Editorial

The promise and challenges of deep learning models for automated histopathologic classification and mutation prediction in lung cancer

Pradnya D. Patil[1], Brian Hobbs[2], Nathan A. Pennell[1]

[1]Department of Hematology and Oncology, [2]Department of Quantitative Health Sciences, Taussig Cancer Institute, Cleveland Clinic, Cleveland, OH, USA

Correspondence to: Nathan A. Pennell, MD, PhD. Associate Professor, Department of Hematology and Oncology, Taussig Cancer Institute, Cleveland Clinic, 9500 Euclid Avenue, CA-6, Cleveland, OH 44195, USA. Email: penneln@ccf.org.

Provenance: This is an invited Editorial commissioned by the Section Editor Long Jiang (Department of Thoracic Oncology, Second Affiliated Hospital, Institute of Respiratory Diseases, Zhejiang University School of Medicine, Hangzhou, China).

Comment on: 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;24:1559-67.

Submitted Nov 26, 2018. Accepted for publication Dec 06, 2018.
doi: 10.21037/jtd.2018.12.55
View this article at: http://dx.doi.org/10.21037/jtd.2018.12.55

The histologic subclassification of tumors as either squamous cell carcinoma or adenocarcinoma is used by convention to guide systemic chemotherapy for patients diagnosed with non-small cell lung cancer (NSCLC) (1,2). The current gold standard for histopathologic classification of tissue specimens is visual microscopic inspection of tissue specimens by pathologists. In cases where morphological appearance is not adequate for classification, discriminatory immunohistochemical (IHC) stains are often required. Visual inspection of pathology slides is a labor-intensive process and diagnosis may be further delayed if IHC stains are required for definitive diagnosis. Over the last 20 years, further insights into the heterogeneous nature of NSCLC, particularly adenocarcinoma, have motivated new research areas endeavoring to elucidate additional actionable characteristics of tumor cells and the surrounding tumor microenvironment. In this era of precision medicine, identification of targetable somatic mutations is an essential step in determining optimal systemic therapy in patients with adenocarcinoma of pulmonary origin. Despite the many advancements in methodologies for testing genetic alterations, molecular testing is also often time consuming and can be limited by availability of adequate tissue samples (3,4). Associations between morphologic appearances on histopathology and certain genetic alterations have been previously described, however these findings lack discriminatory power to an extent required to impact traditional diagnostic protocols (5,6).

Advances in computer science, statistics, and data science have produced image classification algorithms devised to recognize patterns intrinsic to various types of objects. These algorithms are being widely used commercially outside the medical field in applications such as image and facial recognition on social media (7). This technology has many potential applications in the medical field as well, including the potential for automated classification of histopathologic specimens. Over the last decade rapid advancements in computing power, the availability of large datasets, and development of improved algorithms have accelerated the pace of innovations in image classification. The use of deep convoluted neural networks (DCNN), in particular, which are modeled after biologic networks involved in image processing (such as the visual cortex) have improved the predictive performance and robustness of image-derived classifiers. These essentially process multiple layers of non-linear information in the form of extracted image-based features to recognize patterns which can then be used to classify images into different categories (8,9). In a recently published study in Nature Medicine, Coudray *et al.* trained a deep convolutional neural image processing network to automatically classify histopathological subsets from digitalized lung specimens as well as to predict

common mutated genes in adenocarcinoma of the lung (10). The performance of their models was quite promising, with accuracy comparable to that of pathologists and findings that were reproducible irrespective of the methodologies used for tissue preservation and processing. Crucial for histopathologic image processing, the authors used a pre-trained image recognition model (Inception V3) which has the ability to process data at multiple resolutions. This model has demonstrated good performance in the presence of both limited hyperparameter specification and computational capacity, and thus valuable to broader dissemination of big data analytics (11).

In this study, 1,634 whole slide images (1,176 of tumor tissue and 459 of normal lung) obtained from The Cancer Genome Atlas (TCGA) database were used to train and validate image classification models. Training the classifiers using whole slide images, the authors eliminated potential confounders that could arise by training and validating the model on tiles obtained from the same specimen. Since the whole slide images were too large to be utilized as an input for the neural network, the images were split into non-overlapping tiles for analysis. The authors first developed a classification model that could discriminate between the tiles consisting of normal lung tissue, adenocarcinoma and squamous cell carcinoma. The performance of this model was assessed by area under the receiver operating characteristic curve (AUC), where an AUC of 1 reflects perfect class discrimination, while an AUC of 0.5 implies that the classifier is no more informative than random guessing. The AUC of their model was reported as 0.97, which is higher than what has been achieved in previously reported studies, some of which used conventional feature-based image processing and machine-learning methods (11,12). When compared to the classification of whole slide images in the training set by two thoracic pathologists and one anatomic pathologist, the deep-learning model had a slightly higher agreement with the TCGA classification, although this did not reach statistical significance (AUC 0.82 *vs.* 0.78 for the consensus of the three pathologists). Out of the slides that were misclassified by the model, over half were also inaccurately identified by at least one pathologist. In addition, 45 of the 54 TCGA images that had been misclassified by at least one pathologist was accurately classified by the model.

