Medical Research Archives

European Society of Medicine

OPEN ACCESS

Published: November 30, 2023

Citation: Brehler M, Walhagen P, et al., 2023. Difficulties and Recommendations for AI-Based Prediction of Prostate Cancer Aggressiveness in Digital Pathology, Medical Research Archives, [online] 11(11). https://doi.org/10.18103/mra.v11i10.4586

DOI
https://doi.org/10.18103/mra.v11i10.4586

ISSN: 2375-1924

REVIEW ARTICLE

Difficulties and Recommendations for AI-Based Prediction of Prostate Cancer Aggressiveness in Digital Pathology

Michael Brehler[1], Peter Walhagen[2], Christer Busch[2,3], Stefan Bonn*[,1,2], Ewert Bengtsson*[,2,4]

[1]Institute of Medical Systems Biology, Center for Biomedical AI (bAIome), Center for Molecular Neurobiology Hamburg (ZMNH), University Medical Center Hamburg-Eppendorf, Hamburg, Germany
[2]Spearpoint Analytics AB, Stockholm, Sweden
[3]Uppsala University, Faculty of Medicine, Department of Surgical Sciences, Uppsala, Sweden
[4]Uppsala University, Department of Information Technology, Centre for image analysis Uppsala, Sweden

*These authors contributed equally. Correspondence to sbonn@uke.de and ewert.bengtsson@it.uu.se

ABSTRACT

Prostate cancer is among the most common cancers in men with around 1.4 million new cases each year world-wide. A vital part in the diagnosis of prostate cancer is the evaluation of its severity using biopsies and histopathology. Recent progress in artificial intelligence-based image analysis has led to a flurry of algorithms for the automated analysis of prostate cancer histopathological data focusing on the detection of cancerous areas, the grading of cancer severity, and patient outcome. Some of these approaches have reached human expert-level performance and digital models trained directly on patient outcomes might surpass human performance in the future.

Although these results hold great promise for the future usage of digital pathology in clinical settings, several bottlenecks remain to be addressed. Especially the robustness, reliability and trustworthiness of predictions must be guaranteed across a wide range of variation in protocols and instrumentation. While human experts are relatively robust to technical and biological variation in biopsies, artificial intelligence-based systems tend to struggle with differences in staining intensity, color, scanner type, and image resolution, impeding the clinical usage of digital models.

In this work we highlight salient problems and minimal requirements of computational pathology for future use in clinical settings, while focusing on prostate cancer as a use case. In particular, we highlight data and model problems and solutions that include data variability, dataset size, and data annotations, as well as model robustness to data heterogeneity, model prediction confidence, and the explainability of model decisions. While model and data requirements for successful computational pathology in clinics will be highlighted, legal, ethical, and deployment requirements will not be addressed in this review.

In summary, we provide a short overview of the field, salient problems, and potential solutions to harvest the full potential of digital pathology for prostate cancer in clinical practice.

The potential of artificial intelligence in prostate cancer digital pathology

The rapid development in artificial intelligence (AI) over the last years is finding many potential uses in medicine. One of the promising application fields is for grading prostate cancer (PC), which is one of the most common cancers among men with around 1.4 million new annual cases world-wide[1]. Cancers vary widely in how aggressively they grow, and it is important to grade the tumor to determine the type of treatment that finds the best balance between risk of deadly progression and morbidity caused by the treatment. The gold standard of grading is by manual inspection of tissue from biopsies and assigning a Gleason grade, which signifies cancer severity[2,3]. The grading system was updated in 2005 by the International Society of Urological Pathology (ISUP)[4]. With the introduction of high-throughput digital slide scanners, digital pathology opens up the possibility to subsequently use computer models to automate parts of the tissue analysis and grading process.

Prostate cancer digital pathology (PCDP) can be broadly categorized into three different approaches. The first approach, image segmentation, is the delineation of glands, nuclei, and other biological features of interest that are associated with prostate cancer. Subsequent to the segmentation, the extracted features may be used to classify the severity of the cancer. With the introduction of new AI approaches in the last couple of years several groups have achieved good performance on internal validation cohorts, reaching an AUC of 0.99 for epithelial cell detection for instance, but data on external validation cohorts is largely missing[5–9]. In this context it is pivotal to define the difference between external and internal validation data. While internal validation data is not used for training of the PCDP, it is usually from the same cohort as the training data and therefore bears similar characteristics (staining, thickness, resolution, distribution). External validation data, on the other hand, is not used for training and stems from a different clinical source, bearing different characteristics. Hence, a true benchmark for the usability of a PCDP system needs to be evaluated on external validation data. In general, many studies on PCDP segmentation were conducted on very small data sets, usually in the range of 10 to 100 training samples, potentially due to the time-consuming process of creating manual annotations[10–12]. Given the absence of external validation data, the need for expert annotations, and the small data set sizes, it is still questionable if PCDP-guided segmentation could be robust enough to be of clinical use.

