cancer should be possible based on hematoxylin and eosin-stained sections. In a later study from the same group, 18 pathologists were asked to read 610 biopsies from a total of 218 institutions twice, first without AI and then together with AI [19]. They found an improvement of both sensitivity and specificity, but it was noted that there appeared to have been no washout period between the readings.

A study by Eloy and colleagues failed to find an improved accuracy when AI was used [40]. Four pathologists were asked to review WSIs paired with immunohistochemical stains of all slides from 41 men. When the slides were evaluated again with AI assistance after a washout period, no improvement of sensitivity or specificity was determined. Da Silva and colleagues found that AI assistance improved the sensitivity for detecting cancer, while the specificity was lower than that achieved with both conventional microscopy and unassisted assessment of digital pathology [15].

In a study by Jung and colleagues, 593 prostate biopsies were analyzed by a pathologist with AI support and the results compared against the original report with a review by a panel of 3 experts in uropathology as gold standard [12]. Sensitivity and specificity for cancer detection were comparable or superior to that of the original reports.

In a small study, Vazzano and colleagues used an expert in urological pathology as ground truth for the diagnosis of prostate cancer in 30 selected core needle biopsies (20 malignant and 10 benign) [9]. The biopsies were scanned by 3 different scanners and then diagnosed by pathologists who were given access to the 90 generated WSIs with AI mappings of areas suspicious for cancer. Percent cancer length estimated by the pathologists correlated with percent cancer areas predicted by AI ($R^2 = 0.765–0.799$ for the three scanner models). The sensitivity was high (99%), but the specificity was lower at 93%–98%.

Tolkach and colleagues analyzed biopsies from 423 men across 3 cohorts with an in-house AI algorithm [10]. Similar to others, they found a high sensitivity (97%–100%) but lower specificity (87%–98%). As expected, false positive diagnoses included atypia suspicious for cancer and benign mimics such as granulomatous prostatitis.

It has been argued that the use of AI may reduce the need of immunohistochemistry and second opinion [41]. In support of

this, Eloy and colleagues found a 20% reduction of the need of immunohistochemistry when AI was used to confirm the diagnosis [40].

A common denominator for many of the studies comparing human pathologists and AI is a high sensitivity but lower specificity. This is not surprising if the aim is to utilize AI for detecting areas suspicious for cancer rather than for producing a definitive diagnosis of cancer. However, an obvious risk with this strategy is that AI may lead to an overdiagnosis of cancer in the hands of an inexperienced pathologist, which is similar to the results that are seen following the overuse of immunohistochemistry for prostate cancer for prostate cancer detection [42].

4.2 | Reduction of Workload

One of the arguments for using AI in diagnostic pathology is the potential to save time for the pathologists, either by reducing the amount of time spent examining each slide or by eliminating slides that should not require evaluation by a pathologist. In a study by Jung and colleagues, the time spent in diagnosis was reduced from 55.7 s/case to 36.8 s, although it was not explained how this time reduction was achieved [12]. Huang and colleagues claimed that the time used to diagnose and grade cancer and estimate its extent was reduced even further, from 4–6 min/slide to <1 min [18]. This remarkable reduction of evaluation time must reasonably have been achieved by trusting and largely accepting the results of AI and spending limited time verifying the results. Since the specificity is relatively low in many AI studies it seems that this approach may lead to false positive diagnoses. Other studies have reduced the review time by a more modest 13.5%–22% [35, 40]. Despite this, time-saving could still be a reasonable target for some diagnostic aspects such as the screening of lymph nodes where the likelihood of finding metastatic deposits is low. For this, diagnostic AI must have a very high sensitivity in order not to miss any focus of cancer. In a study on the detection of lymph node metastases of breast cancer, with immunohistochemistry as gold standard, the performance of AI was superior to a panel of 11 pathologists reading the slides under time constraint [43].

Giving priority to sensitivity in the detection of cancer would enable the use of AI as a safety system to ensure that pathologists do not overlook small foci of cancer. On the other hand, if the aim is to generate a diagnosis that is most likely to be correct, then the trade-off between sensitivity and specificity needs to be more balanced.

