Overcoming the challenges to implementation of artificial intelligence in pathology

Jorge S. Reis-Filho, MD PhD FRCPath (1) and

Jakob Nikolas Kather, MD MSc (2, 3, 4)

(1) Experimental Pathology, Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY, USA.

(2) Department of Medicine I, University Hospital and Faculty of Medicine, Technical University Dresden, Dresden, Germany.

(3) Else Kroener Fresenius Center for Digital Health, Technical University Dresden, Dresden, Germany.

(4) Pathology & Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, United Kingdom.

Correspondence:

Jakob Nikolas Kather, MD, MSc, Else Kroener Fresenius Center for Digital Health, Technical University Dresden, Fetscherstrasse 74, 01309 Dresden, Germany. jakob-nikolas.kather@alumni.dkfz.de

Abstract

Pathologists worldwide are facing remarkable challenges with the increasing workloads and the lack of time to provide consistently high-quality patient care. The application of artificial intelligence (AI) to digital whole slide images has the potential of democratizing the access to expert pathology and affordable biomarkers, by supporting pathologists in the provision of timely and accurate diagnosis as well as supporting oncologists by extracting prognostic and predictive biomarkers directly from tissue slides. The long-awaited adoption of AI in pathology, however, has not materialized, and the transformation of pathology is happening at a pace that is much slower than that observed in other fields (e.g., radiology). Here, we provide a critical summary of the developments in digital and computational pathology in the last ten years, outline key hurdles and ways to overcome them, and provide a perspective for AI-supported precision oncology in the future.

Keywords: Artificial intelligence, deep learning, machine learning, pathology, oncology

Introduction

In the last ten years, artificial intelligence (AI) has been demonstrated to be a useful tool in histopathology image analysis.[1,2] In particular, AI can extract much information directly from hematoxylin and eosin (H&E) stained sections. Multiple approaches have been tested for the deployment of AI in pathology, including “strongly” supervised approaches, which emulate what pathologists do, and weakly supervised approaches which, theoretically, can go above and beyond what pathologists do.[3] Strongly supervised AI methods are mostly used for automation, can reduce variability in cancer typing and grading and automate immunohistochemistry scoring, and, thus, can help pathologists arrive at more precise and consistent diagnoses. Weakly supervised methods can use AI to predict a ground truth with is derived from the tissue slide itself - for example, predicting the, the presence of prostate cancer from slides by using just a single label per slide, which can be affected by the subjectivity or lack of precision of a given diagnosis.[4,5] Exceeding this, weakly supervised AI can be trained on an orthogonal ground truth, for example, information derived from molecular diagnostics or clinical follow-up. Hence, weakly supervised AI can define new biomarkers: it can predict genetic alterations[6,7] and clinical endpoints[1,3] – tasks that are currently not routinely possible for pathologists (**Figure 1A**).

The promise of AI in pathology

On the surface, AI is ready for prime time; however, in reality, limitations of AI have hindered its broad adoption (**Figure 1B**). Despite the enthusiasm with the utilization of digital pathology and AI, why has AI not become a reality yet? Here, we discuss what we perceive to be key limitations to the transformative potential of AI in pathology and potential strategies to overcome them.

Challenges in the adoption of AI in pathology

Paradigm shifts

The first challenge is conceptual and cultural, given that the adoption of AI in pathology requires two fundamental paradigm shifts, namely the introduction of digital pathology for diagnosis and pathological assessment of cancers, as well as the transition from a human-based diagnosis/assessment system to one where AI will render the final diagnosis or provide the final results for a given biomarker. New technologies require the discontinuation of established practices, and this can cause distress for users. For example, the introduction of microscopes for the diagnosis and characterization of diseases was