The second approach is PCDP-guided cancer detection, which aims at classifying the presence of cancer in a biopsy, subsection of a biopsy, or even at a pixel level. The performance of the model and pathologist is measured as the area under the receiver operating characteristic (ROC AUC), for example, with values ranging between 0.5 (random prediction) and 1.0 (perfect prediction). Many studies have focused on this task, and several have reached excellent performance that rival human experts[13]. Campanella et al, for instance, reached a biopsy-level cancer detection ROC AUC of 0.991 and 0.932 on internal and external validation cohorts, respectively[14]. The external validation cohort consisted of over 17,000 whole slide images (WSI). The overall good performance of biopsy-level cancer detection algorithms on external validation data achieved by several independent groups suggests that this approach might be suitable for clinical use.

The third procedure is PCDP-guided Gleason grade prediction, which aims to grade the severity of prostate cancer on a biopsy in accordance with the ISUP standard. In this case the PCDP performance is measured as the concordance of Gleason grading with clinical experts, measured as quadratic Cohens Kappa (qk). Several recent studies showed human expert-level performance on internal validation data sets, while having slightly lower performance on external validation data[6,15–17]. Bulten et al. achieved a PCDP qk of 0.85 on internal validation data, whereas the qk on the external validation data shrunk to 0.72 and 0.71[6]. The benchmark standard in Gleason grade prediction has been performed recently in the PANDA (prostate cancer grade assessment) challenge, which featured heterogeneous biopsy data from several international clinics for training and testing and more than 1,000 groups that developed PCDP models[18]. The best model achieved a qk of 0.88 on internal and 0.83 on the external validation, which is on par with the concordance between expert pathologists (0.82). This study suggests that, given enough high-quality training data, PCDP-guided Gleason grade prediction can be performed at a performance that rivals clinical experts, even on external validation data.

The above observation that PCDP models can potentially reach expert performance when trained on sufficiently large and heterogeneous data is also reflected in the recent appearance of several commercial PCDP solutions[8,19–21].

While these results strongly suggest that PCDP models will become a mainstay in clinical PC evaluation, the decreased performance of nearly all PCDP systems on external validation data still poses the question how trustworthy and robust these systems really are. Furthermore, it is still unclear if PCDP models can outperform human experts in segmentation, cancer detection, or cancer aggressiveness grading.

Contemporary difficulties and potential solutions

In this section we delve into the challenges that the use of AI in the context of PCDP faces and discuss potential solutions. The three critical pillars discussed are, namely, those rooted in data, model robustness, and the overarching goal of predicting patient outcomes.

DATA

The first step in deploying an AI model to help either in the decision-making process or for full task automatization is the acquisition and inspection of data the model will be trained on. A major problem faced in the deployment of AI solutions for PCDP is the inherent time-delay of follow-up data. If a study with a specific cohort is planned it would take 5-10 years to gather a new dataset including follow-up patient data. This limits the AI model construction to already existing data sets. In the following we will focus on several data-driven problems that need to be taken into account.

Data variability. Arguably, the lack of robustness to data variation is the most prevalent factor that prohibits the usage of PCDP systems in clinical practice to date. Creating a digital slide involves a series of sequential steps: formalin fixation and paraffin embedding of tissue, sectioning, staining and slide scanning. Each of these steps encompasses numerous parameters that differ between clinics, research institutions and even within the same lab over time. Figure 1 depicts several variations in color caused by different protocols. In the worst case, different staining techniques might highlight different biological entities, such as membranes, nuclei, or glands, for instance. While this heterogeneity poses relatively little problems to expert pathologists, it can heavily affect PCDP models.

In recent years, numerous approaches dealing with data variation and prevention of overfitting, the training of a model so that it fits the training data so closely to the point that it captures noise rather than underlying patterns and relationships, have been proposed in the literature. Most of these methods worked well under lab conditions but fell short when confronted with the wide variation in specimen preparation encountered in clinical routine. It is thus essential both to know when present tissue samples fall outside the model domain, the domain of data the model was trained on, and ensure that models give accurate predictions for these cases or reject a prediction.

