Science Bulletin 68 (2023) 648–652

Contents lists available at ScienceDirect

Science Bulletin

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

News & Views

How to keep artificial intelligence evolving in the medical imaging world? Challenges and opportunities

Huadan Xue [a,1], Ge Hu [b,1], Nan Hong [c,*], N. Reed Dunnick [d,*], Zhengyu Jin [a,*]

[a] Department of Radiology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, China
[b] Medical Research Center, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, China
[c] Department of Radiology, Peking University People's Hospital, Beijing 100044, China
[d] Department of Radiology, University of Michigan, Ann Arbor 48109, USA

Medical imaging is involved in all processes of clinical practice. Approximately 70% of diagnostic information originates from radiologic images, which also account for 90% of the digital data volume of a hospital. However, the number of physicians has increased only modestly compared to the rapid growth in the number of medical images. In China, data from medical imaging increases by 30% every year, whereas the number of radiologists increases by only 4% annually. Artificial intelligence (AI), which is representative technology of the fourth industrial revolution, may alleviate the increasing pressure and job burnout, and further improve the diagnostic efficiency of radiology services [1]. Despite the urgent and realistic demand for AI technology, many challenges remain in the development and translation of AI products. The rate of the scientific translation of AI research into clinical applications is extremely low. Furthermore, AI models that are applied in clinical settings exhibit unreliable performance and are often impractical [2]. Therefore, radiologists may not have access to suitable medical imaging AI models to solve specific clinical problems. This paper analyzes and discusses this problem according to two aspects: the data sources and the AI algorithm (Fig. 1).

Medical imaging data exhibit the qualities of big data, such as diverse types, frequent updates, a large scale, and complex processing methods. Thus, radiology is expected to be one of the first specialties to take full advantage of AI and to be most affected by developments therein [3]. However, medical imaging data also have an obvious long-tail effect; that is, most diseases are small data that are scattered in different centers, thereby forming "data islands" that lack effective interoperability. As existing AI technology remains data-driven, the training data of a single center cannot satisfy the AI performance requirements [2]. AI models should ideally be trained by combining data from multiple centers; however, barriers to data sharing often exist.

The construction of multimodal medical image databases (or datasets) with large sample sizes is the primary means of solving the problem of data islands, but many challenges are faced in this process. First, database construction is a time- and resource-consuming task, as each case requires the mutual cooperation of patients, physicians, technicians, engineers, and other information technology experts, and major investments in manpower as well as financial and material resources are necessary [1]. Second, there is a high technical threshold for database construction. The process varies substantially depending on the acquisition, cleaning, annotation, and loading of medical images from different sites. The construction and management of databases with different goals (model training, performance testing, clinical evaluation, and quality control) may also differ significantly. Finally, the database must be constantly adjusted to dynamic changes in clinical and social needs to preserve its expected value [4,5].

A medical image database should be deployed uniformly at the national level, and led by a multidisciplinary team of experts with strong professional and organizational abilities. The framework should be designed from top to bottom, and all aspects of the database construction should be described by the standards or expert consensus [5]. A particular challenge in establishing standards for a medical image database is image annotation. High-quality and trustworthy human-generated labels are a time-consuming, labor-intensive, and expensive process. Any biases in the process can be transferred to the outcomes of the AI systems. Fortunately, an integrated iterative annotation technique proposed by previous research has the potential to solve this problem. Through the interaction between humans and automatically generated annotations and the "human-in-the-loop" strategy, the labeling ability of this technology was validated on a dataset of prostate glands while reducing the annotation burden. Another problem with image labeling is the inconsistency between expert consensus. Take pulmonary nodule annotation as an example, some researchers eliminated the region corresponding to vessels, bronchus, and air from the nodule, but others only delineated the outer contours. The solution to this problem requires further exploration on the impact of annotation details on downstream tasks (Text S1 online).

At present, the construction of large-sample medical image databases is the consensus among AI research teams globally [6]. Many public image datasets (or databases), such as the National Lung Screening Trial (NLST) dataset, Liver Tumor Segmentation (LiTS) dataset, Multi-Modality Whole Heart Segmentation

[*] Corresponding authors.
E-mail addresses: hongnan1968@163.com (N. Hong), rdunnick@med.umich.edu (N. R. Dunnick), jinzy@pumch.cn (Z. Jin).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.scib.2023.03.031
2095-9273/© 2023 Science China Press. Published by Elsevier B.V. and Science China Press. All rights reserved.

