MINI-REVIEW

Searching Images for Consensus

Can AI Remove Observer Variability in Pathology?

Hamid R. Tizhoosh,* Phedias Diamandis,† Clinton J.V. Campbell,† Amir Safarpour,* Shivam Kalra,† Danial Maleki,* Abtin Riasatian,* and Morteza Babaie*

From the Kimia Laboratory,* University of Waterloo, Waterloo; the Department of Laboratory Medicine and Pathobiology,† University of Toronto, Toronto; and the Department of Pathology and Molecular Medicine,† Faculty of Health Sciences, McMaster University, Hamilton, Ontario, Canada

Accepted for publication

Address correspondence to Hamid R. Tizhoosh, Ph.D., University of Waterloo, Waterloo, Ontario, Canada. E-mail: tzhoosh@uwaterloo.ca.

Observer variability in pathology is described as the degree of variation between the diagnostic interpretations when a set of cases are examined by two or more independent clinicians. While there are many ways to estimate these errors, in most circumstances, this estimation requires large-scale case reviews by two or more blinded pathologists examining the same cases independently. While straightforward in theory, this quantification has proven difficult, as it is a costly and time-consuming process and appears to vary across different sample types and pathologists. Overall, cancer diagnoses tend to be more highly variable than non-neoplastic cases, especially as the number of diagnostic criteria continues to evolve in the era of molecular diagnostics.1,2 Variability also appears to depend on specific tumor type and tumor frequency, with breast and gynecologic malignancies and less common atypical tumor types showing higher observer variability than skin and gastrointestinal lesions and more common tumor types, respectively.3−5 Similarly, observer factors are an important consideration. Specific training and disease distribution may substantially change from a pathologist’s original training or working environment, leading to specific diagnostic biases.6 In some studies in which pathologists shared difficult cases with intradepartmental colleagues, the diagnosis was changed in as many as 13% of cases.5,7

The overall reported error rates in pathology are relatively low and are currently of little clinical significance at a population level.8,9 However, the transition to more personalized care and the reliance on more precise diagnostics for guiding patient management will likely make previously inconsequential errors more crucial, especially for individual patients.10 Therefore, subtle differences between diagnostic considerations currently too subjective for consistent agreement between independent pathologists may soon lead to life-saving decisions in personalized medicine. A false positive diagnosis could result in a patient receiving inappropriate or unnecessary toxic chemotherapy in oncology. Conversely, false negative results may delay or deny the benefit of effective therapy in another patient. Unfortunately, many errors appear to be the result of pathologists missing a pathologic finding on a slide or not

No Funding.

Disclosures: S.K. is a Ph.D. student at the University of Waterloo and had a Mathematics of Information Technology and Complex Systems internship at Huron Digital Pathology (St. Jacobs, ON, Canada). H.T. has received consultant’s fees from Huron Digital Pathology in 2020.

This article is part of a mini-review series on the applications of artificial intelligence and deep learning in advancing research and diagnosis in pathology.
considering an alternative diagnosis (false negative), and may be related to distracting factors that are increasing in modern clinical practice.\textsuperscript{3,4,11,12} Recently, several potential diagnostic misinterpretations in expert-annotated cohorts have been highlighted, suggesting that interobserver variability may be an under-reported issue that requires addressing modern informatic solutions.\textsuperscript{13} As new tools for addressing interobserver variability are designed, an understanding of why and where most variability exists can help in prioritizing the development of context-specific informatics tools to improve quality assurance in pathology.

Interobserver variability also remains problematic in a major field of pathology, hematopathology, specifically in myelodysplastic syndromes (MDSs). Diagnosing MDSs is one of the most challenging areas in hematopathology and involves multiple clinical and pathologic components.\textsuperscript{14} The central and most important aspect in making a diagnosis of MDSs is cytomorphology (visual features of bone marrow cells).\textsuperscript{14,15} In suspected cases of MDS, individual bone marrow cells are assessed by a hematopathologist for subtle and nuanced variations of normal cytomorphologic features known as myelodysplasias.\textsuperscript{16} While the World Health Organization has attempted to standardize the assessment of myelodysplasias in MDS,\textsuperscript{15} significant interobserver variation remains.\textsuperscript{16--20} Up to 12% of MDS cases might be misdiagnosed in smaller, less experienced centers due to a failure to recognize morphologic dysplasias.\textsuperscript{21} Consequently, there is a recognized need for standardization in the assessment of cytomorphologic features in MDS.\textsuperscript{16,22,23} Given the success of machine learning in the analysis of histomorphologic features in many nonhematopoietic pathology domains,\textsuperscript{24--26} there is a clear opportunity to develop machine learning—based approaches to support standardized cytomorphology in MDS.