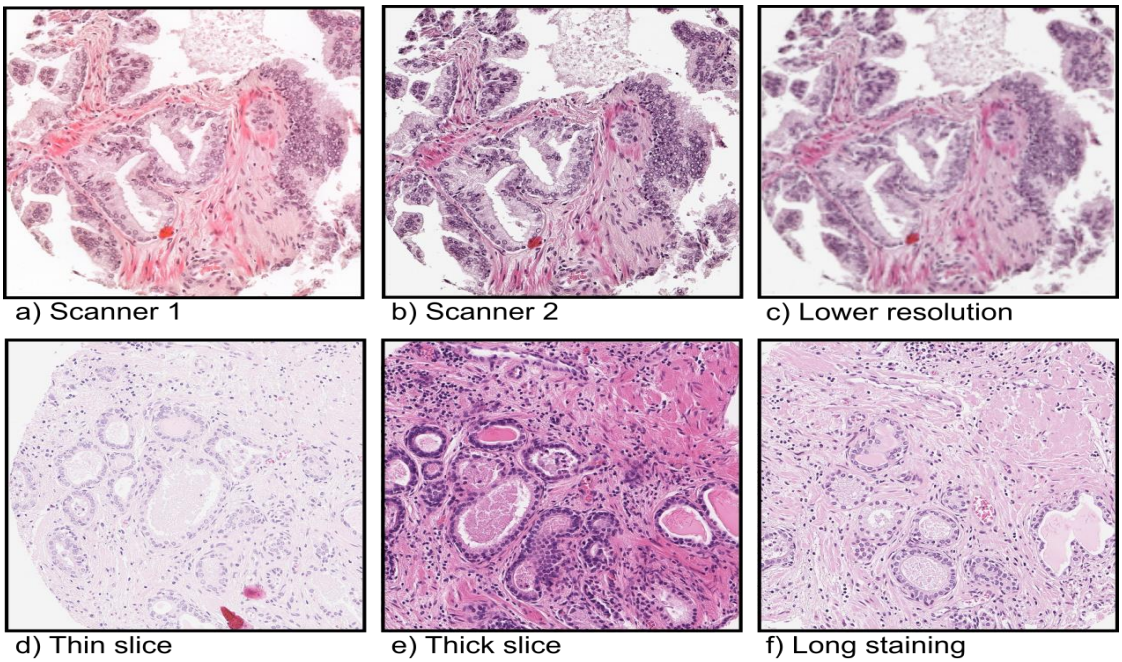

Figure 1 Potential variations in color for H&E-stained prostate tissue samples. a) and b) Visible differences in intensity and color saturation can be caused by using different slide scanner models or c) by the scanner acquired image resolution. d) to f) show typical variations caused by tissue preparation protocols, such as tissue slicing thickness (d & e) and staining duration (f).

Image augmentation and normalization. The most commonly used approaches in clinical histopathology trying to mitigate data variability are color augmentation (Fig. 2b) and stain normalization (Fig. 2c), both of which are used extensively for whole slide H&E images. Both methods aim to reduce stain color variation to assimilate training and test distributions. Recent studies looked at the effects of stain color augmentation and normalization and showed that stain color augmentation drastically improves performance, while normalization is in fact negligible for classification performance[22]. Even though both approaches did find their way into clinical practice they do not fully solve the underlying problem of adapting an unrelated domain. Normalization, which aims at the appearance of images, relies on the selection of meaningful reference slides[23,24]. A general problem with image normalization is, however, also present in histopathology data: spatial features are not taken into account, which can lead to structure colors not being preserved if the selected reference slides did not include the same structure. Data augmentation on the other hand, takes advantage of the stain variation itself by changing the intensity of color during training. Since complete data sets including all possible variations are not available, data augmentation can be used to expand the variety of training images artificially. Image augmentation methods include but are not limited to color adaptations and geometric transforms like flipping or rotation of training images. The degree of change of color intensity can therefore be considered as an additional hyperparameter. Again, a problem with this approach is the variation of unseen data which could, although drastic changes in color are not realistic for histology, lie outside of the color augmentation space used during training.

Color transfer. An ideal histopathological decision support system should provide ways to "push" samples that lie outside the training distribution back into the model domain. This push of test data into the same domain as the training data can be achieved by for example color transfer methods (see Fig. 2d). Histogram matching is a relatively simple technique that is employed to ensure that the color and staining characteristics of images remain consistent with a desired reference by modifying the color and intensity distribution to align it with a reference. The effectiveness of this method heavily depends on the quality and appropriateness of the chosen references. If the reference image itself contains artifacts or inaccuracies or if the number of reference points is too limited, it may propagate these issues to the adapted image or adapt images in an inadequate way[25]. In addition, histogram matching may not take spatial variations in staining intensity or color into account. The loss of spatial information can be critical in cases where local staining patterns are diagnostically relevant. A common method outside the medical field for image-to-image translation are CycleGANs, a type of generative adversarial network for unpaired image-to-image translation[26]. Basically, CycleGANs learn a mapping from the training domain to the test domain. Although this data-driven approach allows CycleGANs to capture a broader range of staining variations and adapt images more effectively while potentially preserving anatomical features, the approach is prone to give confident results that are not justified by the training data and is therefore not simply applicable in the clinical settings.

