



OPEN

Optimized digital workflow for pathologist-grade evaluation in bleomycin-induced pulmonary fibrosis mouse model

Toshiki Goto^{1,4}, Akira Sano^{2,4}✉, Shinichi Onishi³, Natsuko Hada¹, Rui Kimata², Saori Matsuo³, Sohei Oyama¹, Atsuhiko Kato³, Hideaki Mizuno¹ & Masaki Yamazaki³✉

Idiopathic pulmonary fibrosis (IPF) is a progressive and ultimately fatal disorder of unknown etiology, characterized by interstitial fibrosis of the lungs. Bleomycin-induced pulmonary fibrosis mouse model (BLM model) is a widely used animal model to evaluate therapeutic targets for IPF. Histopathological analysis of lung fibrosis is an important method for evaluating BLM model. However, this method requires expertise in recognizing complex visual patterns and is time-consuming, making the workflow difficult and inefficient. Therefore, we developed a new workflow for BLM model that reduces inter- and intra-observer variations and improves the evaluation process. We generated deep learning models for grading lung fibrosis that were able to achieve accuracy comparable to that of pathologists. These models incorporate complex image patterns and qualitative factors, such as collagen texture and distribution, potentially identifying drug candidates overlooked in evaluations based solely on simple area extraction. This deep learning-based fibrosis grade assessment has the potential to streamline drug development for pulmonary fibrosis by offering higher granularity and reproducibility in evaluating BLM model.

Idiopathic pulmonary fibrosis (IPF) is a progressive and ultimately fatal disorder of unknown etiology, characterized by interstitial fibrosis of the lungs^{1,2}. Repetitive alveolar epithelial injury triggers the early development of fibrosis and leads to tissue remodeling and the fibrosis seen in end-stage pulmonary fibrosis. The Ashcroft fibrosis score is a metric used to assess fibrotic changes in clinical samples under microscopic evaluation³. This assessment is traditionally performed by pathologists with a 10x objective in multiple fields of view, with scores ranging from 0 (normal lung) to 8 (total fibrous obliteration). The final score averages those of individual microscope fields.

To develop therapeutics for IPF, a widely used non-clinical model for IPF is bleomycin-induced pulmonary fibrosis mouse model (BLM model) of lung injury. Unfortunately, while no animal model fully recapitulates the histologic pattern of IPF or exhibits features of progressive disease⁴, BLM model, unlike human IPF, has characteristics such as rapidity of its development, inflammation preceding fibrosis, and self-resolution nature⁵. However, the model is currently considered “the best-characterized animal model available for preclinical testing” at this moment. Ashcroft fibrosis score, originally designed for human evaluation, includes late-phase IPF features such as “honeycomb lung”. Honeycomb lung appears rare in mouse models, but has been forcibly induced for the evaluation of BLM model^{6,7}. Furthermore, one intractable problem with Ashcroft fibrosis score is that it is time-consuming and requires highly trained pathologists who are nevertheless prone to intra- and inter-observer variations^{6,8}. Therefore, a new high-throughput and reproducible scoring system for assessing lung fibrosis in BLM model would be greatly aid the development of therapeutics^{9–11}.

Recently, several reports have been published on image analyses for quantifying fibrosis areas in lung^{8,12,13}. However, IPF grading requires highly nuanced interpretation of pathological images, including the accumulation of myofibroblasts in clusters called fibroblastic foci, and extensive extracellular matrix (ECM) deposition within the interstitium that results in the destruction of alveolar architecture¹⁴. Moreover, several reports have shown that tissue density affects fibrosis development^{12,15}. As such, simple measurements of the fibrosis area may not

¹Research Division, Chugai Pharmaceutical Co., Ltd., 216 Totsuka-cho, Totsuka-ku, Yokohama-shi, Kanagawa 244-8602, Japan. ²ExaWizards Inc., 4-2-8 Shibaura, Minato-ku, Tokyo 108-0023, Japan. ³Translational Research Division, Chugai Pharmaceutical Co., Ltd, 216 Totsuka-cho, Totsuka-ku, Yokohama-shi, Kanagawa 244-8602, Japan. ⁴Toshiki Goto and Akira Sano contributed equally to this work. ✉email: akira.sano@exwzd.com; yamazakimsk@chugai-pharm.co.jp

be enough to capture the severity of IPF. Therefore, we developed a workflow in BLM model and generated deep learning models for grading lung fibrosis. The models function with minimal variability and, furthermore, provide a continuous score, offering opportunities for research with higher granularity.