5 | Challenges in the Implementation of AI in Prostate Pathology

AI is often evaluated on data sets from a single laboratory. However, there is often a variation in the cutting and staining of sections between laboratories and this will most likely have an impact on AI pathology [36, 44]. In one study, a considerable lack of reproducibility of color balance in hematoxylin and eosin stains was noted both between five laboratories in the United Kingdom and over time in the same laboratory, even when using an autostainer [45]. Moreover, the choice of scanner also affects the results of the AI interpretation [9, 21]. Duenweg and colleagues digitized sections from 30 radical prostatectomy specimens using 3 different scanner models and found significant differences in color intensity and algorithmically estimated tissue density in the WSIs [46]. Faryna and colleagues compared 113 biopsy cases scanned by 5 different scanner models and graded by a panel of 10 pathologists with an expertise in urological pathology against 2 commercially available AI algorithms [36]. The quadratically weighted κ of the majority vote of 5 ISUP grades compared with AI grading varied from 0.860 to 0.900. Pantanowitz and colleagues chose to recalibrate their AI tool for the local scanner used in their external validation [7]. While local recalibration seems to be a possible path to addressing the challenge of generalizing AI systems and maintaining an appropriate balance between sensitivity and specificity on fully external data, it may also have a negative effect. Specifically, local recalibration of AI may obscure its potential to improve reproducibility and also decrease its accuracy if, for example, the AI system has been trained on data obtained from specialists in uropathology, and is subsequently recalibrated to the grading practices of local general pathologists. Centralization of laboratory work is not a realistic solution to the standardization problem, since there will always be laboratories that prefer maintaining their independence.

One of the challenges in the evaluation of prostate pathology is the plethora of differential diagnoses, in particular benign mimics that may look very similar to prostatic carcinomas [42]. Benign proliferations have an atypia that is mainly architectural, often with closely packed small glands, but with no or minimal nuclear atypia. Such lesions include adenosis, sclerosing adenosis, postatrophic hyperplasia, and verumontanum gland hyperplasia. Anatomical structures such as seminal vesicle, ejaculatory duct, and Cowper's glands also fall into this category, although the former may at times display a considerable nuclear atypia. However, even lesions with relatively large glands such as clear cell cribriform hyperplasia may cause differential diagnostic concern. A problem with the use of AI for identification of these rare diagnostic entities is that training of AI requires very large data sets and these can be difficult to obtain. A possible approach is to permit AI to express not only the most likely diagnosis but also a level of diagnostic uncertainty [47].

The interaction between AI and human pathologists will need to be closely monitored. As with all new technologies, there is a risk that users will either have too little or too much confidence in it.

6 | Additional Diagnostic Information

In addition to determining diagnosis and grade, pathologists are expected to report on a number of features in prostate biopsies that have been shown to correlate with prognosis [48, 49].

Tumor extent in needle biopsies may be reported either as a percentage of the core or millimeter cancer length [48], and AI has the potential to assist in assessment. Ström and colleagues found a strong correlation between tumor extent assessed by a pathologist and the AI estimation with a Spearman's rank

coefficient of 0.96 [20]. The reporting of tumor extent may, however, conflict with the reporting of diagnosis or grade. For example, if two different AI models are trained for different tasks, there is no guarantee that they always produce consistent results. It is entirely possible that the diagnostic model may indicate that cancer is present, but the tumor extent model shows 0 mm cancer length. These challenges are even greater with the reporting of the extent of Gleason Pattern 4 and the reporting of the Gleason score, as this is dependent on the proportion of a specific tumor grade present in the biopsy. The practicing pathologist should be aware of these challenges relating to the designing of AI algorithms.

Perineural invasion of prostate cancer predicts outcome after radical prostatectomy [50] and guidelines recommend its reporting in needle biopsies [49]. The identification of perineural invasion is, however, tedious and hampered by lack of reproducibility [51]. It has been demonstrated that AI can assist with the identification of perineural invasion of prostate cancer in needle biopsies with an AUC of 0.98 [39]. Eloy and colleagues, however, found that the use of AI did not significantly improve the detection of perineural invasion [40]. Similarly, they did not find a significant improvement of the diagnosis of cribriform patterns or intraductal cancer. Contrary to this, Eminaga and colleagues demonstrated a moderate to substantial concordance between pathologists and AI in the detection of cribriform patterns, perineural invasion, and lymphovascular invasion ($\varkappa$ 0.49–0.71) [17].

