AI reality check when evaluating difficult to grade prostate cancers

Liron Pantanowitz¹ · Rohit Mehra¹ · L. Priya Kunju¹

Received: 13 January 2021 / Revised: 13 January 2021 / Accepted: 24 January 2021 / Published online: 2 February 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH, DE part of Springer Nature 2021

With the perpetual digitization of pathology glass slides, globally large archives of whole slide images have been amassed. Such digital datasets have, in turn, accelerated the development of machine and deep learning algorithms in anatomical pathology. One area, in particular, that has been at the forefront of this advance in computational pathology is prostate pathology. This is not surprising given the need to meet the demand for constant growth in prostate core needle biopsies (CNB) encountered in clinical practice. In this context, artificial intelligence (AI)-based tools have been designed to perform several diagnostic tasks typically undertaken by pathologists using their light microscopes. These include the detection of prostate adenocarcinoma, Gleason grading, quantification of tumor volume, and identification of perineural invasion [1, 2]. The benefits of deploying these AI algorithms in routine pathology practice include potential improvement in diagnostic accuracy as well as automation to reduce pathologist workload and likely shorter sign-out turnaround time. Another boon of utilizing computer-aided systems in practice is standardization and possibly reducing inter-observer variability of Gleason grading [3], a welcome benefit given the current inter-pathologist variability reported with grading prostate adenocarcinoma [4, 5].

In a recent publication in *Virchows Archiv*, Egevad and colleagues eloquently shared their findings after challenging an AI-assisted grading system with prostate CNB enriched with cases that had problematic grading of adenocarcinoma [6]. Their cases were selected from the reference Pathology Imagebase (<https://isupweb.org/pib/>) hosted by the International Society of Urological Pathology (ISUP). This online image repository, where photomicrographs from cancer cases are independently vetted and graded by 23

international experts, is used for education and to promote the international standardization of grading. From this reference dataset, the authors selected glass slides of 87 CNB containing problematic or challenging to grade prostatic adenocarcinoma cases. Not surprisingly, in such challenging to grade cases, the genitourinary pathology experts failed to reach a two-thirds consensus in 36 (41.4%) of their chosen cases. While the weighted kappa was 0.77 (range 0.68–0.84) for consensus cases, it was only 0.50 (range 0.40–0.57) for these non-consensus selected cases.

The glass slides from the aforementioned problematic CNB cases were then digitized using an Aperio ScanScope AT2 scanner and subsequently graded using an AI algorithm trained in Gleason grading. Details regarding the development of this AI system were previously published [7]. The deep learning convolutional neural networks of this ensembled AI system were originally trained largely on digitized slides from CNB randomly selected from a Swedish (STHLM3) diagnostic study. None of the 87 problematic cases from the study reported by Egevad et al. was included in the original dataset used for training or validation of the AI algorithm. Previously, the developers of the AI system reported that their algorithm assigned grades that yielded a mean pairwise kappa of 0.62, which was within the range of matching values for the expert panel of pathologists (0.60–0.73) [7]. However, when Egevad et al. applied this AI algorithm to the non-consensus cases selected for their newer series, the AI algorithm only attained a weighted kappa value of 0.53. Compared to humans who were inconsistent at grading these difficult cases, the reproducibility of the AI system was slightly better in this non-consensus group.

While the AI system employed in this study for grading prostate cancer previously demonstrated high performance, it is disappointing that when challenged to grade problematic cancers the results were mediocre. Egevad et al. argue that AI is thus probably no better than an expert uropathologist in such cases. They also correctly state that AI is likely to be as good as the training set based upon which it has been developed. Furthermore, it appears that other investigators employing similar deep learning algorithms to grade prostate adenocarcinoma have experienced equal difficulty achieving

All authors contributed equally to this commentary.