Looking at the immensity of observer variability as a pervasive source of error and ineffectiveness, the question becomes pressing: Which artificial intelligence (AI) paradigm, and to what extent, can bring about a real solution?

### Supervised versus Unsupervised Solutions

In AI, different learning algorithms can be utilized, depending on the available data and the research problem at hand. The level of “supervision” used for solving the problem is certainly a key factor in categorizing AI methods. Supervised, semi-supervised, and unsupervised learning can generally be distinguished. When there is a labeled data set in which each sample is tagged/delineated either quantitatively or qualitatively, supervised learning is often a reasonable candidate solution. The goal of supervised learning is to use inputs for predicting the matching outputs. According to the output type, the prediction task can be called regression or classification for qualitative and quantitative outputs.

Supervised-learning applications in computational pathology can fall into several categories: classification,\textsuperscript{27} segmentation,\textsuperscript{28} detection,\textsuperscript{29} immunostaining scoring,\textsuperscript{30} cancer staging,\textsuperscript{31} and survival prediction.\textsuperscript{32,33} Although supervised-learning approaches sometimes achieve and even surpass human-level precision, the required labor-intensive and costly manual annotation and delineation procedures hinder the utilization of such methods in clinical practice. In addition, any annotated and delineated data set used for supervised learning is subject to observer variability, limiting its practical generalization beyond the experimental setting.

On the other hand, in unsupervised learning, the goal is to identify and infer salient patterns and underlying structure from the input, without a supervisor (ie, the experienced pathologist) to provide a correct answer. In other words, unsupervised learning attempts to transform the complexity of data into low-dimensional spaces that capture prominent information about the histopathologic findings. Given that there is no straightforward measure of success in unsupervised learning, it is not easy to validate the unsupervised predictions. Clustering and anomaly detection,\textsuperscript{34} data compression, and dimensionality reduction techniques are examples of the utilization of unsupervised learning in pathology.\textsuperscript{35}

Semi-supervised learning falls between supervised and unsupervised learning, as it combines a small amount of labeled data with a large amount of unlabeled data. Similarly, semi-supervised learning benefits from the advantages of both supervised and unsupervised learning. The labeled data are used for identifying specific groups present in the data. In contrast, the unlabeled data are used for boundary definition between the latter groups and to help identify unspecified data. Similar to supervised and unsupervised learning, semi-supervised learning is also used as a tool for representation learning in pathology.\textsuperscript{36} However, in case of any labeled data use, observer variability biases and limits learning scope, regardless of the degree of supervision.

Considering central challenges in computational pathology, such as the insufficient amount of labeled data, ground-truth errors due to observer variability, and weak data labeling due to expensive annotation and delineation processes and privacy and confidentiality concerns, semi-supervised and unsupervised learning methods hold more promise for the future. It is still not known exactly how unsupervised learning can be employed to overcome observer variability for building consensus, but it is more aligned with the philosophy of “strong AI” to be rather free from supervision. Both paradigms may be subject to biases of the AI expert/developer who selects a specific algorithm—supervised or not—to perform the desired task. However, user involvement would mostly put a cap on the performance and not introduce a new source of inconsistency, as computer algorithms are generally consistent.

### Image Search

Advances in AI research have shown great promise in assisting health care professionals. However, innovative
algorithms with reliable performance are necessary for gaining trust and adoption in clinical settings. Histopathology is the gold standard for the diagnosis of many diseases such as cancer, inflammations, and infections. With the widespread adoption of digital pathology, histopathology can greatly benefit from the applications of AI. When pathologists diagnose a difficult case, or when a new pathologist is in training, the identification and description of histologic features in images may become a common source of uncertainty. The conventional solutions are for pathologists to ask colleagues or to laboriously browse reference textbooks, hoping to find an image with similar visual characteristics. The general computer vision solution to similar problems is content-based image retrieval (CBIR), a research field with almost three decades of research.