Obtaining large data sets. In addition to the previously described robustness issues, the cost of collection, management, and storage of PCDP data quickly grows with the number of patients included in a cohort, limiting available sample sizes in practice due to missing or limited infrastructure. At the same time, clinical and research institutions working on the same research problem and willing to join their efforts are not allowed to share sensitive patient data due to data protection regulations. Although first steps have been made by initiatives like *Bigpicture* (https://bigpicture.eu/) funded by the EU Innovative Medicines Initiative trying to enable researchers to share data and dealing with possible regulations in different countries, contemporary data access restrictions still limit the usage of 'big biomedical data' across research sites.

Federated learning. A promising approach to train models on a large amount of data are the classes of Federated Learning and Secure Multi-Party Computation approaches[27–29]. Federated Learning is a privacy-enhancing technique that addresses the problem of data privacy and governance by training collaboratively on e.g., distributed data sets of different sites without exchanging the data itself. It implies collaborative model training on decentralized data owned by multiple participants, which does not disclose it to any other computing party, but exchange model parameters computed from their local data. In Secure Multi-Party Computation, all computations are performed on secret shares of the data distributed across computing parties, such that no party can recover private data belonging to any other party. However, there are several reasons why these approaches are not more widely used. Mainly, institutions need to invest time and resources to

develop the necessary infrastructure and protocols as the implementation is complex and data from different clinics may not be compatible due to the aforementioned lack of data standardization.

Adding synthetic data. Due to the extremely large number of clinics, scanners, personnel training and the lack of a general standard of tissue preparation, using federated learning to incorporate training data from all possible domains is very hard to achieve. Increasing the robustness of the model by training on a complete data set that includes all biases, alone, is therefore not an option. If expanding the existing training data by adding data from other domains is not possible, the generation of synthetic data was proposed as a potential solution[30]. By simulating realistic tissue samples, the existing data set can be expanded and thus enhance the diversity of cases for training of a model drastically (see Fig. 2b). An additional benefit is the development and validation of models without compromising patient privacy. For synthetic data generation Generative Adversarial Networks (GANs) can be seen as the gold standard. However, the use of GANs in histopathology comes with its own challenges. While creation of synthetic images works reasonably well, replicating the level of detail inherent in histopathological WSIs requires the generation of structures on different scale levels (glandular to cell level). For histology data a pyramid of GAN structures was proposed, each responsible for generating a different level of detail[31]. Although the use of GANs for creating synthetic data can help in extending a limited data set, GANs also create synthetic artifacts reducing data quality and therefore require time-consuming manual quality control to be applicable for clinical use. Another drawback is that synthetic data still follows underlying assumptions of data distribution. If these assumptions do not accurately capture all nuances of real world data the problem of previously unseen scenarios is still present and bias is introduced into the AI model.

