Artificial intelligence for automating the measurement of histologic image biomarkers

Toby C. Cornish

Commentary

Artificial intelligence has been applied to histopathology for decades, but the recent increase in interest is attributable to well-publicized successes in the application of deep-learning techniques, such as convolutional neural networks, for image analysis. Recently, generative adversarial networks (GANs) have provided a method for performing image-to-image translation tasks on histopathology images, including image segmentation. In this issue of the *JCI*, Koyuncu et al. applied GANs to whole-slide images of p16-positive oropharyngeal squamous cell carcinoma (OPSCC) to automate the calculation of a multinucleation index (MuNI) for prognostication in p16-positive OPSCC. Multivariable analysis showed that the MuNI was prognostic for disease-free survival, overall survival, and metastasis-free survival. These results are promising, as they present a prognostic method for p16-positive OPSCC and highlight methods for using deep learning to measure image biomarkers from histopathologic samples in an inherently explainable manner.

Find the latest version:

https://jci.me/147966/pdf

Artificial intelligence for automating the measurement of histologic image biomarkers

Toby C. Cornish

Department of Pathology, University of Colorado School of Medicine, Aurora, Colorado, USA.

Artificial intelligence has been applied to histopathology for decades, but the recent increase in interest is attributable to well-publicized successes in the application of deep-learning techniques, such as convolutional neural networks, for image analysis. Recently, generative adversarial networks (GANs) have provided a method for performing image-to-image translation tasks on histopathology images, including image segmentation. In this issue of the *JCI*, Koyuncu et al. applied GANs to whole-slide images of p16-positive oropharyngeal squamous cell carcinoma (OPSCC) to automate the calculation of a multinucleation index (MuNI) for prognostication in p16-positive OPSCC. Multivariable analysis showed that the MuNI was prognostic for disease-free survival, overall survival, and metastasis-free survival. These results are promising, as they present a prognostic method for p16-positive OPSCC and highlight methods for using deep learning to measure image biomarkers from histopathologic samples in an inherently explainable manner.

Multinucleation as a prognostic biomarker

HPV is an oncogenic virus associated with squamous dysplasia and invasive carcinoma of a variety of body sites, most notably the oropharyngeal and anogenital regions (1). HPV-related carcinomas at these sites are recognized as molecularly and clinically distinct from their non-HPV–related counterparts (2). Notably, although patients with HPV-associated oropharyngeal squamous cell carcinoma (OPSCC) have better outcomes than those with non-HPV–associated OPSCC, there remains considerable variation in outcomes within the HPV-associated cohort. Since treatment for OPSCC entails substantial morbidity, personalizing the intensity and choice of treatment by stratifying the HPV-related risk for each patient is highly desirable.

Multinucleation has previously been described as a prognostic biomarker in HPV-positive OPSCC. In 2012, Lewis et al. found that patients whose tumor showed multinucleation had worse outcomes than did those whose tumor lacked multinucleation (3). Lewis and authors defined multinucleation as the presence of at least one high-power field of view containing three or more tumor cells with more than one nucleus. In a multivariate analysis, multinucleation was associated with worse disease-specific survival, with a HR of 11.9 ($P = 0.02$) (3). However, while Lewis et al. demonstrated that multinucleation conveyed prognostic information (independent of tumor stage, histologic type, extracapsular extension, and smoking history), multinucleation is not a required reporting element for HPV-associated OPSCC (3). Like other histologic features that show promising results in the literature, failure to adopt these methods clinically may in part be due to difficulties with efficiency and reproducibility when performing the task. Given that a single focus of multinucleation is sufficient to qualify a tumor as positive, one can appreciate the time required to carefully inspect all tumor slides at high power and the likelihood of false negatives. Additionally, the amount of tumor submitted may bias this metric.

In this issue of the *JCI*, Koyuncu et al. build on the above work by transforming multinucleation from a human-measured biomarker to a biomarker based on quantitative image analysis (4). Deep learning was key to the image analysis pipeline that Koyuncu and colleagues put forth, but unlike many approaches to using deep learning to extract prognostic information from whole-slide images (WSIs), the results are inherently explainable and can be confirmed directly by a practicing pathologist (4).

