



Editorial

Enhancing Dataset Quality for AI in Radiology: Challenges and Solutions

Vinayak Rengan¹ Devansh Lalwani² Swapnil Bhat³ Pravin Meenashi Sundaram⁴

¹Department of Pediatric Surgery, SMS Medical College, Jaipur, Rajasthan, India

²Seth GS Medical College & KEM Hospital, Mumbai, Maharashtra, India

³AI Researcher, Miko.ai, Mumbai, Maharashtra, India

⁴Department of Renal Transplant Surgery, Sheffield Kidney Institute, Sheffield Teaching Hospitals, Sheffield, United Kingdom

J Gastrointestinal Abdominal Radiol ISGAR 2025;8:3–4.

The quality of datasets used in artificial intelligence (AI) applications in radiology is pivotal for the development of robust and accurate AI models. Various issues impacting dataset quality have been identified, affecting the performance and generalizability of AI tools in clinical settings. First, many datasets lack diversity in demographic, geographic, genetic, and disease representation. This inclusivity deficit can lead to AI models that perform poorly across different populations, potentially exacerbating health disparities.¹ Second, label accuracy in datasets is a significant concern. Studies have shown that labels in some large public datasets do not accurately reflect the visual content of images, leading to AI models trained on incorrect data and thus compromising their reliability.²

Moreover, the preparation and curation of medical imaging data are both costly and time-intensive. Many datasets are derived from small sample sizes and limited geographic areas, resulting in AI algorithms with poor generalization capabilities outside their training environments.³ Additionally, ethical considerations in dataset construction are often overlooked, leading to biases related to patient information, capture conditions, and class imbalances. These biases can affect AI model performance and raise ethical concerns regarding their use in clinical practice.⁴

Addressing these issues requires a concerted effort to improve data curation practices, enhance dataset diversity and accuracy, and incorporate ethical considerations into dataset development for AI applications in radiology. Several strategies can be employed. Enhancing dataset diversity and representation is crucial. Ensuring that datasets are diverse and representative of various demographics, disease states, and imaging modalities can improve AI model generalizability. This involves collecting data from a wide range of geographic locations and patient populations.⁵

Improving data annotation and curation is also vital. Accurate and expert-level annotation of imaging data can be achieved by involving multiple radiologists in the annotation process and using consensus or adjudication methods to resolve discrepancies. Employing longitudinal and multimodal datasets can provide richer information for training AI models.⁵ Utilizing advanced machine learning techniques such as self-supervised learning, federated learning, and multimodal learning can address issues related to limited annotated data and data privacy. These methods allow for learning from unlabeled data and combining different types of data (e.g., clinical and imaging data), enhancing the robustness of AI algorithms.⁵

Adopting ethical and bias-reduction practices is essential. Implementing strategies to identify and mitigate biases in datasets includes using tools for ethical analysis and ensuring that data collection and curation processes are transparent and adhere to ethical standards.⁵ Data augmentation techniques can also help overcome the issue of scarce data for certain conditions or imaging types by artificially increasing the size and diversity of training datasets through transformations or generating synthetic data.³

Federated learning (FL) in radiology is a promising approach to enhance dataset diversity and representation while addressing issues related to limited annotated data and data privacy. In radiology, FL allows for the collaborative training of machine learning models using data from multiple institutions without the need to share sensitive patient data directly. This method leverages a larger, more diverse dataset that includes various imaging modalities, patient demographics, and disease characteristics, which are crucial for developing robust AI models.^{6–9} FL is particularly beneficial in environments where data privacy is paramount and regulations may restrict the sharing of medical data across borders or

Address for correspondence
Devansh Lalwani, MBBS, Seth GS
Medical College & KEM Hospital,
Acharya Donde Marg, Mumbai,
Maharashtra 400012, India
(e-mail: devanshalalwani@gmail.com).

DOI <https://doi.org/10.1055/s-0044-1790232>.
ISSN 2581-9933.

© 2024. The Author(s).

This is an open access article published by Thieme under the terms of the Creative Commons Attribution License, permitting unrestricted use, distribution, and reproduction so long as the original work is properly cited. (<https://creativecommons.org/licenses/by/4.0/>)

Thieme Medical and Scientific Publishers Pvt. Ltd., A-12, 2nd Floor, Sector 2, Noida-201301 UP, India

institutions. By training models locally on the data available at each site and only sharing model updates or parameters, FL can significantly reduce privacy and security risks.¹⁰ Moreover, FL can improve AI model performance on heterogeneous data, which is often a challenge when models are trained on data from a single source. This approach enhances the generalizability of AI models across different institutions while maintaining the confidentiality and integrity of the data.¹¹

In summary, improving the quality of datasets used in AI applications in radiology involves a multifaceted approach. By enhancing dataset diversity, improving data annotation and curation, utilizing advanced machine learning techniques, adopting ethical practices, and leveraging FL, the reliability and clinical applicability of AI tools can be significantly improved.

Funding

None.

Conflict of Interest

None declared.

References

- 1 Tripathi S, Gabriel K, Dheer S, et al. Understanding biases and disparities in radiology AI datasets: a review. *J Am Coll Radiol* 2023;20(09):836–841
- 2 Oakden-Rayner L. Exploring large-scale public medical image datasets. *Acad Radiol* 2020;27(01):106–112
- 3 Willemink MJ, Koszek WA, Hardell C, et al. Preparing medical imaging data for machine learning. *Radiology* 2020;295(01):4–15
- 4 Arias-Garzón D, Tabares-Soto R, Bernal-Salcedo J, Ruz GA. Biases associated with database structure for COVID-19 detection in X-ray images. *Sci Rep* 2023;13(01):3477
- 5 Hong GS, Jang M, Kyung S, et al. Overcoming the challenges in the development and implementation of artificial intelligence in radiology: a comprehensive review of solutions beyond supervised learning. *Korean J Radiol* 2023;24(11):1061–1080
- 6 Arias-Garzón D, Tabares-Soto R, Bernal-Salcedo J, Ruz GA. Biases associated with database structure for COVID-19 detection in X-ray images. *Sci Rep* 2023;13(01):3477
- 7 Darzidehkalani E, Ghasemi-Rad M, van Ooijen PMA. Federated learning in medical imaging. Part I: toward multicentral health care ecosystems. *J Am Coll Radiol* 2022;19(08):969–974
- 8 Guan H, Yap PT, Bozoki A. Federated learning for medical image analysis. A survey. *Liu M. Pattern Recognition* 2024;151:110424
- 9 Arasteh ST, Kuhl C, Saehn MJ, et al. Enhancing domain generalization in the AI-based analysis of chest radiographs with federated learning. *Sci Rep* 2023;13(01):22576
- 10 Aouedi O, Sacco A, Piamrat K, Marchetto G. Handling privacy-sensitive medical data with federated learning. challenges and future directions. *IEEE J Biomed Health Inform* 2023;27(02):790–803
- 11 Yan R, Qu L, Wei Q, et al. Label-efficient self-supervised federated learning for tackling data heterogeneity in medical imaging. *IEEE Trans Med Imaging* 2023;42(07):1932–1943