CBIR systems use a search engine to retrieve similar images when the pathologist provides a query image, instead of using text for search in an archive. CBIR systems of medical images have been well researched. Only with the emergence of digital pathology and deep learning, research has begun to focus on image search and analysis for histopathological images. There are two major hurdles in the extensive usage of CBIR systems for digital pathology. First, pathology images exhibit an intractable level of variability in visual features which makes their identification, compared with natural images, much more challenging. The computational representation of histopathologic images for search purposes requires capturing the sematic high-level patterns of whole slide images (WSIs). Second, WSIs are gigapixel images of huge dimensionality (ie, larger than 50,000 × 50,000 pixels). Most research is focused on resolving these issues to make CBIR systems feasible for use in digital pathology.

Recent digital pathology studies have reported the success of supervised AI algorithms for use in classification and segmentation. Compared to other AI algorithms, this success is related to the ease of design and in-laboratory validation in generating highly accurate results. However, compared to other methods of computer-vision algorithms, CBIR offers a new approach to computational pathology. To facilitate image search, CBIR algorithms essentially describe the image content with nontextual attributes, generally with a vector of real numbers known as a feature vector. A set of AI algorithms can be trained to transform an image into a feature vector to serve as its representation. If a feature vector encompasses the descriptive visual properties of an image, then searching for similar images becomes a "nearest-neighbor" problem, that is, a matching task to find similar histologic features. Images with similar content can be retrieved based on a comparison of their feature vectors and not on the associated textual metadata or manual delineations subject to observer variability, and often come from a small number of pathologists. Image comparison is generally possible if a feature vector encodes the semantic structures of an image invariant to scale, rotation, translation, and deformation to some degree. Such rich and descriptive features, identified in an unsupervised manner, represent images for comparison and matching, which is the core task of any CBIR system.

Recently, the search engine Yottixel was proposed, enabling image search in large archives of histopathology images at both the patch/tile and WSI levels. The underlying technology behind Yottixel consists of a series of unsupervised AI algorithms, including clustering techniques, deep networks, and gradient barcoding. Yottixel indexes a WSI by converting tissue patterns into a set of barcodes, a process that is both storage friendly and computationally efficient. Every WSI is then represented as a bunch of barcodes (BoB), and thereby the image search can be performed by simpler and fast binary matching. After a large-scale validation of the search engine, the Yottixel results demonstrated that the image archives size and diversity play a role in image search. The largest public repository of WSIs, The Cancer Genome Atlas (provided by the National Cancer Institute/NIH) has been used for a
first validation. Almost 30,000 WSI files of 25 primary anatomic sites and 32 cancer subtypes were processed by dismantling of the large slides into almost 20 million image patches/tiles that were then individually indexed by employing approximately 3 million barcodes. The validation results show that image search can indeed provide a base for computational consensus. The primary diagnoses of evidently diagnosed cases can be used after they are identified by the search engine and listed as the best matches for the query WSI to decide based on the majority vote. Finding similar images opens the way for the development of new approaches to resolving observer variability through what might be called virtual peer review: Image search provides access to the reports of evidently diagnosed cases with similar histopathologic features.

The Google team recently published another search tool for use in digital pathology, similar image search for histopathology (SMILY). SMILY uses a deep-learning model, trained using 500 million natural, nonpathology images (dogs, trees, human-made objects, etc.), to compress images into a feature vector. During the training process, the network learned to distinguish similar images from dissimilar ones by computing and to compare their embeddings. This model is then used for generating a database of image patches and their associated features using a subset of WSIs from The Cancer Genome Atlas data set. SMILY uses a dense sampling, which may be advantageous in some situations, such as mitotic counting or finding some specific lesion within the same WSI, but is quite computationally expensive.

Any search framework depends on robust and expressive image representation, which are called deep features. Hence, extracting these deep features is the main step toward capable image search engines. Table 1 provides an overview of the most commonly used deep networks for this purpose.

One should bear in mind that the purpose of the image search as an unsupervised framework is not to get pathologists to agree on individual case diagnoses through interactions with each other (although this may very well be a possible and promising venue to investigate). Rather, the goal of any computer-driven solution is to maximize diagnostic accuracy. Whereas the high accuracy of supervised AI is subject to limited generalization due to the scarcity and variability-prone nature of labels, it is expected that unsupervised AI offers a more reliable path toward high accuracy by acting as a mediator for building computational consensus using evidently diagnosed cases from large-scale archives.