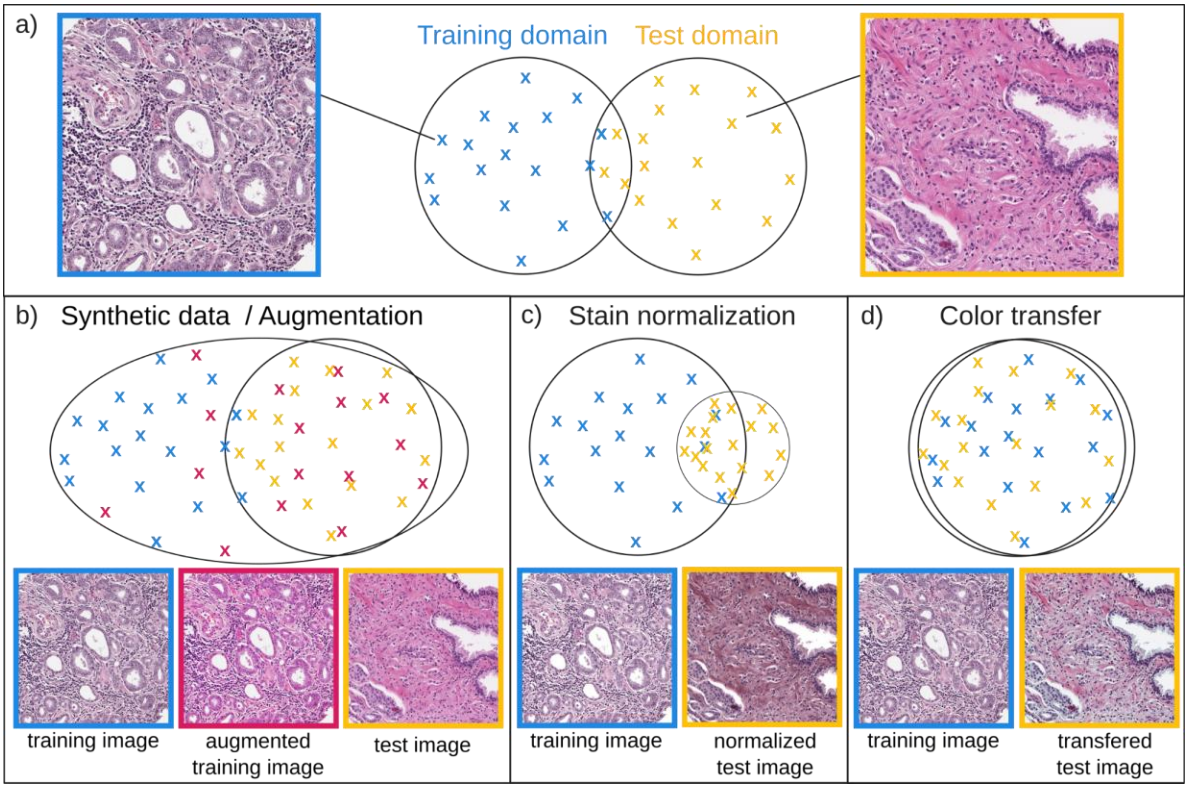

Figure 2 Schematic overview of the three most common adaptation categories to compensate domain shift with exemplary prostate tissue images. a) if the distribution of model training data (blue) has little to no overlap with the unseen test data distribution (yellow) AI model results are unpredictable for this unseen test data. To adjust for domain shift, data can be synthetically augmented (b), data can be normalized to decrease data variance caused for example by color intensity shifts (c), or data can be 'pushed' in distribution by color transfer (d), for instance.

Using additional information. The Gleason grade is per definition based on the glandular architecture of the tissue as visible in the sample. In contrast, an AI may also base its grade on other image components such as the nuclear chromatin texture/structure and changes in the structure of the stroma. It is well known that nuclear chromatin structure carries information about malignancy

grade[32,33]. To what extent the nuclear chromatin patterns add information to the prediction of patient outcome can be studied by image processing operations that detect cell nuclei, segment those and crop out a small region around each nucleus. A reasonably large sample of such nuclear images can be assembled into a stitched image of each patient. Those artificial images can then be used to train and test an AI-based model to predict outcome based only on those images. Similarly cropped images of connective tissue from patients can be assembled and used to train an AI model to predict outcome. It is likely that those models will be less powerful than the model trained on the WSI, but such a study will clarify the extent to which other information available in the tissue image than the glandular architecture that is the basis of Gleason grades contributes to the predictive power. Additional non-image related information such as PSA or other parameters available at the time of diagnosis could also be added to an AI model. Systematic studies of how these different components together with imaging form a reliable basis for AI training are, however, not available.

MODEL ROBUSTNESS

Before a model can be deployed clinically, it should provide a measure of quality of the prediction in an interpretable manner and if possible, adapt to encountered data variations. These model-based challenges and potential solutions will be discussed in the following.

Domain shift. In the medical domain, the basic assumption in AI training, that training and test data follows the same distribution usually does not hold true. This creates a domain shift between the source training data and the target clinical data on which the model is deployed. When this domain shift gets too large, the model might fail on unseen data sets in unpredictable ways. Larger sample sizes with data coming from several domains are therefore highly desired to train a robust and accurate model and by constraining the clinical use case, e.g., to a certain scanner, domain shift can be contained to some extent. However, the general applicability of the model suffers greatly, and some sources of domain shift or bias are still inevitable.

Anomaly detection. Tackling the problem of detection of unseen samples lying outside of the model distribution can be done with out-of-distribution (OOD) detection. Data that lies outside the training distribution is detected and potentially highlighted for the user or excluded from prediction. Most traditional AI models, particularly deep learning models, may produce predictions with very high confidence on unseen data. This bears the risk of misdiagnosis. Conformal prediction can be used to address the issue of assessing the reliability of predictions in clinical applications by quantifying the uncertainty associated with the prediction. In the context of PCDP conformal prediction can assist pathologists and clinicians in making more informed decisions and flagging unreliable predictions of the AI system for human inspection. Olsson et al. showed in a recent study that conformal prediction could, with small sample sizes, detect systematic differences in external data leading to worse predictive performance[34].

Domain-invariant features. A general problem with domain adaptation, approaches that adapt a domain to a preexisting one to deal with domain shift, is that they build a mapping from the training domain to the test domain using fixed feature representations of the domains. Instead of trying to mitigate differences of the different PCDP domains, a promising approach is domain adversarial training. This method tries to use image properties that are both discriminative of the underlying clinical question, e.g., survival prediction of prostate cancer patients, and at the same time domain-agnostic. This can help to avoid issues related to variations in staining intensity, color balance, and other image properties[35,36].