met with considerable resistance by physicians, perhaps best exemplified by the great microscopy debate that took place in the Paris Academy of Medicine in the 19th Century.[8] Similarly, abolishing the traditional microscope and moving to routine digitalization of glass slides was successful in a few institutions in the 2010s, but is still met with broad resistance. Even the first step for digitalization of pathology, the transition from a traditional histology workflow to a "radiologist-like" workflow, in which the user looks at images on a computer screen, is still not a reality yet. And indeed, why should pathologists move from microscopes to computer screens if the current workflows are inexpensive and effective, and the training of new pathologists is primarily based on micoscope-based diagnoses? Digital pathology measuring tools and remote work are strong incentives, but the ultimate incentive could be the development of AI-based biomarkers. Once clinical evidence supports the predictive and prognostic power of these biomarkers, and clinical guidelines recommend AI biomarkers, pathology departments will, inevitably, have to become digital, otherwise assays essential for patient care will not be available. Hence, we contend that the evaluation and, ultimately, the validation of AI biomarkers in samples from prospective clinical trials will likely serve as a catalyst for the digitalization of histopathology. At present, however, access to the algorithms being developed is limited, given the limited digitalization of pathology. In some countries such as Sweden, the UK or the Netherlands, large-scale efforts are underway or have been completed to digitize most large pathology departments. In many other countries, digitalization of pathology is not yet a national priority and has not begun on a large scale.

Quality control, biases and ground truth

A second challenge is related to the quality and diversity of the source data (Figure 1C). Tissue fixation, cutting and staining procedures vary between laboratories and cause differences in morphology. This heterogeneity of input data is a challenge for AI methods. There are two fundamental approaches to address this. The traditional approach holds up the "garbage in, garbage out" paradigm: according to this approach, pre-analytical and data handling workflows should be standardized perfectly. However, this is not always possible, for example, whenever algorithms are trained based on subjective ground truth data. An alternative to striving for perfect standardization is to accept the diversity of pathology slides and to accept some diversity in the ground truth labels and train large models on diverse data. An intermediate way is to accept varying quality of training data, but to mandate local calibration of the AI model at every institution to ensure the data are "in domain". Many "weakly supervised" AI training

methods require training on thousands of slides and are, therefore, more data intensive than the traditional "strongly supervised" approaches.[3,4] Computational methods to augment data are helpful, including style transfer[9] or other synthetic data generation methods.[10,11] In addition, federated learning[12] and swarm learning[13] are emerging technologies that can help algorithms to access sufficiently diverse training data. Subtler, and, possibly, more important issues emerge during the process of training the AI model, and include overfitting, systematic biases, performance drift, and an imperfect ground truth.[14–16] These can be immensely challenging to detect but do adversely influence the performance of AI systems, even leading to undetected AI malpractice over long periods. There is no universal remedy for these, but adherence to Good Machine Learning Practices[17] during development of AI methods helps to mitigate some of the risk. The single most important measure is to gather empirical evidence for the generalization of AI systems on external cohorts representing different patient populations. Also, it is important that end users critically evaluate the results of AI assays, like they do with any diagnostic assay, and place it in context with information obtained through other techniques. For example, clinically approved methods for cancer detection in pathology slides are developed to assist pathologists, but in case of a discordance between the AI model and the human pathologist, the pathologist makes the final decision.

Validity as a biomarker

A third challenge is the technical validity of AI assays, which should be considered de facto biomarkers. Biomarkers constitute a characteristic that is objectively measured/ evaluated as an indicator of normal biological/ physiological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention, and must be fit for their intended purpose. AI biomarkers clearly fall under this definition. For example, in breast cancer, AI has been used to classify benign vs atypia vs ductal carcinoma in situ, to predict hormone receptor and HER2 status, PAM50 subtypes and other genomic features as well as outcome directly from routine pathology slides.[7,18,19] Similarly in colorectal cancer, AI has been used to predict microsatellite instability (MSI)[7,20,21] and other genetic alterations[22], as well as molecular subtypes[23] and outcomes[24] from H&E slides. Some studies have expanded these findings to any tumor type in a "pan-cancer" approach[18,25,26]. Whilst these biomarkers do not reach perfect concordance with the ground truth methods, and, therefore, cannot replace current sequencing methods completely, they can be used as pre-screening tests to reduce the load of molecular tests.[27,28] Independently of their