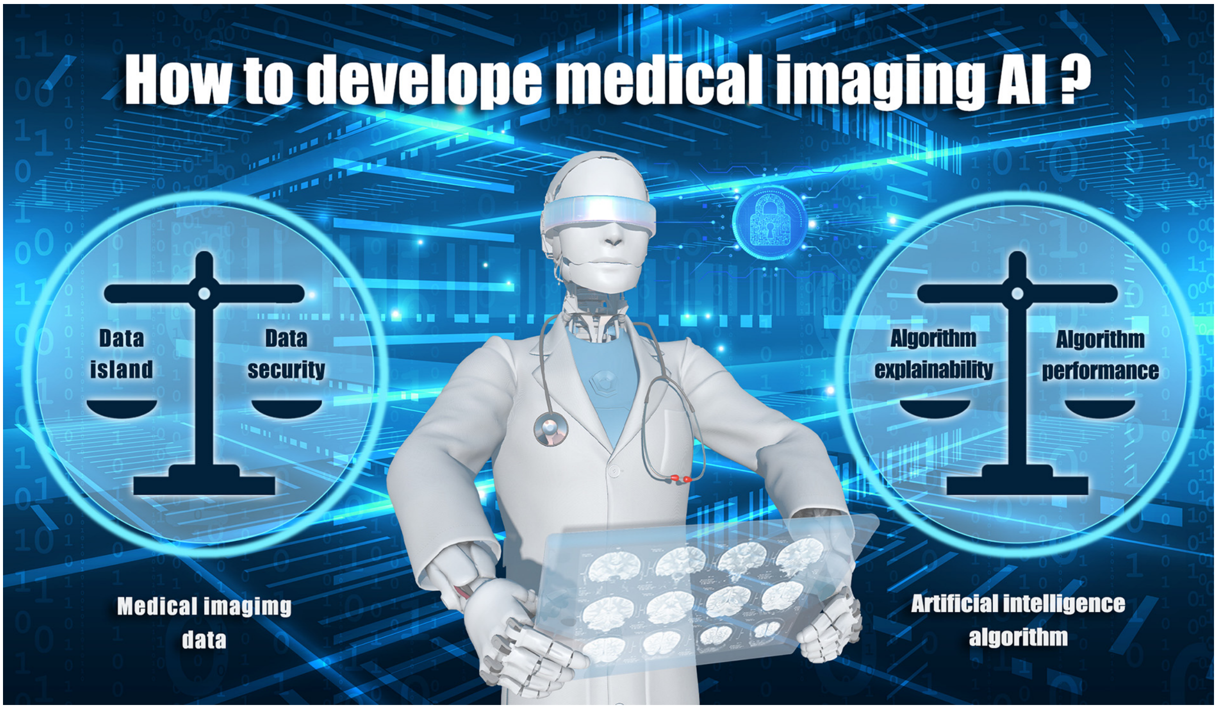


Fig. 1. How to develop medical imaging AI.

(MMWHS) dataset, The Cancer Imaging Archive (TCIA) database, MedPix database, and National COVID-19 Chest Imaging Database (NCCID), provide enormous image resources to researchers who are interested in medical imaging AI. In July 2022, the National Health Commission Capacity Building and Continuing Education Center of China announced the establishment of the Radiographic Image Database Construction Project, which officially launched the systematic construction of the medical imaging databases in China. These large databases, which have already been built or are currently under construction, offer opportunities for solving data island problems. The NLST dataset had been extensively used for the development of machine learning applications. Based on the NLST dataset, researchers constructed a risk prediction model called the Lung Cancer Prediction Convolutional Neural Networks (LCP-CNN) to discriminate malignancy in pulmonary nodules and validated its performance on external validation. The NCCID is a centralized database containing mainly chest X-rays and computed tomography scans from patients with COVID-associated respiratory syndrome. This database addressed many common pitfalls highlighted in a recent meta-analysis of COVID-19 imaging models, and will be used to support the development of machine learning (ML) technologies (Text S2 online).