Materials and methods

Dataset

We collected 173 whole slide images (WSI) (ndpi, x20) for C57BL6/J mouse lung fibrosis slides stained with Sirius red. Although the slides of left lung were prepared in two different facilities, the Sirius red staining was performed following a standard procedure. These slides were obtained from five independent archived animal studies with mice treated with bleomycin including control mice (Table 1). In all studies, animal procedures were conducted in accordance with the Institute's Guide for the Care and Use of Laboratory Animals, and all experimental protocols were approved by the Institutional Animal Care and Use Committee at Chugai Pharmaceutical Co., Ltd. All C57BL/6J male mice were purchased from The Jackson Laboratory Japan, Inc. (Kanagawa, Japan) and were acclimated for 1 week before the start of treatments. Lung fibrosis was induced by the following three methods. Study 1 and 2 were established via intrathecral administration using 6-week-old mice. Bleomycin (Nihonkayaku, Tokyo, Japan) was administered intratracheally once at 0.75 mg/kg. Lung fibrosis was evaluated at day 21 from the administration of the bleomycin. Study 3 and 4 were established via osmotic pump using 8-week-old mice. 60 mg/kg Bleomycin was administered through subcutaneously by implanting osmotic infusion pump (ALTEZ 1007D) for 10 days. Lung fibrosis was evaluated at day 28 after pump implantation. Study 5 were established via subcutaneous administration using 8-week-old mice. Bleomycin was administered 200 µg/100 µL subcutaneously in total 10 shots (5 days consecutively administration, 2 days of withdrawal, and following 5 days consecutively administration). Lung fibrosis was evaluated at day 28 from 1st administration of the bleomycin. Lung sampling was performed under isoflurane anesthesia after exsanguination euthanasia in all studies. The study is reported in accordance with ARRIVE guidelines.

Histopathological evaluation

Pathologists assessed fibrosis grade for each animal study using light microscopy or WSI to establish ground truth. Fibrosis grades were categorized into four levels: none; 0% (Grade0), weak; less than 10% (Grade1), moderate; between 10% and 20% (Grade2), severe; more than 20% (Grade3) (Fig. 1a; Table 1). The grade was defined based on the area of fibrosis spreading within the lung tissue in each WSI. These pathologists provided the label for generating our deep learning models; however, to fully evaluate the performance of our models in comparison to pathologists, two additional pathologists (P#3 and P#4) also graded fibrosis using WSIs independently (Study 1 and 2 in Table 1).

Workflow

First, WSIs were preprocessed with HALO AI image analysis software (Indica Labs, NM, USA, v3.3 or v3.4) to distinguish the alveolar area for analysis. Using the “classifier” function, a pathologist annotated “alveolar” and “bronchus”. The software then iteratively trained and output the classified results, and the “alveolar” area was defined as the object of analysis. Next, the fibrosis area within “alveolar” area was extracted using the “Area quantification v2.1.11” module. These results, including images of the detected fibrosis area, annotation data, and the original WSI file, were exported.

The resolution of unmodified WSIs was too high for use as inputs for our models; however, reducing image size could result in the loss of information needed for fibrosis analysis. Therefore, we used a pre-processing strategy to preserve necessary information while focusing on areas critical for determining the fibrosis grade. The WSIs were cropped into rectangular images of the alveolar areas and were partitioned into 1024 × 1024-pixel patches. The background regions within these patches were then removed. Each patch was scored according to the area of fibrosis region detected by HALO AI. Next, we generated 5 × 5 tiled images using the top 25 scored patches from each WSI as input images to test and evaluate our models. For training the models, we also generated tiled images using 25 patches randomly selected from the top 50 scored patches, and each patch was randomly rotated and flipped. (Fig. 1b). We generated 50 tiled images corresponding to a single WSI, and during each training epoch, a random tiled image was selected.