position in a diagnostic cascade, however, AI biomarkers need to be assessed with exactly the same rigor as "traditional" biomarkers. Many academic studies of AI biomarkers, however, are based on small datasets and/ or the analysis of tissue microarrays, utilize the digital whole slide images from The Cancer Genome Atlas as the primary dataset, and/or report incomplete performance metrics which can obscure deficiencies.[29] Even worse, when an AI method is turned into a commercial product, it does not even necessarily have to demonstrate a high generalizability in order to be approved for clinical routine use. Reaching a demonstrable “clinical-grade” performance requires training data in the order of thousands to tens of thousands of patients.[4,30] The most fundamental piece of evidence required to demonstrate the robustness is a true external validation (i.e. an application of the trained model to a dataset which is completely independent of the training dataset).[5] Another fundamental requirement for AI biomarkers is reproducibility. This has increasingly come into the focus of computational pathology research[5,31] and several large-scale studies have evaluated AI systems on multiple cohorts, demonstrating its clinical validity. An important element of the analytical validity of AI algorithms is to test its reproducibility at the level of deployment, including endeavors testing the accuracy, reproducibility and consistency for the deployment of the algorithm for its use by an individual pathologist, by different pathologists at the same institution and by different pathologists at other institutions.[20,24,32] We argue that like any biomarker, AI biomarkers need empirical proof demonstrating that they are fit-for-purpose in all intended use cases. We would also contend that the lessons learnt in the process of incorporating genomics biomarkers[33–35] could serve as a framework for the assessment of the analytical validity, clinical validity and clinical utility of AI-based biomarkers. This should be combined with the development of levels of evidence for the validity of this new category of biomarkers.

Regulatory approval

A fourth challenge is the complexity and the rapid changes in regulatory approval. For AI-based diagnostic assays to be used clinically, they must go through the regulatory process for medical devices. This process differs between the United States, the EU, and other large markets. In the EU, the relevant rule set since May 2020 is the medical device regulation (MDR) and the In Vitro Diagnostic Medical Devices Regulation (IVDR) while in the US, the relevant ruleset for any laboratory test is defined in the Clinical Laboratory Improvement Amendments (CLIA) statute. Due to the involved nature of regulatory

processes, the clinical deployment of AI methods developed by academic groups is remarkably challenging; in fact, partnerships with an existing company or the development of a 'spinoff' company from academic groups are approaches rather commonly being employed.[36] It is increasingly clear, however, that obtaining regulatory approval does not mean that the algorithm is actually being used in clinical routine or will result in clinical adoption. Several companies struggle to commercialize their AI methods, leading to a plethora of "orphan" products which have been formally approved, but not incorporated in pathology practice. Without the approval of algorithms that can de facto improve pathology practice (not incremental improvements), these approvals may not translate into wide adoption of algorithms. We argue that only by the approval of transformative AI-based biomarkers, there would be a clear incentive for pathology departments to undergo the required digital transformation that will ultimately enable the adoption of AI in pathology. Germane to the successful adoption of AI algorithms in pathology is clarity in terms of the type of regulatory approval sought, as well as in regard to the required levels of analytical validity for the use of these algorithms as laboratory developed tests (LTDs).

Financial challenges

Converting a pathology laboratory to digitized workflows is costly. Hardware cost and setting up incurs a high fixed cost, while slide scanning and backing up data incur a smaller but persistent variable cost. Furthermore, fixed costs repeat every couple of years as devices reach their expiry date and a new technical generation of devices becomes available. It is important to consider, but still mostly unclear, how these technologies will be priced and whether they will be covered by insurance. Unlike many other laboratory equipment (e.g., massively parallel sequencers) where the actual costs of the hardware are included as part of the cost of the consumables needed, whole slide scanners at present require an initial investment. From the perspective of health insurance providers or single payers, reimbursement of AI technologies will depend on their potential to reduce costs, improve clinical trial evaluation as well as patient outcomes. Ideally, we would quantify how many pathologist-hours automatic scoring systems can save or how many life years are gained by a treatment informed by an AI biomarker compared to the standard of care. These measurements, however, are difficult to obtain in an unbiased manner, and, in their absence, it is unclear how AI-based diagnostic assays and biomarkers should be priced. Conversely, however, AI provides a unique opportunity to deliver expert pathology, with algorithms benchmarked against the top experts in the field or orthogonal data, and to

democratize access to biomarkers in the context of healthcare systems with less abundant resources. In fact, good performance of AI biomarkers can be achieved just with a basic microscope and a mobile phone, illustrating the potential of these approaches in providing equity and inclusion for diagnostic pathology and biomarker assessment in more remote and less affluent regions.[37]