7 | Prediction of Prognosis

The strength of deep learning algorithms is the ability to probe features that are helpful for the building of classifiers beyond our current morphological knowledge [13]. The use of AI for the assessment of Gleason scores is not necessarily the best way to utilize AI for the prediction of prognosis. Despite the strong correlation between the Gleason score and the outcome of prostate cancer, there may be other features that perform as a greater predictor of outcome. In a tissue microarray study on the prediction of biochemical recurrence after radical prostatectomy, a deep-learning system added independent prognostic information to the prediction provided by ISUP grades [52]. An AI model was trained and validated on 16 204 prostate biopsy slides from randomized clinical trials split into 80% training and 20% validation cases with 10-year outcome data [53]. The model outperformed predictions by the National Cancer Center Network (NCCN) riskgroups by 9.2%–14.6%. A model for the prediction of response to hormonal therapy was developed by Spratt et al. [54]. Pretreatment biopsies from patients enrolled in trials on radiotherapy with or without androgen deprivation therapy were used to develop and validate a predictive model. Hormonal therapy only had effect in patients predicted to be responders and, thus, adverse side effects could be avoided in men where AI showed a low likelihood of treatment benefit.

8 | How Will AI be Used by Pathologists?

A recent survey among 24 pathologists with an expertise in the implementation of AI in pathology confirmed many assumptions about the promises of AI. Despite this, the survey also showed that the experts agreed with the statement that hurried pathologists may often take “shortcuts” by accepting AI interpretations without verifying the diagnosis morphologically [55]. Thus, the clinical implementation of AI in pathology comes with a risk that the pathologists will automatically accept the diagnosis suggested by AI, thus potentially leading to diagnostic errors such as overdiagnosis of cancer. This has medicolegal implications and the suggestion that pathologists should be legally responsible for diagnoses made with the help of AI was strongly supported by the expert pathologists [55].

If the medicolegal responsibility was moved from the pathologist to the software developer, the diagnostic accuracy of the AI models would need to be carefully audited by international organizations, requiring an extensive control system that might hamper the development of AI. A formal responsibility shared between developers and users would, on the other hand, be legally complex.

9 | A Policy Proposal for the Use of AI in Pathology

The challenge for the pathology community is to use AI in the diagnostic work without losing competence and without making diagnostic errors. Pathologists should be required to enter their own diagnosis and in cancer cases the tumor grade, before obtaining access to the AI results. By doing this, pathologists will need to stay focused and maintain their professional competence.

After accessing the AI diagnosis, it should be possible for the pathologist to amend their initial diagnosis if necessary. Importantly, if AI has highlighted a cancer focus that was obviously overlooked by the pathologist, it would be inappropriate if amendments were not permitted.

As noted earlier, the pathologist should be legally responsible for the final diagnosis. Despite advances, it will most likely be impossible to avoid AI-generated errors. If the pathologist is not responsible for the surveillance of the AI process the question of legal responsibility could become very complicated. It is not reasonable to assume that computer engineers who developed AI algorithms would have a medical responsibility for unexpected errors, especially as the balance between sensitivity and specificity in the detection of cancer may be set to favor sensitivity.

The entire diagnostic process should be logged into the system stating whether the diagnosis was generated by AI or by the human pathologist, and whether an AI-generated diagnosis was modified. This is important for the legal documentation, but also for the improvement of AI analyses.

By limiting the disadvantages of AI and applying transparent reporting policies, we believe that AI will develop into a useful tool that could enhance diagnostic safety and also minimize tedious routine work in histopathology. This would allow the pathologists to focus their attention on difficult diagnostic decisions such as the diagnosis of unusual tumor morphologies and the detection of tumor mimics.

Author Contributions

Conception and design of the study – Lars Egevad. Literature search – Lars Egevad and Andrea Camilloni. Drafting the manuscript – Lars Egevad, Brett Delahunt, Hemamali Samaratunga, Martin Eklund, and Kimmo Kartasalo.

Acknowledgments

L.E. was funded by grants from The Swedish Cancer Foundation (Grant No. CAN 20 1358 PjFs and 23 2641Pj) and The Stockholm Cancer Society (Grant No. 204043). K.K. received funding from the SciLifeLab & Wallenberg Data Driven Life Science Program (KAW 2024.0159), David and Astrid Hägelen Foundation, Instrumentarium Science Foundation, KAUTE Foundation, Karolinska Institute Research Foundation, Orion Research Foundation, and Oskar Huttunen Foundation.

Conflicts of Interest

L.E., M.E., and K.K. are shareholders in Clinsight AB, which develops AI algorithms.

