OXFORD

JNCI: Journal of the National Cancer Institute, 2023, **115(6)**, 608–612
https://doi.org/10.1093/jnci/djad048
Advance Access Publication Date: 17 March 2023
Commentary

Overcoming the challenges to implementation of artificial intelligence in pathology

Jorge S. Reis-Filho, MD, PhD, FRCPath,[1] Jakob Nikolas Kather, MD, MSc[2–4,*]

[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 and Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK

*__Correspondence to:__ Jakob Nikolas Kather, MD, MSc, Else Kroener Fresenius Center for Digital Health, Technical University Dresden, Fetscherstrasse 74, 01309 Dresden, Germany (e-mail: jakob-nikolas.kather@alumni.dkfz.de).

Abstract

Pathologists worldwide are facing remarkable challenges with increasing workloads and 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 directly extracting prognostic and predictive biomarkers from tissue slides. The long-awaited adoption of AI in pathology, however, has not materialized, and the transformation of pathology is happening at a much slower pace than that observed in other fields (eg, radiology). Here, we provide a critical summary of the developments in digital and computational pathology in the last 10 years, outline key hurdles and ways to overcome them, and provide a perspective for AI-supported precision oncology in the future.

In the last 10 years, artificial intelligence (AI) has been demonstrated as a useful tool in histopathology image analysis (1,2). In particular, AI can directly extract much information from hematoxylin and eosin–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 equal or surpass 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 derived from the tissue slide itself: for example, predicting the presence of prostate cancer from slides by using 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 currently not routinely possible for pathologists (Figure 1, A).

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 1, B). Despite the enthusiasm with the utilization of digital pathology and AI, why has AI not yet become a reality? 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 2 fundamental paradigm shifts: the introduction of digital pathology for diagnosis and pathological assessment of cancers, as well as the transition from a human-based diagnosis or 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 at 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 yet a reality. And indeed, why should pathologists move from microscopes to computer screens if the current

Received: December 21, 2022. **Revised:** March 02, 2023. **Accepted:** March 11, 2023
© The Author(s) 2023. Published by Oxford University Press. All rights reserved. For permissions, please email: journals.permissions@oup.com

