Science Bulletin 70 (2025) 437–440

Contents lists available at ScienceDirect

Science Bulletin

journal homepage: www.elsevier.com/locate/scib

News & Views

Collaborative artificial intelligence and clinical evaluation of interstitial lung diseases: a call for interdisciplinary partnerships

Hongyi Wang [a,b], Rongguo Zhang [c], Xiaojuan Guo [d], Han Kang [e], Min Liu [b,f], Ulrich Costabel [g], Chen Wang [a,b], Huaping Dai [a,b,*]

[a] China-Japan Friendship Hospital (Institute of Clinical Medical Sciences), Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100029, China
[b] National Center for Respiratory Medicine; State Key Laboratory of Respiratory Health and Multimorbidity; National Clinical Research Center for Respiratory Diseases; Institute of Respiratory Medicine, Chinese Academy of Medical Sciences; Department of Pulmonary and Critical Care Medicine, Center of Respiratory Medicine, China-Japan Friendship Hospital, Beijing 100029, China
[c] Academy for Multidisciplinary Studies, Capital Normal University, Beijing 100048, China
[d] The Department of Radiology, Beijing Chaoyang Hospital of Capital Medical University, Beijing 100020, China
[e] Institute of Advanced Research, Infervision Medical Technology Co., Ltd., Beijing 100025, China
[f] The Department of Radiology, China-Japan Friendship Hospital, Beijing 100029, China
[g] Center for Interstitial and Rare Lung Diseases, Pneumology Department, Ruhrlandklinik, University of Duisburg-Essen, Essen, 45239, Germany

Interstitial lung diseases (ILD) encompass over 200 lung disorders marked by inflammation and/or fibrosis, posing significant diagnostic and treatment challenges [1]. The American Thoracic Society recommends a multidisciplinary approach for accurate ILD diagnosis [2]. In the real world, this multidisciplinary approach is time-consuming and resource-intensive. Clinicians hope to speed up the diagnostic process with the help of artificial intelligence (AI).

AI models have advanced ILD assessment by improving diagnostic efficiency and accuracy. However, different backgrounds of clinicians and data scientists cause inefficient collaboration in AI model development, which may hinder AI advancement and implementation. This paper offers a comprehensive overview of the varying roles in developing AI models for ILD management perceived by clinicians and data scientists, promoting further interdisciplinary partnerships.

What is the role of clinicians for the AI application in ILD? Data collection is the most fundamental task for clinicians. Accurate and comprehensive data enables precise evaluation. In addition to data collection, clinicians should also assume three key roles. First, clinicians provide new clinical issues. The research topics prevailing in contemporary AI studies predominantly stem from traditional clinical concepts such as diagnosis and prognosis, which may not cover all requirements in ILD management. Table S1 online presents a compilation of common classical clinical topics, indicating a trend wherein the novel issues often sprout as refinement or enhancement of specific classical topics. For example, conventional AI models can only classify patients into ILD or not ILD without considering the extent of fibrosis or other specific lesion (i.e. ground-glass opacities, and consolidation) on HRCT (high-resolution computed tomography) (Table S1 online). Later, the quantitative fibrosis score was explored as a useful measurement for ILD patients (Table S1 online). Besides improving AI models used in specific classical topics, redefining patients is the easiest way to facilitate innovation. For instance, an AI model for predicting prognosis in idiopathic pulmonary fibrosis may already exist [3], further research may also deal with developing prognostic AI models for any fibrotic ILD [4]. These AI models among redefined populations may help clinicians understand different ILD subgroups to provide a more precise management strategy (Table S1 online). Moreover, extending into new clinical decision-making scenarios often arises naturally from significant advancements within similar topics, which may also represent an another unmet need. For example, research on quantitative assessment of emphysema and lung volume using CT in patients with chronic obstructive pulmonary disease [5] can be applied to similar investigations in patients with pulmonary hypertension [6] and combined pulmonary fibrosis and emphysema [7] (Table S1 online). At times, algorithm-driven advancements can lead to new clinical decision requirements. For instance, the traditional ILD diagnostic pathway may not typically include the consideration of comorbidity. However, with the aid of deep learning, it may become possible to directly screen for ILD diagnoses from electronic records [8]. These algorithmic advancements expand the diagnostic possibilities and offer potential avenues for earlier detection and more accurate ILD assessments. By leveraging the power of deep learning algorithms, healthcare providers can enhance their diagnostic capabilities and potentially improve patient outcomes in the realm of ILD management (Table S1 online).