Most traditional multi-center databases adopt centralized storage and analysis, and use desensitization or de-identification to protect human subject privacy in the sample collection stage. In this mode, full images or processed copies are transferred from one center to another, thereby drawing attention to data security. Issues arising from data security can be solved through relevant policies or AI technologies.

Laws and regulations relating to personal information protection are constantly improved owing to the increasing emphasis on privacy protection. However, the focus on data security has exacerbated the problems that are caused by data islands. Therefore, many policies and regulatory standards for data sharing and privacy have been established worldwide, such as the EU's General Data Protection Regulation (GDPR), the US's Health Insurance Portability and Accountability Act (HIPAA), and China's Personal Information Protection Law (PIPL). The GDPR requires that participants should be provided with explicit information regarding how the personal data will be used, how long it will be used, who will have access to the data, and whether the data will be shared anonymously. Similar to GDPR, HIPAA also allows for research use, disclosure, and data sharing with participant consent. The PIPL clearly requires processors to take corresponding security technologies, such as encryption and de-identification, to ensure the security of personal information (Text S3 online).

The challenges of collecting data from data islands while maintaining data security can also be addressed at a technical level. Federated learning (FL), which was developed by Google [7], is a typical cryptographic distributed ML technique. FL can effectively realize multisite collaborations while satisfying privacy protection and data security requirements by establishing a data federation. As the most common malignant primary brain tumor, the incidence of glioblastoma is extremely low (3/100,000), which means that large and diverse images can hardly be collected to develop robust and generalizable AI models. Researchers conducted a large FL study, involving thousands of multi-parametric MRI scans from multiple centers, to develop an automatic tumor boundary detector for glioblastoma by only sharing numerical model updates. The latest progress in FL is the "No-free-lunch" law that first reveals the intrinsic constraint between model utility and privacy protection of FL from an information-theoretic perspective. By using the "No-free-lunch" law, the security, utility, and efficiency of trustworthy FL can be coordinated while ensuring data privacy protection. This provides a new opportunity to solve the contradiction of medical image information islands and data security (Text S4 online).

However, when the imaging standards are not uniform, it is also difficult to perform multi-center FL. This brings the issue of image standardization. The current diversity of medical image data acquisition standards as well as the lack of a unified understanding of imaging signs create barriers to the interaction between AI models and medical image data [8]. The only solution to this problem is the establishment of unified image standards.

In 2022, the Recommended Practice for Quality Management of Datasets for Medical Artificial Intelligence, led by the National Institutions for Food and Drug Control of China, was officially released by the IEEE Standards Association, thereby becoming the first global standard in the field of AI medical datasets. In 2020, an article on the preparation of medical imaging data for AI algorithm development was published in *Radiology*. Researchers describe a standard process of labeling, curating, and sharing medical image data for ML in the article. In 2016, the international organization FORCE11 formally proposed the FAIR guidelines to provide data assurance for AI research by standardizing the description and traceability of medical imaging data acquisition, processing, and management. These specifications are expected to solve the problem of the inconsistent standardization of clinical image data, and to assist in developing medical imaging AI models and providing translational products in the process of clinical practice (Text S5 online).

It appears that the model performance can be improved significantly if the AI programs are provided with sufficient high-quality image data. However, the real world is a huge and unpredictable open set, and we cannot exhaust all possibilities on the "chessboard" as with AlphaGo. Therefore, when the training data are limited, the perspective should be shifted to more central AI algorithms.

At present, the most important problem with AI algorithms is interpretability (also known as transparency), whereby attempts are made to reveal the mechanisms behind AI systems with a black-box nature [9]. The explainability crisis that results from the enormous code and complicated structure of ML models, particularly deep learning (DL) models, raises human concerns regarding the use of AI in high-stakes scenarios such as healthcare [10]. First, it is difficult to interpret the relationships between image biomarkers and clinical endpoint events for disease diagnosis and treatment; thus, imaging AI applications are highly susceptible to being challenged by experts. Second, physicians are primarily interested in the diagnosis or prediction of patient prognoses based on clinical images. However, the current unexplainable technologies result in weak engagement with doctors. Meanwhile, owing to unknown internal principles of AI, the design, optimization, and upgrading of the models are dependent on the experience of IT experts or engineers; therefore, it is difficult for AI to move from the inherent pattern of training data and accurately respond to open data in the real world. Finally, inexplainable AI raises several social issues, such as the definition of ethics and morality, patient-clinician relationships, the legal responsibility for medical errors, and medical humanistic care [11].