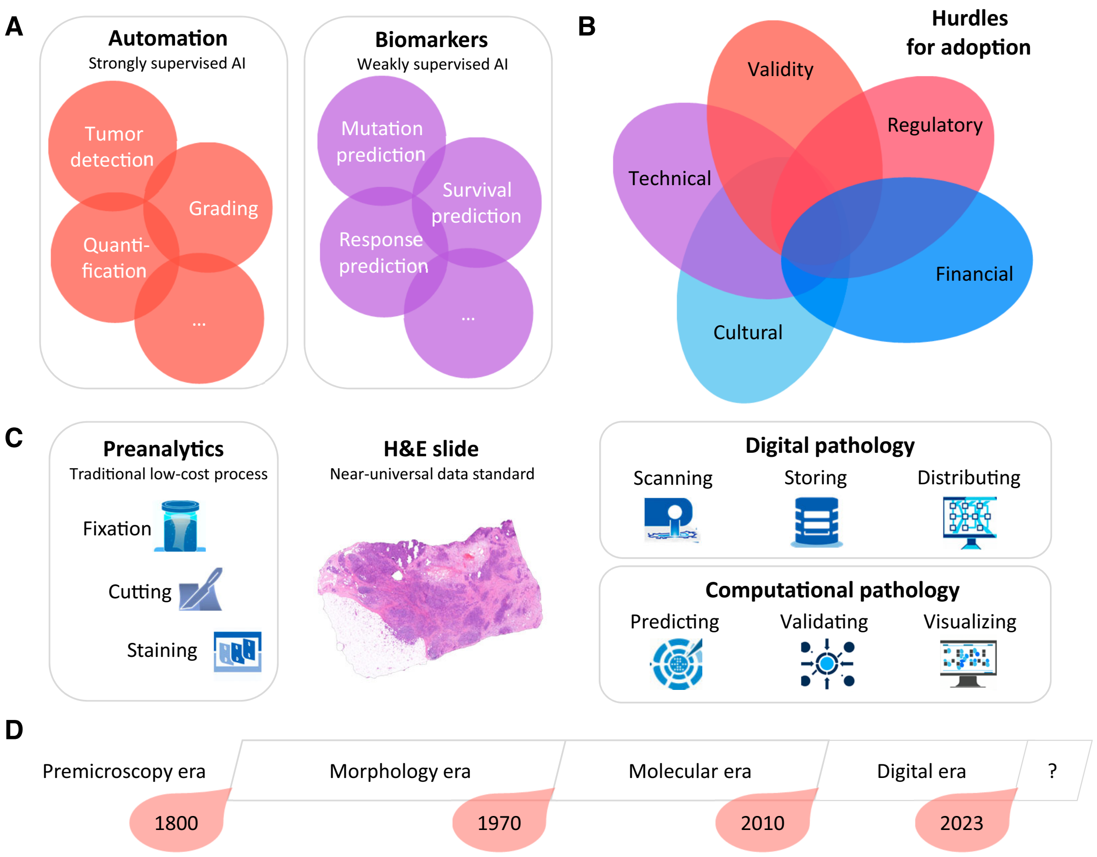

Figure 1. History, potential and challenges of computational pathology. **A**) Key use cases of artificial intelligence (AI) in pathology. Strongly supervised AI has mostly been used for diagnostic purposes or to generate input data for downstream models of prognosis or treatment response. Weakly supervised 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. H&E: Hematoxylin and eosin.

workflows are inexpensive and effective and the training of new pathologists is primarily based on microscope-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 evaluation and, ultimately, 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 United Kingdom, 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 1, C). Tissue fixation and 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 2 fundamental approaches to address this. The traditional approach holds up the “garbage in, garbage out” paradigm: according to this approach, preanalytical and data handling workflows should be perfectly standardized. 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, 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 or evaluated as an indicator of normal biological or 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 and 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 (7,20,21) and other genetic alterations (22) as well as molecular subtypes (23) and outcomes (24) from hematoxylin and eosin slides. Some studies have expanded these findings in a "pan-cancer" approach to any tumor type (18,25,26). Although these biomarkers do not reach perfect concordance with the ground truth methods and therefore cannot completely replace current sequencing methods, they can be used as prescreening 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 that 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 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 (ie, an application of the trained model to a dataset that 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 learned 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 rapid changes in regulatory approval. For AI-based diagnostic assays to be used clinically, they must pass the regulatory process for medical devices. This process differs between the United States, the European Union, and other large markets. In the European Union, the relevant rule set since May 2020 is the Medical Device Regulation and the In Vitro Diagnostic Medical Devices Regulation, and in the United States, the relevant rule set for any laboratory test is defined in the Clinical Laboratory Improvement Amendments 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 that 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 would there 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 regarding the required levels of analytical validity for the use of these algorithms as laboratory developed tests.

Financial challenges

Converting a pathology laboratory to digitized workflows is costly. Hardware cost and set-up incurs a high fixed cost, and 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 is still mostly unclear, how these technologies will be priced and whether they will be covered by insurance. Unlike many other laboratory equipment (eg, 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 and 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 with 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 health-care systems with less abundant resources. In fact, good performance of AI biomarkers can be achieved with only 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 10 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 that was essentially mature. The years 2021 and 2022 revealed that the technological aspects of AI are still expected to massively evolve (Figure 1, D). In particular, the zero-shot capabilities of large language models or diffusion models 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.

Data availability

No research data were used for this article.

Author contributions

Jorge S. Reis-Filho, MD PhD FRCPath (conceptualization; investigation; supervision; validation; writing—original draft; writing—review and editing); Jakob Nikolas Kather, MD, MSc (conceptualization; investigation; validation; visualization; writing—original draft; writing—review and editing).

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).

Conflicts of interest

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.

JSRF, a *JNCI* Deputy Editor and coauthor on this commentary, was not involved in the editorial review or decision to publish this manuscript.