Fine-tuning. A powerful and commonly used technique is fine-tuning a neural network after pretraining it. After initial training, the network is fine-tuned on a target data set, which contains for example out-of-domain staining variations, lighting conditions, or other variations present in the target data set that the network should be adapted to. During fine-tuning, the network's weights are adjusted to better align with the specific characteristics of the target domain. Although fine-tuning provides a certain level of control over the adaptation process, it is important to note that it is a complex process involving a lot of parameters (learning rate, layers to freeze, …) that need additional data from the target domain and careful monitoring to avoid a drop in the model's performance. This method would also need to be reapplied for each new domain that the model is potentially adapted to in the future and is therefore heavily dependent on the target domain.

Interpretability. It is also important that an AI model can show the urologist and pathologist what features in the image material it based its grade on. Trustworthy AI is a key concept here. Using a robust and generalizable model increases trust but neural networks still act as black boxes, not revealing the underlying decision-making processes[37]. The influence of single input features on

a prediction is hard to reconstruct due to the non-linear and complex nature of neural networks, which makes them hard to interpret.

Explainable AI. To overcome this, explanations of what the AI system focuses on during a prediction can be generated to some extent. By doing so, the decision-making process is better comprehensible. For PCDP, image regions that were most relevant for a classification by a model, e.g., glandular structure, can be visualized as heat maps, showing where in the tissue the model found cancer and what image features were used to grade that cancer. This can be achieved by using the attention mechanism of the model. Basically, the attention map of an input image specifies parts of the image that contribute to the decision of the network, making the result more interpretable.

Modeling patient outcome

Current PCDP models mainly focus on segmentation of glands, or directly predict Gleason grades. While these approaches have yielded impressive performance that rival human experts, they rely on subjective human annotations. Even expert pathologist concordance in Gleason grading suffers from high interobserver variability leading to over- or under-treatment[38]. These observations posit that PCDP trained on subjective human annotations will be limited by human performance, while potentially reducing the variability that is observed in human Gleason grading.

In order to achieve objective scores of cancer aggressiveness that do not rely on subjective human annotations, PCDP models can be trained on known outcomes to predict patient survival or the time-to-event. The endpoint for this prediction does not have to be the patient's death, but can also be, for example, the time to relapse. Two studies have recently been published based on a large biobank with prostate tissue samples from 17,700 patients with 10 years of follow up data. Dietrich et al. modeled the exact time of relapse using a Recurrent Neural Network and achieved expert-level performance on the internal validation data, reaching a cumulative dynamic AUC (CDAUC) of 0.77[39]. Walhagen et al. reached similar performance, ROC AUC of 0.79, by predicting recurrence within the first five years post treatment[40]. The results of both studies show the ability to predict cancer outcome with a performance that rivals human experts that use Gleason grading. Interestingly, this could be achieved even with the AI model using much smaller tissue microarray spots (TMAs) for its prediction, while the Gleason grade was based on the whole prostatectomy tissue sample.

While these studies highlight the potential usability to predict patient outcome to obtain objective cancer aggressiveness scores and categories, the robustness of these approaches on external validation data has still to be shown. Furthermore, it is still to be seen if PCDPs that model patient outcomes can in fact surpass human experts in their predictive performance.

Conclusion

In conclusion, this short-review has provided an overview of the challenges and advancements for a potential application of AI for PCDP. We have discussed the work of various research groups trying to overcome the current problems and steering the field into alternative grading of PC samples. Much of that work is not limited to the application in prostate cancer but relevant for possibly all histopathological sample assessment. The progress in the field of AI in medicine and AI-based Gleason grading and prostate cancer aggressiveness prediction in particular are promising and show the way to a potential deployment of advanced grading systems and clinical decision support software. The move away from a subjective scale like Gleason to a more objective survival outcome prediction might pave the way to a personalized approach in the treatment of prostate cancer patients. However, before such systems can be generally adapted in clinical routine, they have to prove their ability to handle the large data variability and that they are able to highlight data outside their training distribution and give an accurate measurement of their confidence in the results. AI models need to be able to communicate on what areas and aspects of the analyzed tissue the decision has been based on. Although the first support systems have been approved by regulatory agencies their application is still limited by the data the system was trained on. Innovative new methods like the learning of domain-invariant features and quantifying the uncertainty with conformal prediction are promising methods to overcome these barriers. By combining collaborative data sharing strategies, establishing robust data standards, and leveraging extensive training on larger datasets, the future of AI in PCDP promises to be exceptionally intriguing.