* Corresponding author.
E-mail address: daihuaping@ccmu.edu.cn (H. Dai).

https://doi.org/10.1016/j.scib.2024.12.035

Second, clinicians should demonstrate the efficacy of AI models in clinical cohorts. Noteworthy examples utilized quantitative CT and diagnostic classification models (Table S1 online). To investigate the role of quantitative CT in ILD diagnosis and severity assessment, a thorough collection of clinical data may be required, encompassing lung function measurements and pertinent established risk factors. Researchers primarily focus on examining the relationship between imaging findings and clinical outcomes, particularly in the context of fibrotic lung diseases such as idiopathic pulmonary fibrosis and progressive pulmonary fibrosis. Data-driven significance thresholds of quantitative CT can help to predict disease prognosis. Compared with quantitative CT, diagnostic or classification models that provide diagnosis conclusions directly from images without any intermediate steps cannot generate explainable features or lesion regions. Thus, clinicians are unable to estimate how AI models know the diagnosis or prognosis, which stops AI models from convincing. High-quality research frequently employs predictive probabilities linked to prognosis or lung function changes to demonstrate the clinical efficacy of AI models [9,10].

Third, clinicians should explain how models work from clinical perspectives. Clinicians are perplexed about how AI models can effectively work, and the confusion among clinicians often arises from a lack of algorithmic principles understanding [11] and disease mechanisms. Data scientists can only provide explanations of algorithmic principles. Understanding the disease mechanisms from algorithmic principles or comprehending algorithmic principles from disease mechanisms can only be achieved through the collaboration of data scientists and clinicians [12]. Recently, there have been notable efforts by clinicians towards achieving interpretability. For instance, in studying systemic sclerosis-related ILD, researchers quantified the fibrosis progression based on radiomics scores derived from HRCT, which reflects the molecular changes [13] (Table S1 online).

What is the role of data scientists in the application of AI in ILD? Building and optimizing models is undoubtedly one of the most important roles for data scientists, but it is not their only contribution. For data scientists, diagnosis is considered a classification task, which essentially involves the process of determining the category label to which the data belongs. According to Table S1 online, radiomics has long been one of the most classic and effective AI strategies for imaging-based diagnosis of ILD. Recent advancements have focused on emphasizing different anatomical or pathological regions as distinct volumes of interest, enabling more precise feature extraction. Traditionally, radiomics has been paired with various machine learning models, such as Multilayer Perceptron, Support Vector Machine, and eXtreme Gradient Boosting. More recently, deep learning models like Inception-ResNet-v2 and RadImageNet have entered the ILD classification landscape. The rise of attention mechanisms (like Vision Transformer) has further become the new frontrunner in ILD classification tasks. For classification tasks, the organization of imaging data and evaluation metrics have remained relatively consistent. At least, imaging data should include DICOM-formatted HRCT scans with corresponding labels. However, incorporating radiologist-annotated regions or volumes of interest or key clinical parameters such as pulmonary function could potentially enhance model performance. Evaluation metrics are typically based on binary classification tasks, as multiclass problems often need to be decomposed into multiple binary tasks. Key metrics include area under the receiver operator characteristic curve (AUC), accuracy, sensitivity, precision, F1, specificity, positive predictive value, and negative predictive value. To address class imbalance in multiclass tasks, macro- and micro-averaging are sometimes employed. AUC is the most valued statistic, often alongside DeLong's test to compare results between clinician experts and AI models or among different models. Similarly, integrated discrimination improvement index, net reclassification improvement index, decision curve analysis, and calibration curves are common practices in clinical binary prediction models. In addition to classification tasks, segmentation has emerged as a key focus in cutting-edge research, serving as a quantitative marker of disease severity or even a secondary outcome in randomized controlled trials. Because segmentation models remain relatively limited due to the challenge of accurately delineating lesion boundaries by radiologists, the data-derived texture analysis method remains dominant. Data-derived texture analysis classifies pixels to achieve effects similar to segmentation models. Data preparation of segmentation tasks is similar to classification tasks, but evaluation relies heavily on the Dice index combined with the association between quantitative markers and clinical outcomes (lung function or survival) considered to assess the segmentation's clinical value.