✉ Liron Pantanowitz
lironp@med.umich.edu

¹ Department of Pathology, University of Michigan, NCRC Bldg 35, 2800 Plymouth Road, Ann Arbor, MI 48109-2800, USA

optimal performance. For example, Arvaniti et al. reported a kappa of 0.55 for Gleason grading [8], Nir et al. reported an accuracy of 79% (with sensitivity = 75.9% and specificity = 77.9%) [9], and Nagpal et al. reported an area under the curve (AUC) of 0.7 for grading prostate adenocarcinoma [10]. While some studies demonstrate high accuracy (90%) in differentiating Gleason pattern 4 and 5 from Gleason pattern 3, the datasets are frequently small [11]. If AI algorithms are going to be adopted in actual clinical practice in pathology laboratories around the world, they will need to perform better, be better trained on larger datasets with higher complexity and contain more unusual variants of cancer, demonstrate suitable generalizability, and require calibration or fine-tuning to each setting. It is well known that prostate cancer grading employs rules based upon inherently subjective morphologic observations. Moreover, training algorithms to grade prostate cancer based on rules which are evolving makes this a uniquely challenging process. Hence, clinical grade computational pathology tools will need to constantly update their learning techniques in order to decrease inter-observer variability and improve their value for general pathologists.

A true test of an AI system is whether it ultimately improves patient care in terms of diagnosis, management, and disease-specific outcome. In this context, Egevad et al. rightfully emphasize that AI algorithms in pathology will be instrumental in offering objective improvement of prostate cancer management if they are trained not just to simulate tasks diagnostic pathologists perform today with light microscopy, but also incorporate genetic and clinical outcome variables in their algorithms which could improve predictions and better guide patient treatment plans. Finally, it is our hope that, by deploying AI in clinical practice, this will eventually also elevate pathologists' expertise in handling complex cases, including difficult to grade prostate cancer cases.

Declarations No IRB was needed.

Conflict of interest Liron Pantanowitz is on the medical advisory board of Ibex.

References

1. Parwani AV (2019) Automated diagnosis and Gleason grading of prostate cancer - are artificial intelligence systems ready for prime time? *J Pathol Inform*. 10:41
2. Pantanowitz L, Quiroga-Garza GM, Bien L, Heled R, Laifenfeld D, Linhart C, Sandbank J, Albrecht Shach A, Shalev V, Vecsler M, Michelow P, Hazelhurst S, Dhir R (2020) Clinical validation and deployment of an AI-based algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies. *Lancet Digital Health* 2(8):e407–e416
3. Marginean F, Arvidsson I, Simoulis A, Overgaard NC, Astrom K, Heyden A et al. An artificial intelligence-based support tool for automation and standardisation of Gleason grading in prostate biopsies. *Eur Urol Focus* 2020; Dec 7; S2405–4569(20)30296-0.
4. Engers R (2007) Reproducibility and reliability of tumor grading in urological neoplasms. *World J Urol*. 25(6):595–605
5. Fine SW, Trpkov K, Amin MB, Algaba F, Aron M, Baydar DE et al (2020) Practice patterns related to prostate cancer grading: results of a 2019 Genitourinary Pathology Society clinician survey. *Urol Oncol* S1078-1439(20):30396–30393
6. Egevad L, Swanberg D, Delahunt B, Ström P, Kartasalo K, Olsson H, Berney DM, Bostwick DG, Evans AJ, Humphrey PA, Iczkowski KA, Kench JG, Kristiansen G, Leite KRM, McKenney JK, Oxley J, Pan CC, Samarasinghe H, Srigley JR, Takahashi H, Tsuzuki T, van der Kwast T, Varma M, Zhou M, Clements M, Eklund M (2020) Identification of areas of grading difficulties in prostate cancer and comparison with artificial intelligence assisted grading. *Virchows Arch*. 477(6):777–786
7. Ström P, Kartasalo K, Olsson H, Solorzano L, Delahunt B, Berney DM et al (2020) Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *Lancet Oncol*. 21(2):222–232
8. Arvaniti E, Fricker KS, Moret M, Rupp N, Hermanns T, Fankhauser C et al (2018) Automated Gleason grading of prostate cancer tissue microarrays via deep learning. *Sci Rep*. 8(1):12054
9. Nir G, Hor S, Karimi D, Fazli L, Skinnider BF, Tavassoli P et al (2018) Automatic grading of prostate cancer in digitized histopathology images: learning from multiple experts. *Med Image Anal*. 50:167–180
10. Nagpal K, Foote D, Liu Y, Chen PC, Wulczyn E, Tan F et al (2019) Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ Digit Med*. 2:48
11. Lucas M, Jansen I, Savci-Heijink CD, Meijer SL, de Boer OJ, van Leeuwen TG et al (2019) Deep learning for automatic Gleason pattern classification for grade group determination of prostate biopsies. *Virchows Arch*. 475(1):77–83

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.