NLP and Pathology

Using a sophisticated CBIR system to search for similar histopathologic features in a large-scale archive may be an impressive task demonstrating the human-like ability of AI to identify images. However, it is of little value if the software shows the pathologist only similar images retrieved. The histologic matching results contribute to consensus and higher accuracy only if pathology reports and treatment outcomes accompany retrieved images of evidently diagnosed cases. Such information may not be stored where images are stored but elsewhere (eg, in laboratory information systems). These metadata are generally available in the form of text documents. Another set of AI algorithms, natural language processing (NLP), is required for processing the textual metadata.

Nowadays, NLP is no longer about translating or interpreting text or speech based on some key words, but about understanding its meaning. The first steps in this path were taken by using convolutional neural networks, recurrent neural networks, graph-based neural networks, and attention models. Using these deep neural models helps resolve the handcrafted-features problem by making this step automated, which is necessary for eliminating another aspect of observer variability in AI research. These features are learned in specific NLP tasks. Despite all of these improvements, the data used for training models were lacking, which forced networks to use shallower architectures (ie, smaller number of weights to adjust) to avoid the necessity of large data sets.
Recent work has shown that pre-trained models on a large amount of text data can be beneficial for many types of downstream NLP tasks. Deep models (i.e., transformers) alongside improvement in training skills have helped to generate a better pre-trained model. In this context, pre-trained means that an NLP method is trained with nonmedical text (abundantly available) and then used for medical cases, perhaps with a reduced need for a large-scale archive of clinical text, such as pathology reports.

The first generation of these pre-trained models, such as Skip-Gram and Glove, tried to capture the semantic meanings of words. Despite some advantages, these models failed to capture higher-level concepts in a text, such as disambiguation and syntactic structures. In the second-generation pre-trained models, such as embeddings from language models (ELMo), a generative pre-trained transformer (OpenAI GPT), bidirectional encoder representations from transformers (BERT), OpenAI GPT-2, language models (Megatron-LM) and OpenAI GPT-3, researchers tried to overcome this problem by using learning of contextual word embeddings. Although these models have mostly been used for general domain text, of late they have been increasingly used for the biomedical domain. For instance, ClinicalBert trained the BERT model with clinical records to predict the probability of patient readmission. Reports and clinical notes contain valuable information that can help supervised search to offer a baseline for consensus.

BioBert attempted to fine-tune the BERT model with varied medical corpora types such as PubMed abstracts, PubMed Central full-text articles, and other general corpora to be used in downstream NLP tasks such as entity recognition, relation extraction, and image search. Si et al compared different NLP methods for clinical concept extractors. Text and image are two significant resources of data. Therefore, many methods use these two types of data together. Image captioning and visual question answering are two examples. This combination can have far-reaching effects in digital pathology. MDNet is one of these AI models that can generate a pathology report for image patches that have been retrieved by symptoms descriptions and visual attention to provide justification for a diagnosis. Using NLP models in a specific domain can be challenging. The medical domain, especially pathology, is one of them. Pathologists can use different words to describe the same observation or use rare words, making it difficult to train an AI model to represent the images correctly. Keyword and topic selection are readily available for pathology reports (Figure 1). Moreover, a lack of a clean, large-scale, and universal data set for this domain is another challenge in using NLP methods for digital pathology.

The availability of NLP systems like BioBERT and MDNet is quite promising. Assuming that a pathologist is looking at the retrieved images through a capable CBIR system, the NLP can be applied to retrieved metadata of evidently diagnosed patients, that is, reports, to build an amalgamated caption or summarized description of the input image for the undiagnosed patient. This would be a computational consensus that can reduce observer variability if trusted by the pathologist.

Summary and Conclusion

Observer variability is not only difficult to quantify but also a major challenge in establishing diagnostic consensus. The progress and the astonishing success of AI in recent years provides a new paradigm for addressing this challenge. Unsupervised AI methods may be proven to be the solution if combined with the existing evidence in both images and reports in hospitals and laboratories. However, the unsupervised linkage of images and reports is impeded by a lack of access by the research community to large-scale digital archives for exploiting the potentials of both image search and natural language processing.