Two main approaches to interpretability are used at present: inherent explainability and post-hoc explainability [10]. Representative examples of inherent interpretability include models or features with explicit definitions and formulas that are closely related to the semantic descriptions of lesions in diagnostics, and thus, can be used to approximate their potential biological meaning. For example, Bayesian inference networks encapsulate expert knowledge for the generation of differential diagnoses in brain MRI and vessel tortuosity by measuring the abnormal shape of tumor-associated vasculature in breast MRI. Post-hoc interpretability focuses on explainable technologies that aim to dissect the decision-making procedure of a model. For example, researchers interpreted the variability in the importance of different anatomic regions (carpus, thumb, and metacarpophalangeal joint) in predicting the bone age in hand radiographs using gradient-weighted class activation mapping (Grad-CAM), which is a typical technology of heat maps (or saliency maps). Furthermore, it was found that models rely on regions outside the lung fields, such as laterality markers, to make predictions regarding COVID-19, through generative adversarial networks (GANs) and counterfactual explanations (CE). In one study, researchers used radiology–pathology coregistration to explain the biological rationale behind the radiomic features that are used to predict the outcomes of non-small cell lung cancer (NSCLC) patients, they found that peritumoral Gabor features were associated with the density of tumor-infiltrating lymphocyte (TIL) on diagnostic biopsy samples. Other approaches, such as agent interpretation and importance ranking interpretation, have also facilitated the understanding of the logic behind medical imaging AI (Text S6 online).

Despite the rapid development of interpretable AI, two questions remain to be addressed when discussing the limitations of the current technologies. First, is the "relationship" that is inferred by AI and that we wish to explain really correct? Researchers found that biases in the training datasets may cause spurious correlations between predictors and outcomes. Patients with severe COVID-19 typically receive chest X-rays in a supine or recumbent position, whereas healthier patients undergo imaging in an upright position. Such datasets will result in spurious correlations when predicting COVID-19 severity based on the position rather than semantic image features [12]. Second, does the current interpretable technique truly explain the logic behind the black box? In one study, researchers quantitatively evaluated seven saliency maps across multiple AI architectures, they found that although Grad-CAM could generally localize pathologies in chest X-rays more effectively than other heat maps, there was still a large gap in the localization performance between Grad-CAM and experts, particularly for pathologies with smaller sizes and complex shapes. This demonstrates that caution should be exercised when leveraging common explainable approaches to understand AI models (Text S7 online).

The "human-centered design" that was introduced by Chen et al. [9] in a recent systematic review is a promising means of overcoming the above limitations. Chen et al. [9] proposed the INTRPRT guideline, which is a design directive for explainable medical imaging AI systems. Human-centered design principles recommended formative user research as the first step towards understanding user needs and domain requirements. For example, unexpected correlations that are caused by the dataset biases of chest X-rays, as we discussed in the previous section, may be anticipated if a radiologist is included in the multidisciplinary team that designs the model. In one study, the utility of potential explanatory information in AI was assessed using a user-centered iterative design system to enable physicians to understand the AI analysis tool for chest X-rays. Similarly, in another study, target users were consulted in the design of an image retrieval system for medical decision making and a system that preserves human agency was developed to guide the search process (Text S8 online).

Having analyzed the importance of interpretability and current explainable techniques, we pose another question: is the interpretability of AI still necessary when we need to provide a diagnosis or treatment for a disease in a short time (for example, COVID-19)? In traditional research and development process, a new drug must undergo multiple stages of clinical trials and data collection to become commercially available. However, when faced with sudden public health events that urgently require therapy, the advantages and disadvantages of accelerating the launch of new drugs require careful consideration. The interpretability of an AI algorithm is similar in that its advantages and disadvantages must be weighed.