## Artificial intelligence applied to histopathology

The resurgence of artificial intelligence (AI) is one of the most exciting recent developments in medicine. Image-based specialties, such as radiology and pathology, have generated particular interest due to the large gains in performance that convolutional neural networks (CNNs), a form of deep learning, bring to the field of computer vision (5). In histopathology, AI is used to perform aspects of computer-aided diagnosis (CAD) to quantify biomarkers and to extract novel predictive and prognostic information from histologic sections. In the latter task, AI is trained to discover latent information residing in histologic and cytologic patterns rather than replicating known, human-derived criteria. While deep learning is promising, it raises several issues when applied to medicine, including the so-called "black box problem" — once a model is trained, it is difficult to explain why, in human terms, the model makes its predictions. Current models of prognosis,

> ▶ **Related Article: https://doi.org/10.1172/JCI145488**
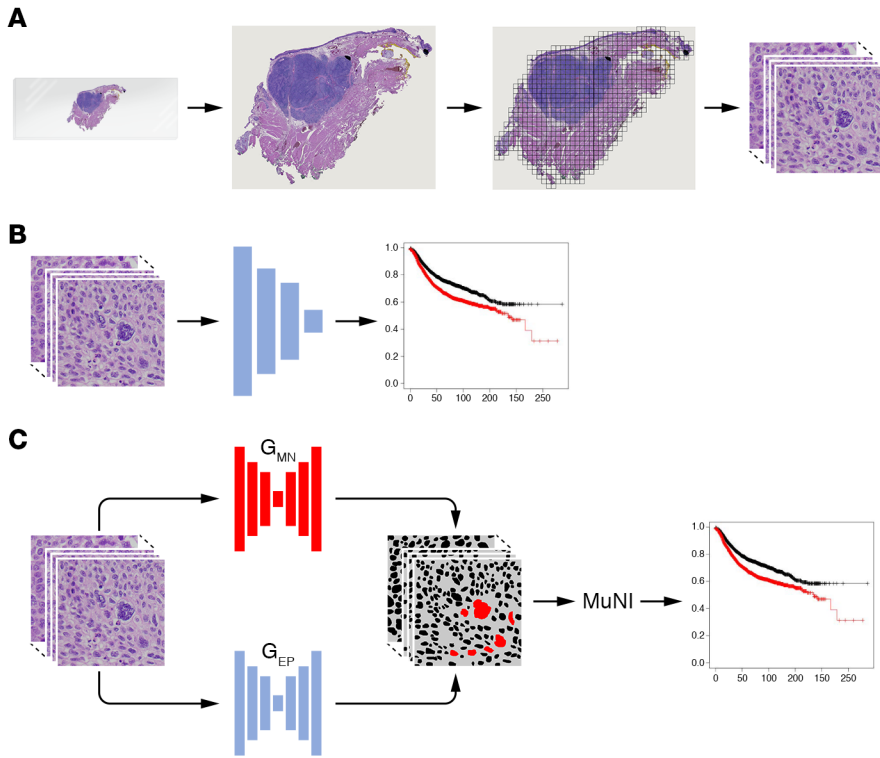
Figure 1. Direct and indirect application of deep learning to prognosis. Generally speaking, there are two different ways to use deep learning to derive prognostic information from histologic slides. (**A**) Both methods begin in a similar fashion, with digitization of stained tumor samples from a patient to create WSIs. The tissue in each WSI is then divided into smaller image patches. (**B**) The direct approach uses a CNN or similar deep-learning model that has been trained using tumor patches as input and patient outcomes as labels. This process permits the model to directly predict patient outcomes but is not easily explainable using current methods. (**C**) The indirect method is illustrated with a simplified representation of the GAN-based method used by Koyuncu et al., but other approaches might use fully convolutional networks or other types of CNNs to accomplish the same task. Two generators (G_{MN} and G_{EP}) translate the patches into segmentation masks, and these masks are combined to identify tumor nuclei (black) and multinucleated tumor nuclei (red). The MuNI is calculated for all tumor patches and serves as an intermediate value that can then be used, along with other clinical data, for prognostication.

for example, rely on human-observable histologic features like proliferation rates, tumor grade, and other features. Contrast this with a CNN-based model that has been trained on WSIs labeled solely with patient outcomes (6). A high-performing model of this type may predict patient outcomes with reasonable accuracy, but how it makes that prediction is not inherently explainable. Thus, deep-learning–based methods are left to be judged on their design and performance, and this can present substantial legal, ethical, and regulatory issues in medicine (7). In response to these concerns, many computer scientists are developing ways to tease out the important features obscured in the deep neural networks of models like CNNs. Although explainable AI is making headway in explaining deep-learning predictions, it remains to be seen if the "black box problem" will hinder the acceptance of deep-learning algorithms.