References

1. S. I. Lambert, M. Madi, S. Sopka, et al., "An Integrative Review on the Acceptance of Artificial Intelligence Among Healthcare Professionals in Hospitals," *NPJ Digital Medicine* 6 (2023): 111.

2. A. Morozov, M. Taratkin, A. Bazarkin, et al., "A Systematic Review and Meta-Analysis of Artificial Intelligence Diagnostic Accuracy in Prostate Cancer Histology Identification and Grading," *Prostate Cancer and Prostatic Diseases* 26 (2023): 681–692.

3. L. Zhu, J. Pan, W. Mou, et al., "Harnessing Artificial Intelligence for Prostate Cancer Management," *Cell Reports Medicine* 5 (2024): 101506.

4. G. Campanella, M. G. Hanna, L. Geneslaw, et al., "Clinical-Grade Computational Pathology Using Weakly Supervised Deep Learning on Whole Slide Images," *Nature Medicine* 25 (2019): 1301–1309.

5. P. Raciti, J. Sue, R. Ceballos, et al., "Novel Artificial Intelligence System Increases the Detection of Prostate Cancer in Whole Slide Images of Core Needle Biopsies," *Modern Pathology* 33 (2020): 2058–2066.

6. S. Perincheri, A. W. Levi, R. Celli, et al., "An Independent Assessment of an Artificial Intelligence System for Prostate Cancer Detection Shows Strong Diagnostic Accuracy," *Modern Pathology* 34 (2021): 1588–1595.

7. L. Pantanowitz, G. M. Quiroga-Garza, L. Bien, 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 Digital Health* 2 (2020): e407–e416.

8. J. C. Santa-Rosario, E. A. Gustafson, D. E. Sanabria Bellassai, P. E. Gustafson, and M. de Socarraz, "Validation and Three Years of Clinical Experience in Using an Artificial Intelligence Algorithm as a Second Read System for Prostate Cancer Diagnosis-Real-World Experience," *Journal of Pathology Informatics* 15 (2024): 100378.

9. J. Vazzano, D. Johansson, K. Hu, et al., "Evaluation of a Computer-Aided Detection Software for Prostate Cancer Prediction: Excellent Diagnostic Accuracy Independent of Preanalytical Factors," *Laboratory Investigation* 103 (2023): 100257.

10. Y. Tolkach, V. Ovtcharov, A. Pryalukhin, et al., "An International Multi-Institutional Validation Study of the Algorithm for Prostate Cancer Detection and Gleason Grading," *NPJ Precision Oncology* 7 (2023): 77.

11. K. Sandeman, S. Blom, V. Koponen, et al., "AI Model for Prostate Biopsies Predicts Cancer Survival," *Diagnostics* 12 (2022): 1031.

12. M. Jung, M. S. Jin, C. Kim, et al., "Artificial Intelligence System Shows Performance at the Level of Uropathologists for the Detection and Grading of Prostate Cancer in Core Needle Biopsy: An Independent External Validation Study," *Modern Pathology* 35 (2022): 1449–1457.

13. Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature* 521 (2015): 436–444.

14. W. Bulten, H. Pinckaers, H. van Boven, et al., "Automated Deep-Learning System for Gleason Grading of Prostate Cancer Using Biopsies: A Diagnostic Study," *Lancet Oncology* 21 (2020): 233–241.

15. L. M. da Silva, E. M. Pereira, P. G. Salles, et al., "Independent Real-World Application of a Clinical-Grade Automated Prostate Cancer Detection System," *Journal of Pathology* 254 (2021): 147–158.

16. D. Dov, S. Assaad, A. Syedibrahim, et al., "A Hybrid Human-Machine Learning Approach for Screening Prostate Biopsies Can Improve Clinical Efficiency Without Compromising Diagnostic Accuracy," *Archives of Pathology & Laboratory Medicine* 146 (2022): 727–734.

17. O. Eminaga, M. Abbas, C. Kunder, et al., "Critical Evaluation of Artificial Intelligence as a Digital Twin of Pathologists for Prostate Cancer Pathology," *Scientific Reports* 14 (2024): 5284.

18. W. Huang, R. Randhawa, P. Jain, et al., "Development and Validation of an Artificial Intelligence-Powered Platform for Prostate Cancer Grading and Quantification," *JAMA Network Open* 4 (2021): e2132554.

19. P. Raciti, J. Sue, J. A. Retamero, et al., "Clinical Validation of Artificial Intelligence-Augmented Pathology Diagnosis Demonstrates Significant Gains in Diagnostic Accuracy in Prostate Cancer Detection," *Archives of Pathology & Laboratory Medicine* 147 (2023): 1178–1185.