Conflicts of Interest Statement

P.W., C.B., E. B., and S.B. work part time for Spearpoint Analytics AB, a company developing AI-based digital pathology solutions. The authors declare no other conflicts of interest.

Funding Statement

M.B. was supported by the 3R (Replace, Reduce, Refine) Start-up Funding Program, awarded by the Medical Faculty of the UKE Hamburg in 2021 and FLIGHT (Federated Learning-Guided digital Health) of the Cross-Disciplinary Labs (CDL) funding program at the House of Computing and Data Science awarded by the University of Hamburg, S.B. was supported by DFG FOR 5068.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*. 2021;71(3):209-249. doi:10.3322/CAAC.21660
2. Gleason DF. Classification of prostatic carcinomas. *Cancer Chemother Rep*. 1966;50(3):125-128. https://cir.nii.ac.jp/crid/1573105974852230656. Accessed September 7, 2023.
3. Mellinger GT, Gleason D, Bailar J. The histology and prognosis of prostatic cancer. *J Urol*. 1967;97(2):331-337. doi:10.1016/S0022-5347(17)63039-8
4. Epstein JI, Allsbrook WC, Amin MB, et al. The 2005 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma. *Am J Surg Pathol*. 2005;29(9):1228-1242. doi:10.1097/01.PAS.0000173646.99337.B1
5. Sethi A, Sha L, Vahadane AR, et al. Empirical comparison of color normalization methods for epithelial-stromal classification in H and E images. *J Pathol Inform*. 2016;7(1):17. doi:10.4103/2153-3539.179984
6. 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(2):233-241. doi:10.1016/S1470-2045(19)30739-9
7. 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(9):1372-1380. doi:10.1001/JAMAONCOL.2020.2485
8. 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 Heal*. 2020;2(8):e407-e416. doi:10.1016/S2589-7500(20)30159-X
9. Li W, Li J, Sarma K V., et al. Path R-CNN for Prostate Cancer Diagnosis and Gleason Grading of Histological Images. *IEEE Trans Med Imaging*. 2019;38(4):945-954. doi:10.1109/TMI.2018.2875868
10. Ren J, Sadimin E, Foran DJ, Qi X. Computer aided analysis of prostate histopathology images to support a refined Gleason grading system. *Proc SPIE--the Int Soc Opt Eng*. 2017;10133:101331V. doi:10.1117/12.2253887
11. Bulten W, Bándi P, Hoven J, et al. Epithelium segmentation using deep learning in H&E-stained prostate specimens with immunohistochemistry as reference standard. *Sci Rep*. 2019;9(1). doi:10.1038/S41598-018-37257-4
12. Bukowy JD, Foss H, McGarry SD, et al. Accurate segmentation of prostate cancer histomorphometric features using a weakly supervised convolutional neural network. *J Med imaging (Bellingham, Wash)*. 2020;7(5). doi:10.1117/1.JMI.7.5.057501
13. Rabilloud N, Allaume P, Acosta O, et al. Deep Learning Methodologies Applied to Digital Pathology in Prostate Cancer: A Systematic Review. *Diagnostics*. 2023;13(16):2676. doi:10.3390/DIAGNOSTICS13162676/S1
14. 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(8):1301. doi:10.1038/S41591-019-0508-1
15. Li Y, Huang M, Zhang Y, et al. Automated Gleason Grading and Gleason Pattern Region Segmentation Based on Deep Learning for Pathological Images of Prostate Cancer. *IEEE Access*. 2020;8:117714-117725. doi:10.1109/ACCESS.2020.3005180
16. Arvaniti E, Fricker KS, Moret M, et al. Automated Gleason grading of prostate cancer tissue microarrays via deep learning. *Sci Reports 2018 81*. 2018;8(1):1-11. doi:10.1038/s41598-018-30535-1
17. TH N, S S, V M, et al. Automatic Gleason grading of prostate cancer using quantitative phase imaging and machine learning. *J Biomed Opt*. 2017;22(3):036015. doi:10.1117/1.JBO.22.3.036015
18. Bulten W, Kartasalo K, Chen PHC, et al. Artificial intelligence for diagnosis and Gleason

grading of prostate cancer: the PANDA challenge. *Nat Med 2022 281*. 2022;28(1):154-163. doi:10.1038/s41591-021-01620-2