Study No.	1	2	3	4	5
Evaluation	G, P#3, P#4	G, P#3, P#4	G	G	G
Center	A	A	A	A	B
Grade0	4	4	8	10	3
Grade1	18	28	16	16	30
Grade2	8	9	6	7	0
Grade3	5	1	0	0	0
Total number	35	42	30	33	33

Table 1. Summary of the dataset. G, P#3, and P#4 indicate ground truth, pathologist #3, and pathologist #4. The ground truth was established by two pathologists. Pathologist #1 conducted Study 1 and 2, while Pathologist #2 conducted Study 3, 4 and 5. Image dataset was prepared at two sites (A and B).

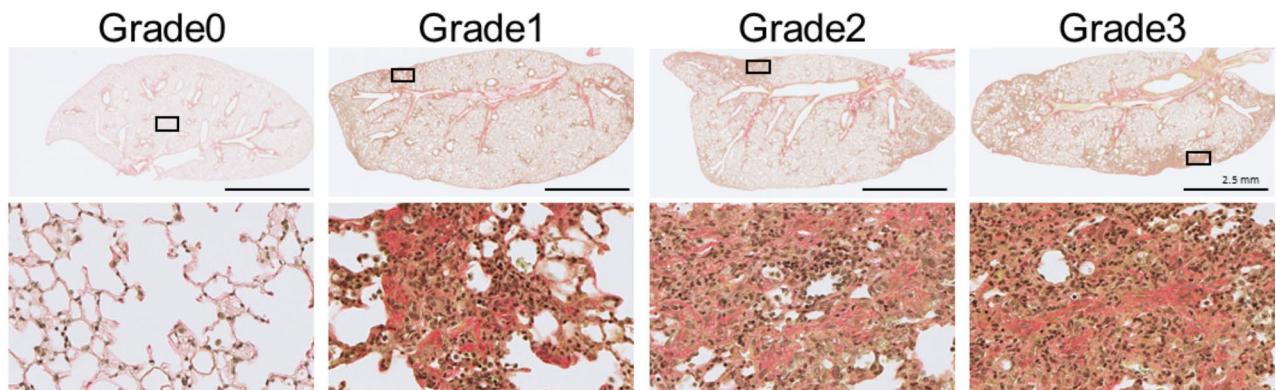
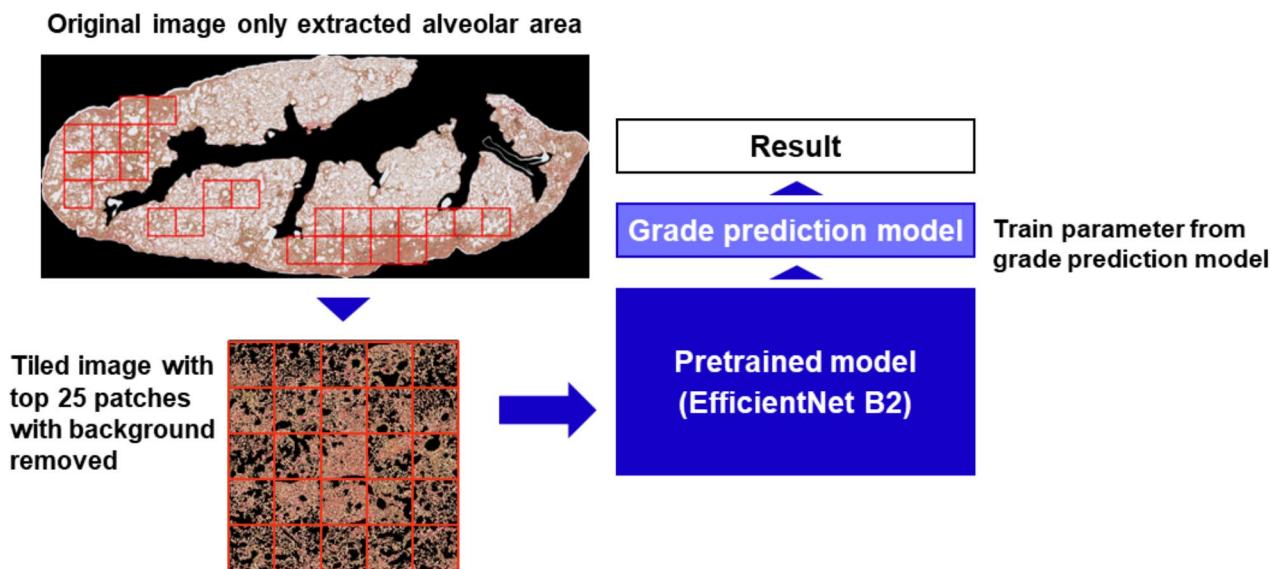
a**b**