Secondly, algorithm development and optimization can be derived from domain knowledge to enhance inference and quantification of disease patterns [14]. For clinicians, a multidisciplinary discussion is used to summarize single conclusions to generate an overall diagnosis in the hospitals. This process organizes data in a specific order highly distinguishing from deep learning models that tend to identify a final label (often diagnosis) via analyzing all input data (often cannot be obtained at the initial visit) simultaneously. In the realm of algorithm development, ensemble learning integrating HRCT images with clinical information using a series of machine learning models aligns well with the multidisciplinary discussion design [15]. However, the challenge lies in determining the sequence and significance of different types of information in ILD diagnosis.

Thirdly, data scientists explain a model from a data perspective. Although various methods have been proposed in recent years to uncover the black box in deep neural networks, no convincing conclusions have been reached. The attention mechanism also aids in the interpretability of ILD-related AI models. Attention mechanism is a concept in machine learning that allows the model to focus on specific parts of the input data. The visualization of attention further enhances the visual and intuitive understanding of both the training process and the reasoning behind the AI model's decisions [16]. By visualizing the attention patterns, clinicians and researchers can gain insights into image areas or features that the model considers to be the most relevant to diagnosis or other tasks (Table S2 online). Visualization of an AI model can also make us see how AI models use input data to generate a final diagnosis, which may reflect pathology, or molecule features in ILD development. Through that visualization, the AI model can also assist clinicians in understanding ILD mechanisms in the future.

Typical successful collaboration cases. The development of the SOFIA (Systematic Objective Fibrotic Imaging Analysis Algorithm) (Table S1 online) model highlights a powerful collaboration between doctors and data scientists. Clinicians posed critical diagnostic challenges in classifying usual interstitial pneumonia (UIP), possible UIP, and inconsistent UIP, which are pivotal for fibrotic lung disease diagnosis. Data scientists responded by harnessing the capabilities of the Inception-ResNet-v2 model to effectively fit these complex imaging datasets. Through this interdisciplinary effort, the SOFIA model was created achieving human-level accuracy and offering reproducible diagnostic support in HRCT evaluations. This collaboration underscores the transformative impact of integrating AI into healthcare. Later, the SOFIA model not only achieves human-level accuracy in high-resolution CT evaluations but also maintains accuracy in predicting progressive pulmonary fibrosis, showcasing the potential of AI to enhance diagnostic precision and patient outcomes. Following its development, clinicians across the United States, Canada, and other regions continued collaborating with data scientists. These partnerships have led to the