19. Sandeman K, Blom S, Koponen V, et al. AI Model for Prostate Biopsies Predicts Cancer Survival. *Diagnostics (Basel, Switzerland)*. 2022;12(5). doi:10.3390/DIAGNOSTICS12051031
20. 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(8):1588-1595. doi:10.1038/S41379-021-00794-X
21. Jung M, Jin MS, Kim C, 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. *Mod Pathol*. 2022;35(10):1449-1457. doi:10.1038/S41379-022-01077-9
22. Tellez D, Litjens G, Bándi P, et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med Image Anal*. 2019;58:101544. doi:10.1016/J.MEDIA.2019.101544
23. Macenko M, Niethammer M, Marron JS, et al. A method for normalizing histology slides for quantitative analysis. *Proc - 2009 IEEE Int Symp Biomed Imaging From Nano to Macro, ISBI 2009*. 2009:1107-1110. doi:10.1109/ISBI.2009.5193250
24. Reinhard E, Ashikhmin M, Gooch B, Shirley P. Color transfer between images. *IEEE Comput Graph Appl*. 2001;21(5):34-41. doi:10.1109/38.946629
25. Khan AM, Rajpoot N, Treanor D, Magee D. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE Trans Biomed Eng*. 2014;61(6):1729-1738. doi:10.1109/TBME.2014.2303294
26. Zhu JY, Park T, Isola P, Efros AA. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *Proc IEEE Int Conf Comput Vis*. 2017;2017-October:2242-2251. doi:10.1109/ICCV.2017.244
27. Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. *NPJ Digit Med*. 2020;3(1). doi:10.1038/S41746-020-00323-1
28. Wolff J, Matschinske J, Baumgart D, et al. Federated machine learning for a facilitated implementation of Artificial Intelligence in healthcare - a proof of concept study for the prediction of coronary artery calcification scores. *J Integr Bioinform*. 2022;19(4). doi:10.1515/JIB-2022-0032
29. Narmadha K, Varalakshmi P. Federated Learning in Healthcare: A Privacy Preserving Approach. *Stud Health Technol Inform*. 2022;294:194-198. doi:10.3233/SHTI220436
30. Xue Y, Ye J, Zhou Q, et al. Selective synthetic augmentation with HistoGAN for improved histopathology image classification. *Med Image Anal*. 2021;67:101816. doi:10.1016/J.MEDIA.2020.101816
31. Li W, Li J, Polson J, Wang Z, Speier W, Arnold C. High resolution histopathology image generation and segmentation through adversarial training. *Med Image Anal*. 2022;75:102251. doi:10.1016/J.MEDIA.2021.102251
32. Li J, Ettel M, Amin A, et al. Interobserver Reproducibility of Quantifying Gleason Pattern 4 Cancer in Prostate Biopsy: Implications for Clinical Practice. *http://www.xiahepublishing.com/*. 2023;3(1):4-9. doi:10.14218/JCTP.2022.00026
33. Veltri RW, Marlow C, Khan MA, Miller MC, Epstein JI, Partin AW. Significant variations in nuclear structure occur between and within Gleason grading patterns 3, 4, and 5 determined by digital image analysis. *Prostate*. 2007;67(11):1202-1210. doi:10.1002/PROS.20614
34. Olsson H, Kartasalo K, Mulliqi N, et al. Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction. *Nat Commun 2022 131*. 2022;13(1):1-10. doi:10.1038/s41467-022-34945-8
35. Wilm F, Marzahl C, Breininger K, Aubreville M. Domain Adversarial RetinaNet as a Reference Algorithm for the MItosis DOmain Generalization Challenge. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)*. 2022;13166 LNCS:5-13. doi:10.1007/978-3-030-97281-3_1/FIGURES/4
36. Lafarge MW, Pluim JPW, Eppenhof KAJ, Moeskops P, Veta M. Domain-Adversarial Neural Networks to Address the Appearance Variability of Histopathology Images. In: Cardoso MJ, Arbel T, Carneiro G, et al., eds. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham: Springer International Publishing; 2017:83-91.
37. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med 2022 281*. 2022;28(1):31-38. doi:10.1038/s41591-021-01614-0

38. Ozkan TA, Eruyar AT, Cebeci OO, Memik O, Ozcan L, Kuskonmaz I. Interobserver variability in Gleason histological grading of prostate cancer. *new pub Med Journals Sweden AB*. 2016;50(6):420-424. doi:10.1080/21681805.2016.1206619
39. Dietrich E, Fuhlert P, Ernst A, et al. Towards Explainable End-to-End Prostate Cancer Relapse Prediction from H&E Images Combining Self-Attention Multiple Instance Learning with a Recurrent Neural Network. *Proc Mach Learn Res*. 2021;158:38-53. https://proceedings.mlr.press/v158/dietrich21a.html. Accessed September 7, 2023.
40. Walhagen P, Bengtsson E, Lennartz M, Sauter G, Busch C. AI-based prostate analysis system trained without human supervision to predict patient outcome from tissue samples. *J Pathol Inform*. 2022;13:100137. doi:10.1016/J.JPI.2022.100137