20. P. Ström, K. Kartasalo, H. Olsson, et al., "Artificial Intelligence for Diagnosis and Grading of Prostate Cancer in Biopsies: A Population-Based, Diagnostic Study," *Lancet Oncology* 21 (2020): 222–232.

21. W. Bulten, K. Kartasalo, P. H. C. Chen, et al., "Artificial Intelligence for Diagnosis and Gleason Grading of Prostate Cancer: The PANDA Challenge," *Nature Medicine* 28 (2022): 154–163.

22. Z. Yang, X. Wang, J. Xiang, et al., "The Devil Is in the Details: A Small-Lesion Sensitive Weakly Supervised Learning Framework for Prostate Cancer Detection and Grading," *Virchows Archiv* 482 (2023): 525–538.

23. D. F. Gleason, "Histologic Grading of Prostate Cancer: A Perspective," *Human Pathology* 23 (1992): 273–279.

24. L. Egevad, B. Delahunt, D. M. Berney, et al., "Utility of Pathology Imagebase for Standardisation of Prostate Cancer Grading," *Histopathology* 73 (2018): 8–18.

25. A. Glaessgen, H. Hamberg, C. G. Pihl, B. Sundelin, B. O. Nilsson, and L. Egevad, "Interobserver Reproducibility of Percent Gleason Grade 4/5 in Prostate Biopsies," *Journal of Urology* 171 (2004): 664–667.

26. A. Glaessgen, H. Hamberg, C. G. Pihl, B. Sundelin, B. Nilsson, and L. Egevad, "Interobserver Reproducibility of Modified Gleason Score in Radical Prostatectomy Specimens," *Virchows Archiv* 445 (2004): 17–21.

27. W. C. Allsbrook, Jr., K. A. Mangold, M. H. Johnson, et al., "Interobserver Reproducibility of Gleason Grading of Prostatic Carcinoma: Urologic Pathologists," *Human Pathology* 32 (2001): 74–80.

28. L. Egevad, A. S. Ahmad, F. Algaba, et al., "Standardization of Gleason Grading Among 337 European Pathologists," *Histopathology* 62 (2013): 247–256.

29. L. Egevad, "Reproducibility of Gleason Grading of Prostate Cancer Can be Improved by the Use of Reference Images," *Urology* 57 (2001): 291–295.

30. J. D. Kronz, M. A. Silberman, W. C. Allsbrook, and J. I. Epstein, "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* 89 (2000): 1818–1823.

31. D. Danneman, L. Drevin, D. Robinson, P. Stattin, and L. Egevad, "Gleason Inflation 1998-2011: A Registry Study of 97,168 Men," *BJU International* 115 (2015): 248–255.

32. L. Egevad, J. Cheville, A. J. Evans, et al., "Pathology Imagebase-A Reference Image Database for Standardization of Pathology," *Histopathology* 71 (2017): 677–685.

33. L. Egevad, D. Swanberg, B. Delahunt, et al., "Identification of Areas of Grading Difficulties in Prostate Cancer and Comparison With Artificial Intelligence Assisted Grading," *Virchows Archiv* 477 (2020): 777–786.

34. W. Bulten, M. Balkenhol, J. J. A. Belinga, et al., "Artificial Intelligence Assistance Significantly Improves Gleason Grading of Prostate Biopsies by Pathologists," *Modern Pathology* 34 (2021): 660–671.

35. D. F. Steiner, K. Nagpal, R. Sayres, et al., "Evaluation of the Use of Combined Artificial Intelligence and Pathologist Assessment to Review and Grade Prostate Biopsies," *JAMA Network Open* 3 (2020): e2023267.

36. K. Faryna, L. Tessier, J. Retamero, et al., "Evaluation of Artificial Intelligence-Based Gleason Grading Algorithms 'in the Wild'," *Modern Pathology* 37 (2024): 100563.

37. G. Nir, D. Karimi, S. L. Goldenberg, et al., "Comparison of Artificial Intelligence Techniques to Evaluate Performance of a Classifier for Automatic Grading of Prostate Cancer From Digitized Histopathologic Images," *JAMA Network Open* 2 (2019): e190442.