Generative adversarial networks for image analysis

While Koyuncu and authors also use a deep-learning approach, they produced a metric that is inherently explainable (Figure 1). Their method uses generative adversarial networks (GANs) to perform image-to-image translation, translating H&E-stained

images directly to image segmentation masks (4). First described by Goodfellow et al. in 2014, GANs did not appear in the biomedical literature until several years later (8). GANs are one type of generative model that uses deep learning. In simple terms, a generative model is a form of unsupervised machine learning that, given a set of input data, attempts to generate fake inputs that could plausibly be part of the distribution of input data. What makes GANs extremely powerful is the use of two deep-learning models (a generator and a discriminator) arranged in an adversarial relationship. The discriminator is trained to distinguish real input images from fake images, whereas the generator is used to generate synthetic images that resemble the images in the input domain. The discriminator and generator models are trained simultaneously and compete repeatedly in a type of zero-sum game. The generator creates fake images that are then provided to the discriminator along with images from the real data set. The discriminator then classifies the images as real or fake, and, based on how well it does, its parameters are updated to improve its performance of the task. Likewise, the parameters of the generator are also updated to improve its performance at fooling the discriminator. Thus, the two adversarial

networks engage in a game of one-upmanship that gradually ratchets up the performance of both models. The ultimate goal of a GAN is to train a generative model capable of synthesizing images that are indistinguishable from the real images, i.e., images that consistently fool the discriminator.

Classic GANs generate images entirely from random noise, which has some uses but may not be appropriate for biomedical imaging. Most applications in biomedical domains therefore use modified versions that provide additional input data when training the generator and discriminator. One of these approaches is conditional generative adversarial networks (cGANs) (9). cGANs still use random noise as an input, but both the generator and discriminator models are conditioned on additional input data. The most common cGAN task in histopathology is image-to-image translation, which uses an image from a different domain to condition the GAN. In this scenario, the generator, rather than creating a wholly fake image from random noise, is instead translating an image from the input (source) domain to the output (target) domain. Training a cGAN requires the availability of pairs of matched images — one for each domain. When paired images are unavailable, another type of GAN, the

cycle-consistent generative adversarial network (cycleGAN) can be used (10). In histopathology, these GAN variants have been used to perform a number of tasks, including image synthesis and data augmentation (11, 12), stain normalization (13), stain-to-stain translation (14), imaging modality adaptation (15, 16), image segmentation (17), ink removal (18), and WSI super-resolution (19).

MuNI: image biomarker or histologic feature?

In their study, Koyuncu et al. describe a metric that they call the multinucleation index or MuNI (4). At the heart of this metric are a pair of cGANs that perform image segmentation on tumor WSIs. GAN epithelial (GAN$_{EP}$) was trained to translate H&E-stained images directly into segmentation masks that divide the images into epithelial and nonepithelial areas. The GAN multinucleation (GAN$_{MN}$) was trained to translate H&E-stained images into masks that segment out cells with multinucleation. The MuNI was then calculated as the total number of multinucleated cells divided by the total number of epithelial cells. Multivariable analysis demonstrated that MuNI predicted disease-free survival (DFS), overall survival (OS), and distant metastasis–free survival (DMFS) independent of age, smoking status, treatment type, and tumor (T) and lymph node (N) stage (4).

While Koyuncu et al. draw heavily on prior work by Lewis et al., the application of AI turns what was previously a histologic feature into a quantitative image biomarker (3, 4). Image biomarkers are not frequently discussed in the context of histopathology, but are a topic of marked interest in radiology, given that specialty's dependence on imaging. Some may argue that metrics like the MuNI are merely manifestations of quantitative digital image analysis (DIA) applied to histologic image features, but these arguments serve only to diminish the important role that DIA and AI should play in the future of digital pathology. Although DIA is currently used in clinical applications, it remains a narrowly deployed tool with limited reim-

bursement in the United States (20). The MuNI, for example, would be ineligible for reimbursement under the current scheme, because existing medical billing codes for quantitative DIA only cover the analysis of immunohistochemical labels, not H&E-stained slides. Clinicians need to pivot the discussion of DIA in histopathology from that of measuring immunohistochemical labels to the broader topic of quantifying image biomarkers, to drive the adoption of clinically validated biomarkers as they emerge. To this end, pathology as a field would be wise to follow the lead of radiology, which recognizes the need to standardize image biomarkers and has begun efforts like the Image Biomarker Standardization Initiative (IBSI) (21). Although these are early days for digital imaging in pathology, interest in quantitative image biomarkers should increase as more clinical laboratories adopt digital pathology.