Many researchers have proposed that the performance of AI in the real world should receive more attention than its interpretability [10]. A case in point is acetaminophen. Despite the mechanism for how acetaminophen works remains only partially understood, it still has been used for more than a century due to its extensive validation in numerous trials. Furthermore, although many explainable techniques have been developed to provide a broad description of AI systems, these explanations are unreliable or provide only a superficial level of interpretation in specific cases [10]. For example, researchers developed a CheXNet model that reached a radiologist-level detective accuracy of pneumonia at chest X-rays [13]. By using a saliency map approach, CheXNet takes a chest X-ray as input, and outputs the areas in the image most indicative of the pathology. However, the model may make decisions based on features that cannot be identified by human eyes, such as pixel-level characteristics. Meanwhile, it is hard to confirm whether the highlighted areas in the heat map (e.g., airspace opacity) really have an important predictive effect on the results, and whether other areas (e.g., heart border or pulmonary artery) that are not shown are really of no value in the decision. In most cases, the black-box only caused comprehension difficulties but did not affect validations or practical applications [14]. Thus, "an essential caveat is why the original model is needed at all if better models are available" [8].

Technical bottlenecks and compromises based on practical application requirements are encountered when exploring AI interpretability, but one criterion is clear: AI models must have good generalizability [5]. This generalizability is reflected in the repeatability or reproducibility of the model performance as well as in the portability [15]. To date, most studies that have evaluated AI applications have not been vigorously validated for reproducibility, following well-defined processes and testing standards to develop high-quality AI systems may be the most executable and actionable solution. Recently, the Microsoft Research teams identified a linear nine-stage ML workflow informed by prior experiences developing AI applications (e.g., natural language processing) and data science tools (e.g., bug reporting). Researchers from NASA similarly proposed a Machine Learning Technology Readiness Levels (MLTRL) framework to simplify the ML workflows and to produce robust, reliable, and responsible AI models. All of these frameworks can be introduced into the healthcare domain, providing opportunities for further development of medical imaging AI (Text S9 online).

In this paper, we provide an overview of the challenges and opportunities of medical imaging AI from the perspective of two major factors: the data sources and the AI algorithm. By reviewing previous studies, we think that an important development direction of AI in the future is to realize real brain-like intelligence. Although the current single-mode learning of AI is still not perfect, its continued development is expected to gradually replace the single-mode human work, such as ChatGPT, which can already replace some language-based work. Thus, multi-modal, multi-channel, and multi-dimensional learning and understanding of information is a challenging task that arouses researchers' attention. The same goes for medical imaging AI. Most existing AI algorithms are "single disease", "single device", or even "single site" models, which cannot meet the actual needs of radiologists for medical image interpretation. Therefore, many challenges remain. How can a model be expanded from a single disease to a complex disease? How can different types of image data be handled? How should the model include additional imaging devices? How can it achieve deeper "comprehensive" and efficient integration with clinical work? These questions will be the key directions of the future development of medical imaging AI. As a final example, it is not sufficient for AI to detect only pulmonary nodules in patients who have undergone a chest CT examination. Only when diseases of the entire thorax can be detected can the clinical and social requirements truly be satisfied [16].

There is still a huge gap between AI and human intelligence, but we believed that the collaboration between academia and industry will play an important role in creating useful medical imaging AI products with good generalizability [17]. The development of AI will also constantly broaden human imagination and unleash greater value in healthcare.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

This work was supported by the Scientific and Technological Innovation 2030—a Major Project of New Generation Artificial Intelligence (2020AAA0109503), the Beijing Science and Technology Planning Project (Z201100005620008 and Z211100003521007), the National Natural Science Foundation (82171934), and the National Key Research and Development Program of China (2021ZD0111105).

Appendix A. Supplementary materials

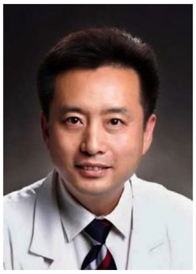
Supplementary materials to this news & views can be found online at https://doi.org/10.1016/j.scib.2023.03.031.