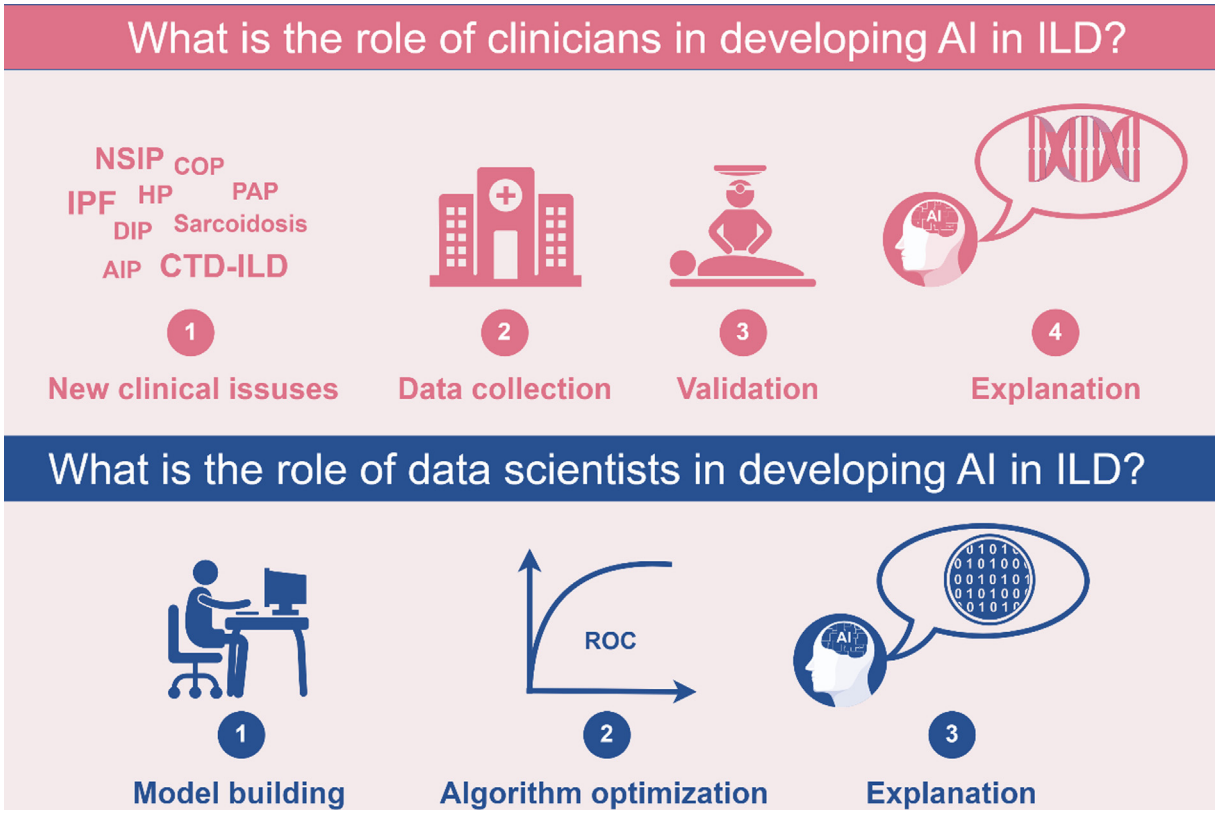

Fig. 1. Different roles of clinicians or data scientists in developing AI of ILD (created by FigDraw). For clinicians, providing new issues, validating models in clinical cohorts, and explaining models from a disease view are major roles except for data collection. For data scientists, building models, optimizing algorithms, and explaining models from data view are three roles in AI research for ILD applications. AI: artificial intelligence; AIP: acute interstitial pneumonia; COP: cryptogenic organizing pneumonia; CTD-ILD: connective tissue disease-associated interstitial lung disease; DIP: desquamative interstitial pneumonia; HP: hypersensitivity pneumonitis; ILD: interstitial lung disease; IPF: idiopathic pulmonary fibrosis; NSIP: nonspecific interstitial pneumonia; PAP: pulmonary alveolar proteinosis; ROC: receiver operator characteristic curve.

creation of a series of reliable AI-assisted diagnostic models (Table S1 online), further advancing the accuracy and efficiency of fibrotic lung disease diagnosis and enabling widespread access to sophisticated diagnostic tools.