Fig. 1. Prediction of the grade of lung fibrosis. (a) Lung fibrosis was defined as having four grades: none, Grade0; weak, Grade1; moderate, Grade2; severe, Grade3. The grade was defined based on the area of fibrosis spreading. The squares in the top figures indicate the areas of high magnification in the lower section. Bar, 2.5 mm. (b) Overview of workflow for predicting the lung fibrosis grade. Whole slide images (WSIs) were cropped into rectangular images encapsulating the alveolar area and were divided into 1024×1024 -pixel patches. Each patch was scored according to the area of fibrosis region detected by HALO AI. Next, we generated 5×5 tiled images using top 25 scored patches from each WSI. Tiled images were used as inputs for our models to test and evaluate.

To evaluate the effectiveness of patch pre-selection, we trained and evaluated another model without using patch pre-selection. As with the patch pre-selection approach, the WSIs were divided into 1024×1024 -pixel patches. Instead of selecting top-scored patches, all patches were used to generate 5×5 tiled images. Consequently, multiple tiled images were generated from a single WSI. For evaluation, the outputs from multiple tiled images of a single WSI were averaged to determine the predicted class.

Our models are convolutional neural networks (CNNs) with an EfficientNet-B2 backbone, pre-trained on ImageNet-1 K, and two fully connected layers as the classifier. We trained the model using a stochastic gradient descent (SGD) optimizer with 300 epochs, learning rate of 10^{-3} , and batch size of 64. We evaluated the models using 5-fold cross-validation. Each fold comprised 143 training data samples and 30 validation data samples. Given the limited number of data samples in this study, particularly with only six samples for Grade3, we did not simply divide the dataset into five parts and assign them to training and validation. Instead, we chose the validation data to ensure sufficient samples for each class per fold. The model training and validation were performed using Python v3.8.10 with Pytorch v1.10.2, Torchvision v0.11.3, Numpy v1.22.2 and Albumentations v1.3.0.

Statistical analysis and evaluation metrics

Statistical analysis and evaluation metrics were performed using Python v3.9.16 with Numpy v1.23.5 and Scipy v1.12.0. The effectiveness of each model compared to the pathologist evaluations was verified based on accuracy. The level of agreement between pathologist-pathologist pairs and machine-pathologist pairs was determined using both agreement coefficient and Cohen's kappa coefficient¹⁶ (Kcoef). The means of the two groups was compared using Welch's t-test.

Result

Validation of the models for grading fibrosis

To validate the quality of our models, we compared the fibrosis grades assigned by pathologists with the scores predicted by our deep learning models using 5-fold cross-validation (Fig. 2). The accuracy of all analyses was high, with the confusion matrix showing no estimation errors more than two classes apart. The models' scores were calculated by multiplying SoftMax values with the grade numbers. The figure shows a correlation between the pathologists' grades and the scores predicted by our models. The average accuracy of the five analyses for the model with patch pre-selection was 83%. In contrast, the average accuracy without patch pre-selection was 75%, demonstrating higher overall accuracy for the former. The model without pre-selection was more likely