References

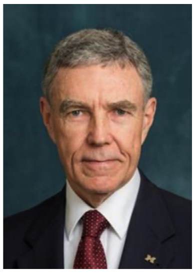
[1] Bi WL, Hosny A, Schabath MB, et al. Artificial intelligence in cancer imaging: clinical challenges and applications. CA-Cancer J Clin 2019;69:127–57.
[2] Desai AN. Artificial intelligence: promise, pitfalls, and perspective. JAMA 2020;323:2448–9.
[3] Santomartino SM, Siegel E, Yi PH. Academic radiology departments should lead artificial intelligence initiatives. Acad Radiol 2022. https://doi.org/10.1016/j.acra.2022.07.011.
[4] Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. New Engl J Med 2021;385:283–6.
[5] Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ 2020;368:l6927.
[6] Denny JC, Collins FS. Precision medicine in 2030—seven ways to transform healthcare. Cell 2021;184:1415–9.
[7] Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. NPJ Digit Med 2020;3:119.
[8] Bera K, Braman N, Gupta A, et al. Predicting cancer outcomes with radiomics and artificial intelligence in radiology. Nat Rev Clin Oncol 2022;19:132–46.
[9] Chen H, Gomez C, Huang C, et al. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. NPJ Digit Med 2022;5:156.
[10] Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 2021;3:e745–50.
[11] Morley J, Floridi L. An ethically mindful approach to AI for health care. Lancet 2020;395:254–5.
[12] Huang S, Chaudhari AS, Langlotz CP, et al. Developing medical imaging AI for emerging infectious diseases. Nat Commun 2022;13:7060.
[13] Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 2018;15:e1002686.
[14] Ma Y, Xiong J, Zhu Y, et al. Deep learning algorithm using fundus photographs for 10-year risk assessment of ischemic cardiovascular diseases in China. Sci Bull 2022;67:17–20.
[15] Sohn E. The reproducibility issues that haunt health-care AI. Nature 2023;613:402–3.
[16] Yacoub B, Kabakus IM, Schoepf UJ, et al. Performance of an artificial intelligence-based platform against clinical radiology reports for the evaluation of noncontrast chest CT. Acad Radiol 2022;29:S108–17.
[17] Spilseth B, McKnight CD, Li MD, et al. AUR-RRA review: logistics of academic-industry partnerships in artificial intelligence. Acad Radiol 2022;29:119–28.

Huadan Xue is the Deputy Director of the Department of Radiology, Peking Union Medical College Hospital, the Director of the Department of Radiology, Peking Union Medical College Hospital, Xidan Branch, the Deputy Secretary-General of Chinese Society of Radiology, and the Committee Member of International Society of Strategic Studies for Radiology. Her research interest focuses on diagnostic imaging of digestive system and female reproductive system, especially on pancreatic diseases.

Ge Hu is an Assistant Researcher at the Medical Research Center, Peking Union Medical College Hospital. She received her Ph.D. degree from Capital Medical University and finished post-doctoral research at Department of Radiology, Peking Union Medical College Hospital. Her research focuses on medical image processing and analysis based on radiomics and deep learning.



Nan Hong is the Vice President of Peking University People's Hospital, Chinese Society of Radiology, Magnetic Resonance Applications Professional Committee of China Association of Medical Equipment, the President Elect of Beijing Society of Radiology, and the Secretary-General of the Chinese Radiologist Association. His research focuses on diagnostic medical imaging of central nervous system, especially the structure and function of magnetic resonance in mental and neurological diseases.



Nicholas Reed Dunnick is the Chair Meritus of the Department of Radiology at the University of Michigan, and Editor-in-Chief of *Academic Radiology*. He has served as the President of Society of Uroradiology and Society of Computed Body Tomography/Magnetic Resonance. He has also served as the President of American Roentqen Ray Society, American Board of Radiology, Academy of Radiology Research, Society of Chairs of Academic Radiology Departments, Association of University Radiologists and Radiological Society of North America. His research focuses on diagnostic oncology and uroradiology.

Zhengyu Jin is the Chair of the Department of Medical Imaging and Nuclear Medicine, Peking Union Medical College, the President of the Chinese Radiologist Association, the Director of Medical Imaging Research Center, Chinese Academy of Medical Sciences, and National Quality Control Center of Radiological Imaging. He has served as the Director of the Department of Radiology, Peking Union Medical College Hospital and the President of Chinese Society of Radiology. His research focuses on diagnostic medical imaging and interventional therapy.