Clinical deployment of AI for prostate cancer diagnosis

In *The Lancet Digital Health*, Liron Pantanowitz and colleagues[1] evaluate an artificial intelligence (AI) system for detecting prostate cancer in whole slide images of core needle biopsies. Their approach localises and quantifies cancer area in addition to providing clinically relevant Gleason groupings, along with perineural invasion identification—a challenge often overlooked in similar studies. Their work is validated across two scanners and institutions, providing insights into how these algorithms perform in the presence of potential batch effects and other pre-analytic variation. The resulting study is distinguished by its real-world evaluation, showing how computer-aided diagnosis (CAD) tools might influence pathology practice in the near future.

The developed CAD system contains a deep learning model trained on 138 cases (a total of 549 haematoxylin and eosin [H&E]-stained slides) from a single hospital and tested on an internal test set of 210 consecutive cases (2501 slides), as well as an external validation set of 100 consecutive cases (1627 H&E-stained slides).[1] Quantitative performance metrics of the CAD system were excellent, with a cancer detection sensitivity of 99·59% (95% CI 98·39–99·90) and specificity of 90·14% (87·76–92·09) on the internal test set. The model was calibrated on slides from 32 separate cases from the external validation site, after which it achieved similarly high sensitivity (98·46%, 94·06–99·61) and specificity (97·33%, 94·43–98·74) on the external validation set.

The practical effects of this CAD system on a clinical diagnostic workflow were evaluated through deployment as a second reader over a 20-month time period at a large health-care provider's centralised pathology centre.[1] During that interval, the system analysed 941 cases (11 429 H&E-stained slides) and triggered alerts when the model predicted a high cancer score versus a benign diagnosis by the pathologist (ie, false negative) or a high algorithmic Gleason score diagnosed as a lower score by the pathologist. These alerts reflect diagnostic discrepancies justifying treatment plan modification and an example is given of the first case detected by the algorithm where cancer had been missed by the pathologist. Interestingly, this computational second reader approach afforded the opportunity to quality control the entire case load with minimal additional pathologist effort, as opposed to the substantial human effort required for the random sub-sampling quality control proposed by laboratory guidelines.[2]

While studies discussing the effects of clinically deploying CAD pathology tools remain sparse, a second reader is likely to be the least disruptive way to include CAD in routine clinical practice. Due to rapid development of the CAD space as a result of favourable legal stances only recently taken by the USA and the EU,[3,4] questions regarding confirmation and automation biases of human experts in computer-aided workflows have not yet had the opportunity to be sufficiently investigated.[5] Additionally, concerns regarding alert fatigue in the case of poorly performing systems might be warranted[6] and thus require dedicated long-term study. The approach taken by Pantanowitz and colleagues minimises reader biasing by initially blinding pathologists to algorithm outputs and relies on the system's high sensitivity to prevent excessive alerts, both of which are required to make the system an aid rather than a hindrance. The review process is then made easier by having the algorithm specify the slide regions driving its diagnosis, efficiently drawing a pathologist's attention to potential artifacts (eg, crushed glands) or challenging areas (eg, small perineural invasions).

In the use of their model for Gleason grading, the authors rightfully point out the difficulty of reaching Gleason grade agreement.[7] They have instead opted to focus on solely the detection of Gleason pattern 5 and abstract away remaining individual patterns into the clinically relevant categories of Gleason score 6 versus Gleason score 7–10. This reduced dependence on precise Gleason scoring, which has historically complicated development and validation of automated grading approaches, helps to drive the high concordance metrics reported (area under the receiver operating characteristic curve >0·94). Additionally, using machine learning to aid in the standardisation of reporting shows how AI could complement pathologists by completing tasks that have only moderate human reproducibility such as cancer grading and area quantification, but are well suited to a tireless, objective algorithm.[8] In this manner, a glimpse into the future is provided wherein algorithms subtly augment pathologists in their existing workflows, while remaining minimally disruptive.

In light of the paucity of pathologists worldwide, along with increasing workloads, it is important to consider patient groups who would most benefit from CAD. While Pantanowitz and colleagues used specialised genitourinary pathologists for validation,[1] access to such specialists is not universal. Algorithms capable of elevating general pathologists' performance to that of highly trained specialists, in particular drawing attention to cases that might warrant specialist review, have the potential to substantially raise the standard of care. This is especially true in areas with pathologist shortages and further enhanced when combined with telemedicine for remote evaluation. While these algorithms will require complex and often expensive infrastructure for the digitisation and storage of slides, as additional CAD algorithms are validated and deployed, the benefit becomes increasingly justifiable. Notably, the development and validation rate of these algorithms is expected to rapidly grow thanks to increased open-source sharing of code[9] and digital pathology datasets,[10] along with the crowdsourcing of annotations.[11]

Pantanowitz and colleagues[1] should be commended for translating a series of well established deep learning best practices into a useful tool for clinical practice. Over a longer time period, through collection of the tool's output along with paired clinical data, it will be possible to study the benefits of increased precision in pathological measurements enabled by these algorithms. In the interim, these types of blinded clinical validation studies provide a blueprint for the development and validation of CAD approaches. This study has contributed to filling gaps in the literature, including evaluating the performance of these algorithms in a realistic setting and discussing the investment in infrastructure and training needed to realise the promise of CAD systems.

*Andrew Janowczyk, Patrick Leo, Mark A Rubin
andrew.janowczyk@case.edu

Case Western Reserve University, Department of Biomedical Engineering, Cleveland, OH 44106, USA (AJ, PL); Department of Oncology, Lausanne University Hospital and Lausanne University, Lausanne, Switzerland (AJ); Department for BioMedical Research (MAR) and Bern Center for Precision Medicine (MAR), University of Bern and Inselspital, Bern, Switzerland

1   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 Digital Health* 2020; **2:** e407–16.
2   Nakhleh RE, Bekeris LG, Souers RJ, Meier FA, Tworek JA. Surgical pathology case reviews before sign-out: a College of American Pathologists Q-Probes study of 45 laboratories. *Arch Pathol Lab Med* 2010; **134:** 740–43.
3   US Food and Drug Administration. FDA allows marketing of first whole slide imaging system for digital pathology. April 12, 2017. https://www.fda.gov/news-events/press-announcements/fda-allows-marketing-first-whole-slide-imaging-system-digital-pathology (accessed June 29, 2020).
4   García-Rojo M, De Mena D, Muriel-Cueto P, Atienza-Cuevas L, Domínguez-Gómez M, Bueno G. New European Union regulations related to whole slide image scanners and image analysis software. *J Pathol Inform* 2019; **10:** 2.
5   Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf* 2019; **28:** 231–37.
6   Ancker JS, Edwards A, Nosal S, Hauser D, Mauer E, Kaushal R, with the HITEC Investigators. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. *BMC Med Inform Decis Mak* 2017; **17:** 36.
7   Allsbrook WC, Mangold KA, Johnson MH, Lane RB, Lane CG, Epstein JI. Interobserver reproducibility of Gleason grading of prostatic carcinoma: general pathologist. *Hum Pathol* 2001; **32:** 81–88.
8   Pavlides M, Birks J, Fryer E, et al. Interobserver variability in histologic evaluation of liver fibrosis using categorical and quantitative scores. *Am J Clin Pathol* 2017; **147:** 364–69.
9   Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. *J Pathol Inform* 2016; **7:** 29.
10   Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging* 2013; **26:** 1045–57.
11   Alialy R, Tavakkol S, Tavakkol E, et al. A review on the applications of crowdsourcing in human pathology. *J Pathol Inform* 2018; **9:** 2.