Outlook

In 2012, Deep neural networks beat any previous handcrafted technology in image processing, and this trend has been a reality in medical image processing since 2017. Hence, the last ten years have been regarded as an inflexion point for AI, and almost as a plateau, with the task being to find new use cases for a technology which was essentially mature. 2021 and 2022 have revealed that the technological aspects of AI are still expected to evolve massively (**Figure 1D**). In particular, the zero-shot capabilities of large language models (LLMs) or diffusion models (DMs) for data generation have yielded astonishing successes, and the commercial and societal disruption resulting from the surrounding software ecosystem is expected to be transformative. It seems plausible that this technological advance will spill over to pathology and lead to previously unimaginable use cases.[10] Diagnostic pathology, however, still seeks to find solutions for the successful implementation of the 2012-2022 generation of AI systems. Hence, it becomes even more important that solutions to these challenges are enacted, and that this process ought to be driven by medical expertise and patient benefit, ultimately resulting in the latest AI technologies being sensibly applied for the benefit of patients and caregivers.

Additional information

Data Availability statement

No research data was used for this article.

Disclosures

JSRF reports a leadership (board of directors) role at Grupo Oncoclinicas, stock or other ownership interests at Repare Therapeutics and Paige.AI, and a consulting or Advisory Role at Genentech/Roche, Invicro, Ventana Medical Systems, Volition RX, Paige.AI, Goldman Sachs, Bain Capital, Novartis, Repare Therapeutics, Lilly, Saga Diagnostics and Personalis. JNK reports consulting services for Owkin, France; Panakeia, UK and DoMore Diagnostics, Norway and has received honoraria for lectures by MSD, Eisai and Fresenius.

Author contributions

JSRF and JNK jointly conceived and wrote this article.

Funding

JSRF is funded in part by the Breast Cancer Research Foundation, a Susan G Komen Leadership Grant, the NIH/NCI P50 CA247749 01 grant and by the NIH/NCI Cancer Center Core Grant P30-CA008748. JNK is funded by the German Federal Ministry of Health (DEEP LIVER, ZMVI1-2520DAT111) and the Max-Eder-Programme of the German Cancer Aid (grant #70113864), the German Federal Ministry of Education and Research (PEARL, 01KD2104C), and the German Academic Exchange Service (SECAI, 57616814).

Figure Legends

Figure 1: History, potential and challenges of computational pathology. (A) Key use cases of AI in pathology. Strongly supervised AI has mostly been used for diagnostic purposes, or to generate input data for downstream models of prognosis or treatment response. Weakly supervised AI can directly yield diagnosis, prognostic or predictive models. **(B)** Challenges of AI in histopathology. **(C)** Histopathology workflows in the AI era. **(D)** Simplified timeline of developments in histopathology.