38. F. Marginean, I. Arvidsson, A. Simoulis, et al., "An Artificial Intelligence-Based Support Tool for Automation and Standardisation of Gleason Grading in Prostate Biopsies," *European Urology Focus* 7 (2021): 995–1001.

39. K. Kartasalo, P. Ström, P. Ruusuvuori, et al., "Detection of Perineural Invasion in Prostate Needle Biopsies With Deep Neural Networks," *Virchows Archiv* 481 (2022): 73–82.

40. C. Eloy, A. Marques, J. Pinto, et al., "Artificial Intelligence-Assisted Cancer Diagnosis Improves the Efficiency of Pathologists in Prostatic Biopsies," *Virchows Archiv* 482 (2023): 595–604.

41. A. Chatrian, R. T. Colling, L. Browning, et al., "Artificial Intelligence for Advance Requesting of Immunohistochemistry in Diagnostically Uncertain Prostate Biopsies," *Modern Pathology* 34 (2021): 1780–1794.

42. L. Egevad, B. Delahunt, B. Furusato, T. Tsuzuki, J. Yaxley, and H. Samaratunga, "Benign Mimics of Prostate Cancer," *Pathology* 53 (2021): 26–35.

43. B. Ehteshami Bejnordi, M. Veta, P. Johannes van Diest, et al., "Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer," *Journal of the American Medical Association* 318 (2017): 2199–2210.

44. E. L. Clarke and D. Treanor, "Colour in Digital Pathology: A Review," *Histopathology* 70 (2017): 153–163.

45. A. Gray, A. Wright, P. Jackson, M. Hale, and D. Treanor, "Quantification of Histochemical Stains Using Whole Slide Imaging: Development of a Method and Demonstration of Its Usefulness in Laboratory Quality Control," *Journal of Clinical Pathology* 68 (2015): 192–199.

46. S. R. Duenweg, S. A. Bobholz, A. K. Lowman, et al., "Whole Slide Imaging (WSI) Scanner Differences Influence Optical and Computed Properties of Digitized Prostate Cancer Histology," *Journal of Pathology Informatics* 14 (2023): 100321.

47. H. Olsson, K. Kartasalo, N. Mulliqi, et al., "Estimating Diagnostic Uncertainty in Artificial Intelligence Assisted Pathology Using Conformal Prediction," *Nature Communications* 13 (2022): 7761.

48. L. Egevad, M. Judge, B. Delahunt, et al., "Dataset for the Reporting of Prostate Carcinoma in Core Needle Biopsy and Transurethral Resection and Enucleation Specimens: Recommendations From the International Collaboration on Cancer Reporting (ICCR)," *Pathology* 51 (2019): 11–20.

49. N. Mottet, R. C. N. van den Bergh, E. Briers, et al., "EAU-EANM-ESTRO-ESUR-SIOG Guidelines on Prostate Cancer-2020 Update. Part 1: Screening, Diagnosis, and Local Treatment With Curative Intent," *European Urology* 79 (2021): 243–262.

50. M. K. Tollefson, R. J. Karnes, E. D. Kwon, et al., "Prostate Cancer Ki-67 (MIB-1) Expression, Perineural Invasion, and Gleason Score as Biopsy-Based Predictors of Prostate Cancer Mortality: The Mayo Model," *Mayo Clinic Proceedings* 89 (2014): 308–318.

51. L. Egevad, B. Delahunt, H. Samaratunga, et al., "Interobserver Reproducibility of Perineural Invasion of Prostatic Adenocarcinoma in Needle Biopsies," *Virchows Archiv* 478 (2021): 1109–1116.

52. H. Pinckaers, J. van Ipenburg, J. Melamed, et al., "Predicting Biochemical Recurrence of Prostate Cancer With Artificial Intelligence," *Communications Medicine* 2 (2022): 64.

53. A. Esteva, J. Feng, D. van der Wal, et al., "Prostate Cancer Therapy Personalization via Multi-Modal Deep Learning on Randomized Phase III Clinical Trials," *NPJ Digital Medicine* 5 (2022): 71.

54. D. E. Spratt, S. Tang, Y. Sun, et al., "Artificial Intelligence Predictive Model for Hormone Therapy Use in Prostate Cancer," *NEJM Evidence* 2 (2023): EVIDoa2300023.

55. M. A. Berbis, D. S. McClintock, A. Bychkov, et al., "Computational Pathology in 2030: A Delphi Study Forecasting the Role of AI in Pathology Within the Next Decade," *EBioMedicine* 88 (2023): 104427.