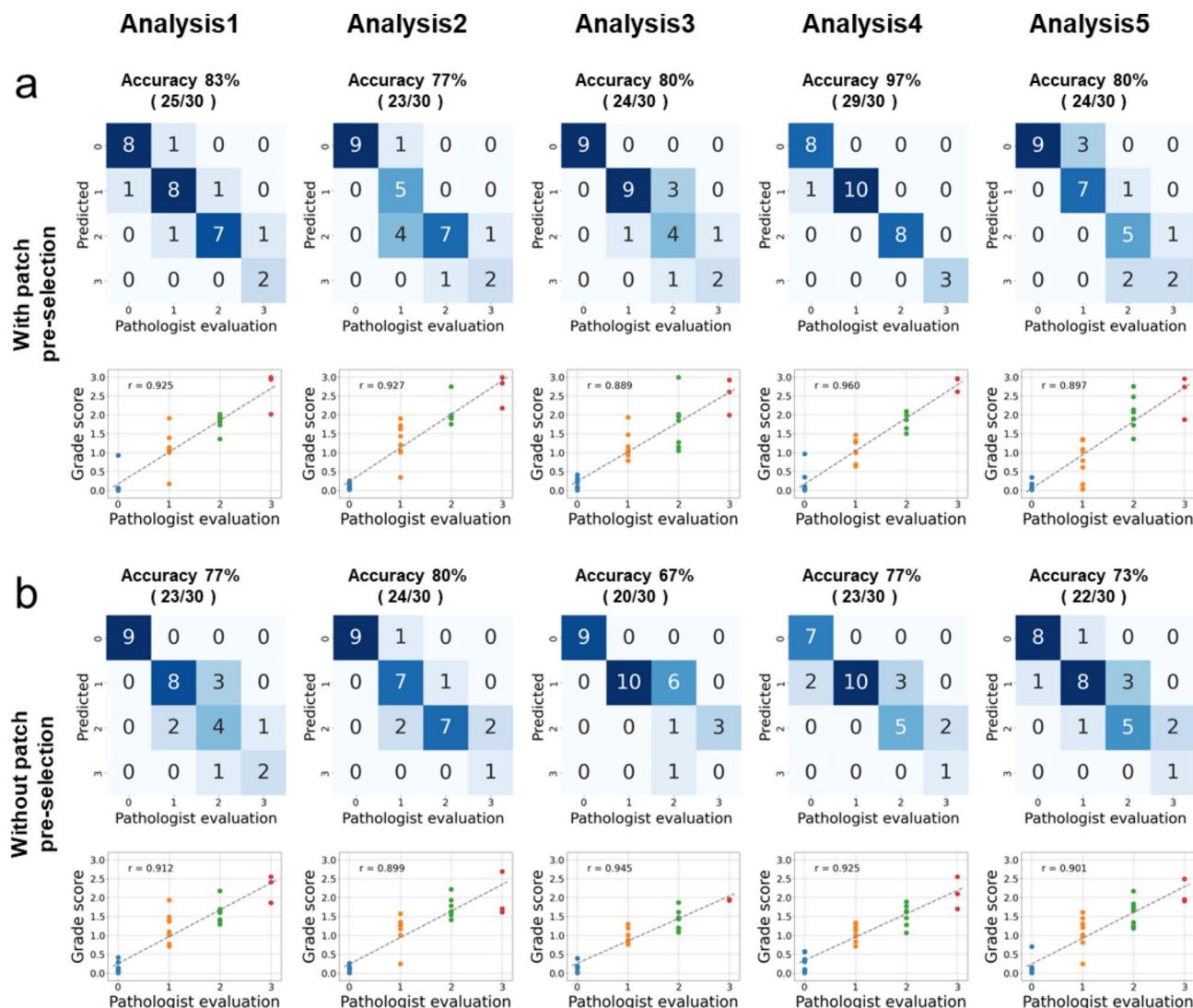


Fig. 2. Validation of the model for grading fibrosis. (a) (top) We calculated the accuracy by comparing the scores of the actual pathologists and those of predicted by our models using 5-fold cross-validation. (a) (bottom) The comparison of the pathologist evaluation and the grade score estimated by the models. The models' grade score was calculated by multiplying SoftMax values and grade numbers. The dashed line was fitted by the least-squares method and r value represents the correlation coefficient between the pathologist evaluation and the calculated grade score. (b) (top) To compare with (a), we calculated the accuracy of the models without patch pre-selection. (b) (bottom) Similarly, we compared the pathologist evaluation and the grade score estimated by the models without patch pre-selection.

