Advocating for Open-Access Imaging Datasets: Building the Foundation for Transparent and Reproducible AI Innovation

Colin Wu[1], Youyou Cheng[1], Edward Pettyjohn, MD[2], Sina Ashraf[1], Justin Kuyn[3], Ravi Patel[1], Zhuangwei Kang, PhD[4], Sarah Pettyjohn, MD[5].

[1]University of the Incarnate Word School of Osteopathic Medicine; [2]University of Illinois College of Medicine Peoria (UICOMP); [3]University of Colorado Boulder; [4]Independent Researcher; [5]Rutgers Robert Wood Johnson Medical School

Purpose

To outline how standardized, open-access imaging datasets can lower regulatory barriers, improve scientific transparency, and serve as the foundation for reproducible AI development in radiology.

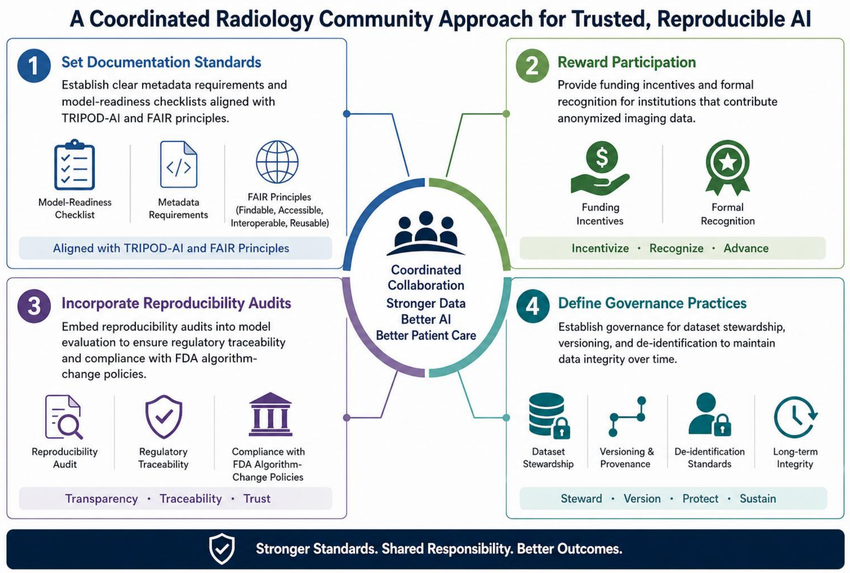
Methods/Materials

Most imaging-AI models are still built on small, private datasets that rarely leave a single institution. The result is familiar: limited generalizability, weak external validation, and slower regulatory review. Each team repeats the same work of cleaning data, writing scripts, and revalidating models at a significant cost of time and money. National programs such as the ACR Data Science Institute, NIH Bridge2AI, and the Cancer Imaging Archive have already shown that open, de-identified data can change this dynamic by enabling safer, faster, and more transparent AI research. What's missing is a consistent framework that makes open-data sharing both practical and rewarding across radiology subspecialties.

Results

We propose a coordinated ACR-led approach that would:

1. **Set documentation standards with clear metadata requirements and model-readiness checklists aligned with TRIPOD-AI and FAIR principles.**
2. **Reward participation through funding incentives and formal recognition for institutions that contribute anonymized imaging data.**
3. **Incorporate reproducibility audits into model evaluation to ensure regulatory traceability and compliance with FDA algorithm-change policies.**
4. **Define governance practices for dataset stewardship, versioning, and de-identification to maintain integrity over time.**



Conclusions

Standardized open datasets would reduce redundant data collection, cut retraining costs, and make studies easier to compare. Reliable, reusable data also improves model validation and helps regulators monitor performance once algorithms enter clinical use. Open imaging data is the groundwork for credible, scalable AI in radiology. With the right standards and incentives, ACR can help make open imaging datasets the status quo for transparent and reproducible AI in radiology.

References

1.Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of TRIPOD-AI and PROBAST-AI. *BMJ Open.* 2021;11:e048008.
2.Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data.* 2016;3:160018.
3.Larson DB, Harvey H, Rubin DL, Irani N, Tse JR, Langlotz CP. Regulatory frameworks for development and evaluation of AI-based imaging algorithms. *J Am Coll Radiol.* 2021;18(3 Pt A):413–424.
4.Lekadir K, Osuala R, Gallin C, et al. FUTURE-AI: Guiding principles for trustworthy AI in medical imaging. *Eur Radiol.* 2022;32:1996–2009.
5.U.S. Food and Drug Administration. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device Action Plan. FDA; 2021.