References

1. Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V. & Madabhushi, A. Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. *Nat. Rev. Clin. Oncol.* **16**, 703–715 (2019).
2. Janowczyk, A. & Madabhushi, A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *J. Pathol. Inform.* **7**, 29 (2016).
3. Shmatko, A., Ghaffari Laleh, N., Gerstung, M. & Kather, J. N. Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. *Nat Cancer* **3**, 1026–1038 (2022).
4. Campanella, G. *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, 1301–1309 (2019).
5. Kleppe, A. *et al.* Designing deep learning studies in cancer diagnostics. *Nat. Rev. Cancer* **21**, 199–211 (2021).
6. Coudray, N. *et al.* Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
7. Kather, J. N. *et al.* Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* **25**, 1054–1056 (2019).
8. Hajdu, S. I. The First Use of the Microscope in Medicine. *Annals of Clinical & Laboratory Science* **32**, 309–310 (2002).
9. Yamashita, R., Long, J., Banda, S., Shen, J. & Rubin, D. L. Learning Domain-Agnostic Visual Representation for Computational Pathology Using Medically-Irrelevant Style Transfer Augmentation. *IEEE Trans. Med. Imaging* **40**, 3945–3954 (2021).
10. Kather, J. N., Ghaffari Laleh, N., Foersch, S. & Truhn, D. Medical domain knowledge in domain-agnostic generative AI. *NPJ Digit Med* **5**, 90 (2022).
11. Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F. K. & Mahmood, F. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering* **5**, 493–497 (2021).
12. Lu, M. Y. *et al.* Federated learning for computational pathology on gigapixel whole slide images. *Med. Image Anal.* **76**, 102298 (2022).
13. Saldanha, O. L. *et al.* Swarm learning for decentralized artificial intelligence in cancer histopathology. *Nat. Med.* (2022) doi:10.1038/s41591-022-01768-5.
14. Howard, F. M. *et al.* The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nat. Commun.* **12**, 4423 (2021).
15. Schömig-Markiefka, B. *et al.* Quality control stress test for deep learning-based diagnostic model in digital pathology. *Mod. Pathol.* **34**, 2098–2108 (2021).
16. Janowczyk, A., Zuo, R., Gilmore, H., Feldman, M. & Madabhushi, A. HistoQC: An Open-Source Quality Control Tool for Digital Pathology Slides. *JCO Clin Cancer Inform* **3**, 1–7 (2019).
17. Center for Devices & Radiological Health. Good Machine Learning Practice for Medical Device Development: Guiding Principles. *U.S. Food and Drug Administration* https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles.
18. Fu, Y. *et al.* Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature Cancer* 1–11 (2020).
19. Binder, A. *et al.* Morphological and molecular breast cancer profiling through explainable machine learning. *Nature Machine Intelligence* **3**, 355–366 (2021).
20. Echle, A. *et al.* Artificial intelligence for detection of microsatellite instability in colorectal cancer-a multicentric analysis of a pre-screening tool for clinical application. *ESMO Open* **7**, 100400 (2022).
21. Yamashita, R. *et al.* Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. *Lancet Oncol.* **22**, 132–141 (2021).
22. Cifci, D., Foersch, S. & Kather, J. N. Artificial intelligence to identify genetic alterations in conventional histopathology. *J. Pathol.* (2022) doi:10.1002/path.5898.

23. Sirinukunwattana, K. *et al.* Image-based consensus molecular subtype (imCMS) classification of colorectal cancer using deep learning. *Gut* **70**, 544–554 (2021).
24. Kleppe, A. *et al.* A clinical decision support system optimising adjuvant chemotherapy for colorectal cancers by integrating deep learning and pathological staging markers: a development and validation study. *Lancet Oncol.* **23**, 1221–1232 (2022).
25. Kather, J. N. *et al.* Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature Cancer* 1–11 (2020).
26. Schmauch, B. *et al.* A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nature Communications* vol. 11 Preprint at https://doi.org/10.1038/s41467-020-17678-4 (2020).
27. Campanella, G. *et al.* H&E-based Computational Biomarker Enables Universal EGFR Screening for Lung Adenocarcinoma. *arXiv [cs.CV]* (2022).
28. Saillard, C. *et al.* Blind validation of MSIntuit, an AI-based pre-screening tool for MSI detection from histology slides of colorectal cancer. *bioRxiv* (2022) doi:10.1101/2022.11.17.22282460.
29. Kleppe, A. Area under the curve may hide poor generalisation to external datasets. *ESMO Open* **7**, 100429 (2022).
30. Echle, A. *et al.* Clinical-Grade Detection of Microsatellite Instability in Colorectal Tumors by Deep Learning. *Gastroenterology* **159**, 1406–1416.e11 (2020).
31. Bizzego, A. *et al.* Evaluating reproducibility of AI algorithms in digital pathology with DAPPER. *PLoS Comput. Biol.* **15**, e1006269 (2019).
32. Lipkova, J. *et al.* Deep learning-enabled assessment of cardiac allograft rejection from endomyocardial biopsies. *Nat. Med.* **28**, 575–582 (2022).
33. Simon, R. M., Paik, S. & Hayes, D. F. Use of archived specimens in evaluation of prognostic and predictive biomarkers. *J. Natl. Cancer Inst.* **101**, 1446–1452 (2009).
34. Dupuy, A. & Simon, R. M. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J. Natl. Cancer Inst.* **99**, 147–157 (2007).
35. CDC Summaries of EGAPP™ Recommendation Statements. https://www.cdc.gov/genomics/gtesting/egapp/recommend/index.htm (2022).
36. Benjamens, S., Dhunnoo, P. & Meskó, B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digit Med* **3**, 118 (2020).
37. Lu, M. Y. *et al.* Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat Biomed Eng* **5**, 555–570 (2021).