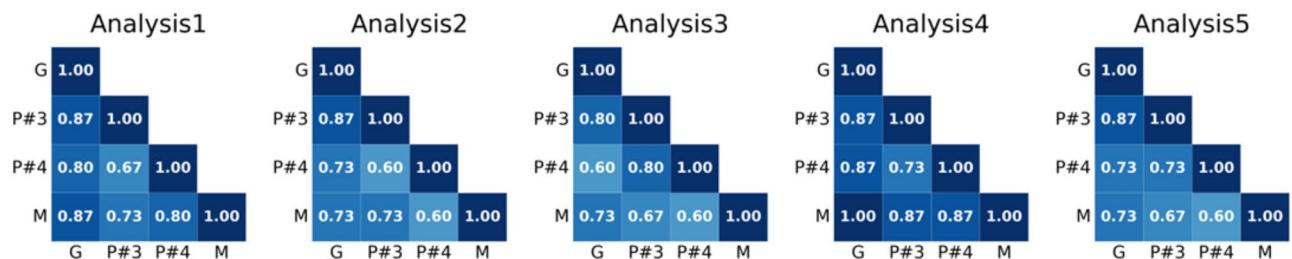
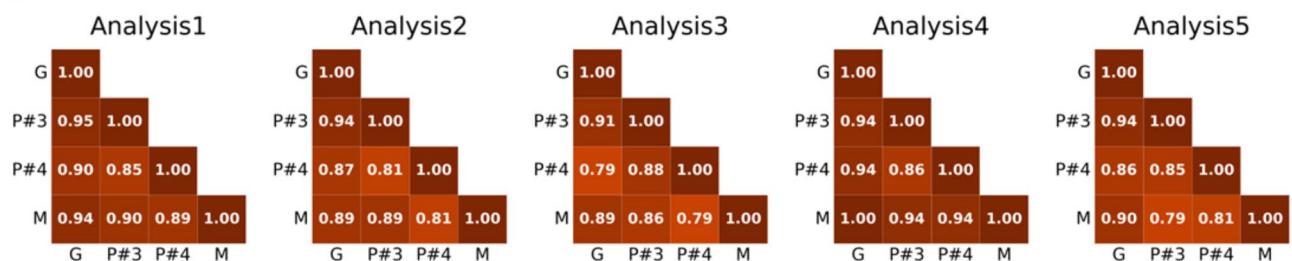
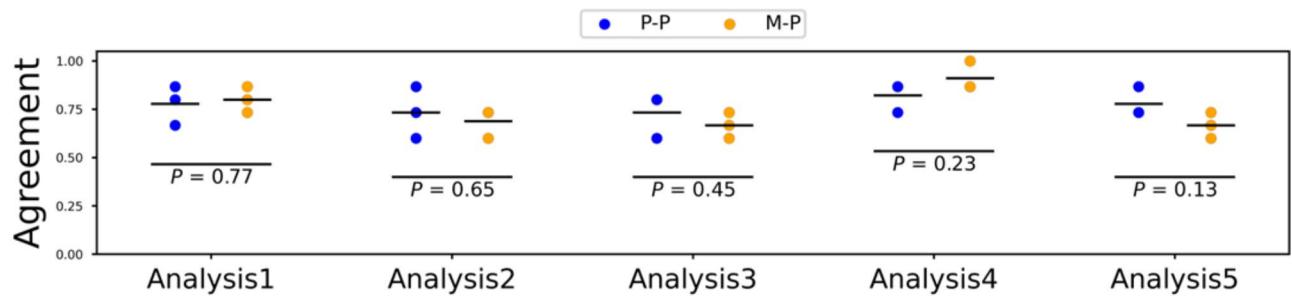
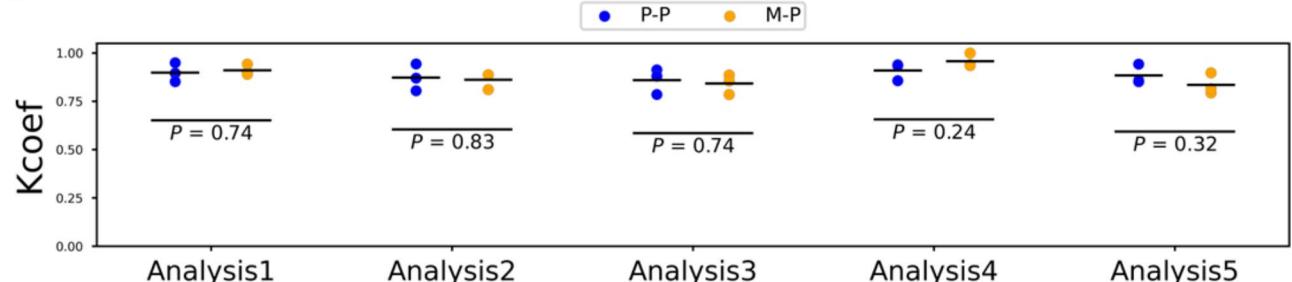
a**b****c****d**