The development of AI models in ILD faces significant challenges, including the labor-intensive and subjective nature of data annotation, limited model generalization across diverse patient populations, and pressing ethical concerns. Model generalization struggles with variability in patient demographics, clinical settings, and imaging protocols. As a result, data annotation often requires expert input, leading to inconsistencies and high costs. Ethical considerations, such as data privacy further complicate adoption. To address these issues, future efforts should focus on creating collaborative platforms that unify multidisciplinary expertise, and facilitating standardized data annotation protocols and shared datasets. Specific funding opportunities from government agencies, healthcare organizations, and AI industries can drive research into innovative solutions, such as semi-supervised learning and large models with attention mechanisms to reduce annotation dependency. Key research questions may include identifying approaches to ensure fairness across populations, integrating explainable AI to improve clinical trust, and establishing ethical guidelines for AI deployment in ILD. Challenging directions include using AI models to quantify molecular-level changes in the context of ILD and designing prognosis prediction algorithms that integrate the latest molecular biology discoveries. Technologically, the next advancements in ILD may focus on training, fine-tuning, or deploying multimodal large models. International multicenter collaboration will be crucial for enabling multi-ethnic studies, where diverse data from different ethnic groups, imaging devices, and imaging standards can further enhance AI-driven ILD diagnosis and quantification. Data scientists and clinicians have different roles in developing an AI model for ILD. To develop better AI models for ILD, it would be advantageous to have clinicians and data science specialists who are involved in ILD-related tasks working together as one team (Fig. 1). Collaborative applications developed for ILD have the potential to become powerful tools not only for the diagnosis and treatment of ILD but also for prevention and rehabilitation. Moreover, they can serve as valuable models for managing other chronic lung diseases, chronic conditions, and even life-cycle care.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2021YFC2500700 and 2016YFC0901101), and the National Natural Science Foundation of China (81870056).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scib.2024.12.035.

References

[1] Maher TM. Interstitial lung disease: A review. JAMA 2024;331:1655–65.
[2] Wijsenbeek M, Suzuki A, Maher TM. Interstitial lung diseases. Lancet 2022;400:769–86.

[3] Nam JG, Choi Y, Lee SM, et al. Prognostic value of deep learning-based fibrosis quantification on chest CT in idiopathic pulmonary fibrosis. Eur Radiol 2023;33:3144–55.
[4] Oh AS, Lynch DA, Swigris JJ, et al. Deep learning-based fibrosis extent on computed tomography predicts outcome of fibrosing interstitial lung disease independent of visually assessed computed tomography pattern. Ann Am Thorac Soc 2024;21:218–27.
[5] Makimoto K, Hogg JC, Bourbeau J, et al. Ct imaging with machine learning for predicting progression to copd in individuals at risk. Chest 2023;164:1139–49.
[6] Dwivedi K, Sharkey M, Delaney L, et al. Improving prognostication in pulmonary hypertension using AI-quantified fibrosis and radiologic severity scoring at baseline CT. Radiology 2024;310:e231718.
[7] Zhao A, Gudmundsson E, Mogulkoc N, et al. Mortality surrogates in combined pulmonary fibrosis and emphysema. Eur Respir J 2024;63:2300127.
[8] Onishchenko D, Marlowe RJ, Ngufor CG, et al. Screening for idiopathic pulmonary fibrosis using comorbidity signatures in electronic health records. Nat Med 2022;28:2107–16.
[9] Humphries SM, Thieke D, Baraghoshi D, et al. Deep learning classification of usual interstitial pneumonia predicts outcomes. Am J Respir Crit Care Med 2024;209:1121–31.
[10] Chung JH, Chelala L, Pugashetti JV, et al. A deep learning-based radiomic classifier for usual interstitial pneumonia. Chest 2024;165:371–80.
[11] Chen H, Gomez C, Huang C-M, et al. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. NPJ Digit Med 2022;5:156.
[12] Vora LK, Gholap AD, Jetha K, et al. Artificial intelligence in pharmaceutical technology and drug delivery design. Pharmaceutics 1916;2023:15.
[13] Le Gall A, Hoang-Thi TN, Porcher R, et al. Prognostic value of automated assessment of interstitial lung disease on CT in systemic sclerosis. Rheumatology (Oxford) 2024;63:103–10.
[14] Kumari S, Singh P. Deep learning for unsupervised domain adaptation in medical imaging: recent advancements and future perspectives. Comput Biol Med 2024;170:32.
[15] Mei X, Liu Z, Singh A, et al. Interstitial lung disease diagnosis and prognosis using an AI system integrating longitudinal data. Nat Commun 2023;14:2272.
[16] Parvaiz A, Khalid MA, Zafar R, et al. Vision transformers in medical computer vision—a contemplative retrospection. Eng Appl Artif Intell 2023;122:106126.