Since real world pathology slides often differ from the TCGA slides in that they have a lower content of tumor cells, artifacts from processing and other features such as blood clots, necrosis and inflammation that can affect the accuracy of automated image recognition models; the authors went on to validate their classifier using an independent set of formalin-fixed paraffin-embedded (FFPE), frozen section and biopsy specimens (n=140, 98 and 102 respectively). To account for some of the previously mentioned complexities of real-world biopsies, areas of high tumor content were manually annotated by pathologists and selected for testing. In these high tumor content areas, the model was able to accurately classify adenocarcinoma from squamous cell carcinoma with encouraging accuracy (AUC ranging from 0.833 to 0.9777), with a higher accuracy at 5× magnification than 20× likely due to the higher number of non-malignant features visible at the higher resolution. Even after replacing the manual annotation of high tumor content areas with an automated selection process, the performance of trained classifier was more or less equivalent. For the cases that were morphologically inconclusive with regards to histology per evaluation by pathologists (one third of the total cases), the model was able to accurately determine the histopathology with an AUC of 0.809 which was only marginally lower than the AUC of the slides that had an obvious morphology.

In another set of experiments, the authors used a similar deep learning approach to distinguish somatic mutations in specimens of lung adenocarcinoma. They first trained and validated the model on specimens from the TCGA with at least 10% mutated tumors and were able to identify six mutations—serine/threonine protein kinase 11 (*STK11*), *EGFR*, FAT atypical cadherin 1 (*FAT1*), SET binding protein 1 (*SETBP1*), *KRAS* and *TP53* with AUCs between 0.733 and 0.856. In 4 out of the 6 genetic mutations, the model-based classifier was found to be associated with allele frequency. In an independent validation set of matched adenocarcinoma specimens with and without *EGFR* mutations (n=29 and 34 respectively), the AUC of the model was 0.75 in those that had EGFR testing performed by sequencing and 0.659 in those with testing performed by IHC. Of note, the classifier had been trained on images from TCGA which were tested using sequencing and therefore included other less common *EGFR* mutations which cannot be detected by IHC.

The findings reported in this study are intriguing and pave the path for further development of such deep learning models for processing digitalized histopathology images. The authors should be commended for not only testing the model in a set of slides with optimal tumor content and lesser artifacts (TCGA), but also pursuing rigorous

validation of the robustness of their model in independent real-world specimens that were procured and processed in a variety of different manners (fresh frozen and FFPE slides). The fact that the discriminative performance of the classifier for histology was equivalent to that of three pathologists suggests that image processing technologies may have evolved to an extent that warrants more widespread interrogations with clinical applications. In addition, the model performed almost as well in slides where morphology alone was considered to be inadequate for classification by pathologists, which further attests to the robustness of the model.

While phenotypic associations of somatic mutations are not routinely used for predicting oncogenic drivers in NSCLC, the authors of this study reported promising predicative capabilities for 6 frequently noted genetic alterations in this population. Although the predictive capability of their genetic predictive model when tested in an independent cohort of patients with known *EGFR* mutation status was less accurate than that of their histology classification models, these findings warrant further testing in larger independent cohorts. Of note, the model was unable to detect *ALK* rearrangements despite the specific morphologic patterns described with this genetic alteration. It is therefore likely that the predictive performance of the DCNN for genetic mutations depends on the degree of phenotypic changes associated with them.

Multiple studies have highlighted the importance of shortening the time to diagnosis and treatment in patients with advanced NSCLC (13). Using traditional diagnostics, it usually takes a few days for the histopathologic diagnosis and at least a week at most institutes to obtain additional molecular testing. The authors state that by using multiple graphics processing units (GPUs), the histologic classification using their model could be executed in seconds. Scanning of the slides to create digitalized images is currently a rate limiting step in this process, however with newer technology that may no longer be the case (14). It is plausible that in the future such deep learning image classification models could be integrated with traditional visual inspection to hasten the time to diagnosis, reduce inter observer variation and aide diagnosis in morphologically inconclusive cases. It is however important to note that while pathologists are trained to account for artifacts and variability that may arise as a consequence of different tissue processing methods, automated models that are trained on standardized slides which lack the ability to accurately classify such non-malignant features. Future research efforts directed at automated classification of such features may further improve upon the performance of these models in real world specimens. Nonetheless due to the inherent extent of variability in these features it is possible that the discriminative abilities of automated models may serve to compliment, but not replace conventional pathological evaluations. The integration of automated models trained on histology slides into routine clinical practice is also likely to be limited by the fact that many patients with NSCLC only have cytology specimens available at diagnosis.