Fig. 3. Model predictions compared with pathologist evaluations. **(a)** Agreement coefficient and **(b)** Cohen's kappa coefficient (Kcoef) for every pair. **(c)** Variations in agreement coefficient and **(d)** variations in Kcoef between pathologist–pathologist pairs (P–P) and machine–pathologist pairs (M–P) are shown. The Kcoef is based on the difference between the observed agreement (p_o) and the probability of chance (p_e) and is calculated as $p_o - p_e / 1 - p_e$. P values were calculated using Welch's t-test ($n=3$). G, P#3, P#4 and M indicate ground truth, pathologist #3, pathologist #4 and machine.

to underestimate grade, particularly predicting Grade 2 for data that pathologists evaluated as Grade 3 than the model with pre-selection. The model with patch pre-selection requires one input image for a single WSI, reducing computational cost.

Correlation prediction from the models and pathologists

To evaluate the performance of our models in comparison to pathologists, we examined the level of agreement coefficient and Kcoef (Fig. 3a and b). We further compared the variations in both agreement coefficient and Kcoef between pathologist–pathologist pairs and machine–pathologist pairs (Fig. 3c and d). No significant

difference was observed, indicating that our models' performance in grading lung fibrosis is on par with that of pathologists.

Discussion

BLM model is a widely used animal model to evaluate therapeutic targets for IPF^{5,17–19}. Histopathological analysis of lung fibrosis is an important method for evaluating BLM model. However, challenges such as inter- and intra-observer variations and disappointing throughput complicate this approach^{6,8}. We believe that digital technologies will be able to streamline and improve the evaluation process.

In this study, we successfully developed deep learning models that mimic evaluation by a pathologist. These models achieved an accuracy comparable to the variations among pathologists. For this dataset, WSIs from Sirius Red stained specimens prepared at multiple centers (Table 1) were used. Although there was a concern that this might affect the accuracy of the models, the verification results showed that the data from the facilities used in this study was consistent and acceptable for this study. However, our workflow should be validated further in order to be applied automatically to any stain of any center due to the limitation of $n=2$ centers in this study.

We carried out patch pre-selection before delving into deep learning. The process of human annotation presents several challenges, as it can be time-consuming and subject to variability²⁰. Meanwhile, self-supervised learning approaches, capable of capturing image features without the necessity for annotation, demand substantial data and computational resources²¹. To surmount these hurdles, we developed models capable of interpreting complex image patterns, akin to the abilities of pathologists. This was accomplished beyond the simple area extraction that the HALO AI image analysis software usually performs. This was motivated by our preliminary study showing that simple area extraction and pathologist evaluations did not sufficiently match. This discrepancy was attributed to the fact that pathologists did not grade solely based on the simple Sirius red-positive area. They also considered qualitative factors such as collagen texture and distribution. The pathologist's perspective has the potential to recognize drug possibilities that might be overlooked in efficacy evaluations based only on the size of the area. We think this qualitative capacity is more important than the operational efficiency and reproducibility provided by our simpler deep learning models.

In conclusion, we developed a deep learning workflow for assessing fibrosis grades in BLM model. This workflow emulates the pathologist's grading process, including qualitative assessment. The use of this workflow is expected to facilitate the development of drugs for pulmonary fibrosis.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Received: 21 May 2024; Accepted: 13 January 2025