Address correspondence to: Toby C. Cornish, Department of Pathology, University of Colorado School of Medicine, Mail Stop B216, 12631 East 17th Avenue, Aurora, Colorado 80045, USA. Email: toby.cornish@cuanschutz.edu.

1. Szymonowicz KA, Chen J. Biological and clinical aspects of HPV-related cancers. *Cancer Biol Med*. 2020;17(4):864–878.
2. Syrjänen S, et al. HPV in head and neck cancer — 30 years of history. In: Golusiński W, et al., eds. *HPV Infection in Head and Neck Cancer*. Volume 206. Springer; 2017:3–25.
3. Lewis JS, et al. Tumor cell anaplasia and multinucleation are predictors of disease recurrence in oropharyngeal squamous cell carcinoma, including among just the human papillomavirus-related cancers. *Am J Surg Pathol*. 2012;36(7):1036–1046.
4. Koyuncu CF, et al. Computerized tumor multinucleation index (MuNI) is prognostic in p16+ oropharyngeal carcinoma: a multisite validation study. *J Clin Invest*. 2021;131(8):e145488.
5. Razzak MI, et al. Deep learning for medical image processing: overview, challenges and future. In: Dey N, et al., eds. *Classification in BioApps*. Springer; 2018:325–350.
6. Zhu W, et al. The application of deep learning in cancer prognosis prediction. *Cancers (Basel)*. 2020;12(3):603.
7. Holzinger A, et al. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev Data Min Knowl Discov*.

2019;9(4):e1312.
8. Goodfellow IJ, et al. Generative adversarial networks [preprint]. http://arxiv.org/abs/1406.2661. Posted on arXiv June 10, 2014.
9. Tschuchnig ME, et al. Generative adversarial networks in digital pathology: a survey on trends and future potential. *Patterns (N Y)*. 2020;1(6):100089.
10. Zhu JY, et al. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks [preprint]. http://arxiv.org/abs/1703.10593. Posted on arXiv August 24, 2020.
11. Van Eycke YR, et al. Strategies to reduce the expert supervision required for deep learning-based segmentation of histopathological images. *Front Med (Lausanne)*. 2019;6:222.
12. Levine AB, et al. Synthesis of diagnostic quality cancer pathology images by generative adversarial networks. *J Pathol*. 2020;252(2):178–188.
13. Zanjani FG, et al. Stain normalization of histopathology images using generative adversarial networks. Presented at: 15th IEEE International Symposium on Biomedical Imaging (ISBI 2018); April 4–10, 2018. Washington, DC, USA. https://doi.org/10.1109/ISBI.2018.8363641. Accessed March 19, 2021.
14. Vasiljević J, et al. Towards histopathological stain invariance by unsupervised domain augmentation using generative adversarial networks [preprint]. http://arxiv.org/abs/2012.12413. Posted on arXiv December 22, 2020.
15. Bayramoglu N, et al. Towards virtual H&E staining of hyperspectral lung histology images using conditional generative adversarial networks. Presented at: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW); October 23, 28, and 29, 2017. Venice, Italy. https://doi.org/10.1109/ICCVW.2017.15. Accessed March 19, 2021.
16. Xing F, et al. Bidirectional mapping-based domain adaptation for nucleus detection in cross-modality microscopy images [published online December 7, 2020]. *IEEE Trans Med Imaging*. https://doi.org/10.1109/tmi.2020.3042789.
17. Mahmood F, et al. Deep adversarial training for multi-organ nuclei segmentation in histopathology images. *IEEE Trans Med Imaging*. 2020;39(11):3257–3267.
18. Jiang J, et al. Image-to-image translation for automatic ink removal in whole slide images. *J Med Imaging (Bellingham)*. 2020;7(5):057502.
19. Li B, et al. Single image super-resolution for whole slide image using convolutional neural networks and self-supervised color normalization. *Med Image Anal*. 2021;68:101938.
20. Cornish TC. Clinical application of image analysis in pathology. *Adv Anat Pathol*. 2020;27(4):227–235.
21. Zwanenburg A, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*. 2020;295(2):328–338.