The genetic mutation predicting capabilities of the models reported in this study are hypothesis generating and warrant further investigation and validation in larger cohorts. However, the performance of these models would have to be comparable to that obtained by sequencing to merit integration into routine clinical diagnostics. In addition, the ability of any such platform to detect all targetable mutations is key, since therapeutic decisions are usually delayed until information about all such driver mutations is available. There is however value to automated mutation prediction in not only hastening the time to treat by allowing early detection of oncogenic drivers, but also potentially forgoing the need for additional biopsies in cases with limited tissue for diagnosis.

In conclusion, the findings reported by Coudray *et al.* forecast an era of integrative diagnostics where deep learning approaches could be integrated with traditional diagnostic methods to aide pathologists. Further refinements o discern non-malignant artifacts would be an important step for generalizability of these models. It is also of the utmost importance that future research endeavors in this field exhaustively interrogate the reproducibility and robustness of their models in the manner reported in this study. Other foreseeable hurdles to the routine application of deep learning techniques include the additional costs and infrastructure requirements for their implementation, although after the initial investment it is likely that the cost of evaluating individual cases would be favorable compared to traditional testing. The economic viability of such an integrated workflow model would have to be assessed before automated histopathologic classification becomes a part of the standard diagnostic process.

Acknowledgements

None.

Footnote

Conflicts of Interest: N Pennell has consulted with Eli Lilly, AstraZeneca, Regeneron, and Merck. The other authors have no conflicts of interest to declare.

References

1. Sandler A, Gray R, Perry MC, et al. Paclitaxel-carboplatin alone or with bevacizumab for non-small-cell lung cancer. N Engl J Med 2006;355:2542-50. Erratum in: N Engl J Med 2007;356:318.
2. Scagliotti GV, Parikh P, von Pawel J, et al. Phase III study comparing cisplatin plus gemcitabine with cisplatin plus pemetrexed in chemotherapy-naive patients with advanced-stage non-small-cell lung cancer. J Clin Oncol 2008;26:3543-51.
3. Lim C, Tsao MS, Le LW, et al. Biomarker testing and time to treatment decision in patients with advanced nonsmall-cell lung cancer. Ann Oncol 2015;26:1415-21.
4. Schneider F, Smith MA, Lane MC, et al. Adequacy of core needle biopsy specimens and fine-needle aspirates for molecular testing of lung adenocarcinomas. Am J Clin Pathol 2015;143:193-200; quiz 306.
5. Lee B, Lee T, Lee SH, et al. Clinicopathologic characteristics of EGFR, KRAS, and ALK alterations in 6,595 lung cancers. Oncotarget 2016;7:23874-84.
6. Warth A, Penzel R, Lindenmaier H, et al. EGFR, KRAS, BRAF and ALK gene alterations in lung adenocarcinomas: patient outcome, interplay with morphology and immunophenotype. Eur Respir J 2014;43:872-83.
7. McAuley J, Leskovec J. Image Labeling on a Network: Using Social-Network Metadata for Image Classification. In: Fitzgibbon A, Lazebnik S, Perona P, et al. editors. Computer Vision – ECCV 2012. ECCV 2012. Lecture Notes in Computer Science, vol 7575. Springer, Berlin, Heidelberg.
8. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 7-12 June 2015; Boston, MA, USA.
9. Rawat W, Wang Z. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. Neural Comput 2017;29:2352-449.
10. 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;24:1559-67.
11. Khosravi P, Kazemi E, Imielinski M, Elemento O, Hajirasouliha I. Deep Convolutional Neural Networks Enable Discrimination of Heterogeneous Digital Pathology Images. EBioMedicine 2018;27:317-28.
12. Yu KH, Zhang C, Berry GJ, et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat Commun 2016;7:12474.
13. Olsson JK, Schultz EM, Gould MK. Timeliness of care in patients with lung cancer: a systematic review. Thorax 2009;64:749-56.
14. Abels E, Pantanowitz L. Current State of the Regulatory Trajectory for Whole Slide Imaging Devices in the USA. J Pathol Inform 2017;8:23.

Cite this article as: Patil PD, Hobbs B, Pennell NA. The promise and challenges of deep learning models for automated histopathologic classification and mutation prediction in lung cancer. J Thorac Dis 2019;11(2):369-372. doi: 10.21037/jtd.2018.12.55