Published online: 17 January 2025

References

1. Yang, H. et al. Clinical and pathological characteristics of immune mediated necrotizing myopathy. *Beijing da xue xue bao Yi xue ban J. Peking Univ. Health Sci.* **51**, 989–995 (2019).
2. Heukels, P., Moor, C., Von der Thüsen, J., Wijzenbeek, M. & Kool, M. Inflammation and immunity in IPF pathogenesis and treatment. *Respir. Med.* **147**, 79–91 (2019).
3. Ashcroft, T., Simpson, J. M. & Timbrell, V. Simple method of estimating severity of pulmonary fibrosis on a numerical scale. *J. Clin. Pathol.* **41**, 467–470 (1988).
4. Tashiro, J. et al. Exploring animal models that resemble idiopathic pulmonary fibrosis. *Front. Med.* **4**, 118 (2017).
5. Jenkins, R. G. et al. An official American thoracic society workshop report: Use of animal models for the preclinical assessment of potential therapies for pulmonary fibrosis. *Am. J. Respir. Cell Mol. Biol.* **56**, 667–679 (2017).
6. Hübner, R.-H. et al. Standardized quantification of pulmonary fibrosis in histological samples. *BioTechniques* **44**, 507–517 (2008).
7. Yin, W., Han, J., Zhang, Z., Han, Z. & Wang, S. Alopine protects mice against bleomycin-induced pulmonary fibrosis by attenuating fibroblast proliferation and differentiation. *Sci. Rep.* **8**, 6265 (2018).
8. Heinemann, F., Birk, G., Schoenberger, T. & Stierstorfer, B. Deep neural network based histological scoring of lung fibrosis and inflammation in the mouse model system. *PLOS ONE* **13**, e0202708 (2018).
9. Kumar, R. K. Morphological methods for assessment of fibrosis. *Methods Mol. Med.* **117**, 179–188 (2005).
10. Cherniack, R. M. et al. Quantitative assessment of lung pathology in idiopathic pulmonary fibrosis. The BAL Cooperative Group Steering Committee. *Am. Rev. Respir. Dis.* **144**, 892–900 (1991).
11. McConnell, E. E. & Davis, J. M. Quantification of fibrosis in the lungs of rats using a morphometric method. *Inhal. Toxicol.* **14**, 263–272 (2002).
12. Gilhodes, J. C. et al. Quantification of pulmonary fibrosis in a bleomycin mouse model using automated histological image analysis. *PLoS One* **12**, e0170561 (2017).
13. Segev, S. et al. A fully automated image analysis method to quantify lung fibrosis in the bleomycin-induced rat model. *PLOS ONE* **13**, e0193057 (2018).
14. Upagupta, C., Shimbori, C., Alsilmi, R. & Kolb, M. Matrix abnormalities in pulmonary fibrosis. *Eur. Respir. Rev.* **27**, 180033 (2018).
15. Biasin, V. et al. Meprin beta contributes to collagen deposition in lung fibrosis. *Sci. Rep.* **7**, 39969 (2017).
16. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Measure.* **20**, 37–46 (1960).
17. Moeller, A., Ask, K., Warburton, D., Gauldie, J. & Kolb, M. The bleomycin animal model: A useful tool to investigate treatment options for idiopathic pulmonary fibrosis?. *Int. J. Biochem. Cell Biol.* **40**, 362–382 (2008).
18. Bethany, B. M. et al. Animal models of fibrotic lung disease. *Am. J. Respir. Cell Mol. Biol.* **49**, 167–179 (2013).
19. Chua, F., Gauldie, J. & Laurent, G. J. Pulmonary fibrosis: Searching for model answers. *Am. J. Respir. Cell Mol. Biol.* **33**, 9–13 (2005).
20. Ehteshami Bejnordi, B. et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama* **318**, 2199–2210. <https://doi.org/10.1001/jama.2017.14585> (2017).
21. Xu, H. et al. A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**, 181–188. <https://doi.org/10.1038/s41586-024-07441-w> (2024).

Acknowledgements

We thank Masakazu Kanamori and Naoshi Horiba at Chugai Pharmaceutical Co., Ltd., as well as Hironori Yamagishi at Chugai Research Institute for Medical Science Inc. for their technical and scientific support. We also thank Jacob Davis at Chugai Pharmaceutical Co. Ltd. for proofreading the manuscript.

Author contributions

T.G. and A.S. designed the study, contributed to data interpretation, and reviewed the manuscript. S.O., N.H. and S.M. provided the original data. R.K., S.O., A.K. and H.M. supervised the writing of the manuscript. M.Y. provided the original data and supervised both the study and the writing of the manuscript. All authors read and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.S. or M.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025