References

tumor—educated neural network. JCO Clin Cancer Inform 2020, 4: 811–821

14. Steensma DP: Myelodysplastic syndromes current treatment algo-

15. Arber DA, Hasserjian RP: Reclassifying myelodysplastic syndromes:
what’s where in the new WHO and why. Hematopathology 2015, 2015:
294–298

16. Della Porta MG, Tragullino E, Boveri E, Ponzoni M, Malcovati L,
Papaemmanuil E, M Rigolin G, Pascotto C, Croci G, Gianelli U,
Milani R, Ambaglio I, Elena C, Ubezio M, C Da Via’ M, Bono E,
Pietra D, Quaglia F, Bastia R, Ferretti V, Cuneo A, Morra E, J
Campbell P, Orazi A, Invernozzi R, Cazzola M: Minimal morpho-
critical criteria for defining bone marrow dysplasia: a basis for clinical
implementation of WHO classification of myelodysplastic syn-
dromes. Leukemia 2015, 29:66

17. Sasada K, Yamamoto N, Masuda H, Tanaka Y, Ishihara A,
Yakunagushi Y, Yatomi Y, Katsuda W, Sato I, Matsu H: Inter-
observer variance and the need for standardization in the
morphological classification of myelodysplastic syndrome. Leuk

18. Font P, Lescotetales J, Benavente C, Bermejo A, Callejas M, Garcia-
Alonso L, Garcia-Marcella A, Gil S, Lopez-Rubio M, Martin E,
Muñoz C: Inter-observer variance with the diagnosis of myelodys-
plastic syndromes (MDS) following the 2008 WHO classification.
Ann Hematol 2013, 92:19–24

19. Font P, Lescotetales J, Soto C, Ricard P, Muñoz-Novas C, Martín-
Clavero E, López-Rubio M, García-Alonso L, Callejas M, Bermejo A,
Benavente C: Interobserver variance in myelodysplastic
syndromes with less than 5 % bone marrow blasts: unilineage vs.
multilineage dysplasia and reproducibility of the threshold of 2%

20. Grasonen JE, Bennett JM, Bain BJ, Brunning R, Vallespi M-T,
Tomanaga M, Zini G, Renault A: Proposal for refining the definition
of dysgranulopoiesis in acute myeloid leukemia and myelodysplastic

ity search engine for large archives of
artificial intelligence and digital
pathology images using artificial intelligence. NPJ digital medicine 2020, 3:1–13

22. Kalra S, Tizhoosh HR, Shah C, Choi C, Damaskinos S, Safapoor A,
Diamandis P, Campbell CJ, Pantanowitz L: Pan-
cancer diagnostic consensus through searching archival histopathology
images with generative adversarial networks. IEEE journal of biomedical and
health informatics., 2019, 1316–1328.

23. Kalra S, Tizhoosh HR, Choi C, Shah S, Diamandis P, Campbell CJ,
Pantanowitz L: Yottixel—an image search engine for large archives of
histopathology whole slide images. Med Image Analysis 2020, 65:
101757

24. Hegde N, Hipp JD, Liu Y, Emmert-Buck M, Reif E, Smilkov D,
Terry M, Cai CJ, Amin MB, Mermel CH, Nelson PQ: Similar
image search for histopathology: SMILY. NPJ digital medicine
2019, 2:1–9

25. Shi X, Xin F, Xu K, Xie Y, Su H, Yang L: Supervised graph
hashing for histopathology image retrieval and classification. Med Image
Analysis 2017, 42:117–128

26. Komura D, Ishikawa S: Machine learning methods for histopath-

27. Tomczak K, Czerwińska P, Wiznerowicz M: The cancer genome
atlas (TCGA): an immeasurable source of knowledge. Contemp
Oncol 2015, 19:A68

WSIs using GNN for scalable diagnostically relevant regions
retrieval. International Conference on Medical Image Computing and
Computer-Assisted Intervention 2019 (MICCAI 2019). Cham,
Switzerland: Springer; 2019

thological whole slide image analysis using context-based CBIR.

30. Adnan M, Kalra S, Tizhoosh HR: Representation learning of histopath-
ology images using graph neural networks. Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition
988–997