References

1. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. *Nat Rev Clin Oncol*. 2019;16(11):703-715.
2. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. *J Pathol Inform*. 2016;7:29.
3. Shmatko A, Ghaffari Laleh N, Gerstung M, Kather JN. Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. *Nat Cancer*. 2022;3(9):1026-1038.
4. 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):1301-1309.
5. Kleppe A, Skrede O-J, De Raedt S, et al. Designing deep learning studies in cancer diagnostics. *Nat Rev Cancer*. 2021;21(3):199-211.
6. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat Med*. 2018;24(10):1559-1567.
7. Kather JN, Pearson AT, Halama N, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med*. 2019;25(7):1054-1056.
8. Hajdu SI. The first use of the microscope in medicine. *Ann Clin Lab Sci*. 2002;32(3):309-310.
9. Yamashita R, Long J, Banda S, Shen J, Rubin DL. Learning domain-agnostic visual representation for computational pathology using medically-irrelevant style transfer augmentation. *IEEE Trans Med Imaging*. 2021;40(12):3945-3954.
10. Kather JN, Ghaffari Laleh N, Foersch S, Truhn D. Medical domain knowledge in domain-agnostic generative AI. *NPJ Digit Med*. 2022;5(1):90.
11. Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F. Synthetic data in machine learning for medicine and healthcare. *Nat Biomed Eng*. 2021;5(6):493-497.
12. Lu MY, Chen RJ, Kong D, et al. Federated learning for computational pathology on gigapixel whole slide images. *Med Image Anal*. 2022;76:102298.
13. Saldanha OL, Quirke P, West NP, et al. Swarm learning for decentralized artificial intelligence in cancer histopathology. *Nat Med*. 2022;28(6):1232-1239.
14. Howard FM, Dolezal J, Kochanny S, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nat Commun*. 2021;12(1):4423.
15. Schömig-Markiefka B, Pryalukhin A, Hulla W, et al. Quality control stress test for deep learning-based diagnostic model in digital pathology. *Mod Pathol*. 2021;34(12):2098-2108.
16. Janowczyk A, Zuo R, Gilmore H, Feldman M, Madabhushi A. HistoQC: an open-source quality control tool for digital pathology slides. *J Clin Oncol Clin Cancer Inform*. 2019;3:1-7.
17. Center for Devices and Radiological Health. Good machine learning practice for medical device development: guiding principles. US Food and Drug Administration. https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles. Accessed January 2, 2023.
18. Fu Y, Jung AW, Torne RV, et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat Cancer*. 2020;1(8):800-810.

19. Binder A, Bockmayr M, Hägele M, et al. Morphological and molecular breast cancer profiling through explainable machine learning. *Nat Mach Intell*. 2021;3(4):355-366.
20. Echle A, Ghaffari Laleh N, Quirke P, 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*. 2022;7(2):100400.
21. Yamashita R, Long J, Longacre T, et al. Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. *Lancet Oncol*. 2021;22(1):132-141.
22. Cifci D, Foersch S, Kather JN. Artificial intelligence to identify genetic alterations in conventional histopathology. *J Pathol*. 2022;257(4):430-444.
23. Sirinukunwattana K, Domingo E, Richman SD, et al.; S:CORT Consortium. Image-based consensus molecular subtype (imCMS) classification of colorectal cancer using deep learning. *Gut*. 2021;70(3):544-554.
24. Kleppe A, Skrede O-J, De Raedt S, 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*. 2022;23(9):1221-1232.
25. Kather JN, Heij LR, Grabsch HI, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat Cancer*. 2020;1(8):789-799.
26. Schmauch B, Romagnoni A, Pronier E, et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nat Commun*. 2020;11(1):3877. doi:10.1038/s41467-020-17678-4
27. Campanella G, Ho D, Häggström I, et al. H&E-based computational biomarker enables universal EGFR screening for lung adenocarcinoma. *arXiv [cs.CV]*. 2022. https://arxiv.org/abs/2206.10573.
28. Saillard C, Dubois R, Tchita O, 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*. 2022;7(2):100429.
30. Echle A, Grabsch HI, Quirke P, et al. Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning. *Gastroenterology*. 2020;159(4):1406-1416.e11.
31. Bizzego A, Bussola N, Chierici M, et al. Evaluating reproducibility of AI algorithms in digital pathology with DAPPER. *PLoS Comput Biol*. 2019;15(3):e1006269.
32. Lipkova J, Chen TY, Lu MY, et al. Deep learning-enabled assessment of cardiac allograft rejection from endomyocardial biopsies. *Nat Med*. 2022;28(3):575-582.
33. Simon RM, Paik S, Hayes DF. Use of archived specimens in evaluation of prognostic and predictive biomarkers. *J Natl Cancer Inst*. 2009;101(21):1446-1452.
34. Dupuy A, Simon RM. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J Natl Cancer Inst*. 2007;99(2): 147-157.
35. CDC summaries of EGAPP^{TM} recommendation statements. 2022. https://www.cdc.gov/genomics/gtesting/egapp/recommend/index.htm. Accessed January 2, 2023.
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*. 2020;3:118.
37. Lu MY, Williamson DFK, Chen TY, et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat Biomed Eng*. 2021;5(6):555-570.