Radiology: Artificial Intelligence

SPECIAL REPORT

Sharing Data Is Essential for the Future of AI in Medical Imaging

Laura C. Bell, PhD • *Efrat Shimron, PhD*

From the Clinical Imaging Group, Genentech, 1 DNA Way, South San Francisco, CA 94080 (L.C.B.); and Department of Electrical and Computer Engineering and Department of Biomedical Engineering, Technion-Israel Institute of Technology, Haifa, Israel (E.S.). Received August 18, 2023; revision requested September 11; revision received November 16; accepted November 20. **Address correspondence to** L.C.B. (email: *bell.laura@gene.com*).

Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence **2024; 6(1):e230337** • **https://doi.org/10.1148/ryai.230337** • **Content code: AI** •

In January, the U.S. federal government declared 2023 to be a "Year of Open Science" and set new requirements and guidelines from the National Institutes of Health on data sharing (1). Although we have a few exemplary open datasets in our radiology artificial intelligence (AI) community, we can and must do better as physicians and scientists for the benefit of our patients. Sharing data, increasingly important for open science, promotes nonmaleficence (a value of the Hippocratic oath) by reducing duplicate imaging, using data more efficiently, and ensuring clear communication to disseminate research and advance medical care. For example, a dataset shared online by one research group as part of a specific study can spark ideas in other researchers and stimulate new research directions. By embracing open science, we have the opportunity to stimulate innovative research while being time-conscious (think of the time it takes to collect and curate data and pretrain AI models!) during a time of reduced public funding; to recover public trust in science and medicine by improving reproducibility (2); and to support diverse and inclusive research environments.

Public sharing of imaging datasets is a crucial element of open science, which most notably has been associated with open-source code. Open science broadly supports the transparent and collaborative sharing of code, data, methodology, and education—tenets that allow for reproducibility within our research. The concept of open science is not new; scientists in the 17th century recognized the need to communicate results efficiently, and journals and conferences were conceived. However, in modern research where science and information technology meet, the tools for how we conduct research and communicate it have changed drastically. Consider, for instance, the recent COVID-19 pandemic, during which the AI community had a chance to demonstrate its value to the public by building predictive models based on information contained within chest radiographs and CT images. By the conclusion of the pandemic's initial year, it was reported that out of the hundreds of AI models published, none were deemed suitable for clinical translation (3). From these initial publications and reports, data sharing was applauded, and large databases were built (4). Nevertheless, it was clear that we still had to learn how to share quality datasets and ensure close collaboration with radiologists and imaging scientists (5,6). If we want AI to succeed in radiology, we must share data and learn *how* to share data.

Sharing medical imaging datasets is an ongoing global effort; numerous datasets and repositories have been around for the last decade or so. For example, a major resource is The Cancer Imaging Archive, an expansive repository housing cancer-related medical imaging datasets with corresponding clinical data that aims at supporting the reproducibility of quantitative imaging metrics (7,8). These various datasets have allowed researchers to build and share postprocessing algorithms for various cancer applications. Another important resource is the UK Biobank Imaging Study (9), which provides an extensive collection of imaging data, paired with comprehensive clinical and genetic information. Furthermore, initiatives such as the Alzheimer's Disease Neuroimaging Initiative (10), Open Access Series of Imaging Studies (11), and the Human Connectome Project (12) offer vast datasets that contribute substantially to neuroscience research and foster a deeper understanding of brain function. Stanford Medicine's Artificial Intelligence in Medicine and Imaging datasets platform (13) offers a diverse range of medical imaging datasets. In MRI, datasets such as fastMRI (14), mridata.org (15), SKM-TEA (16), M4Raw (17), and others (18) offer raw, nonreconstructed MRI data (k-space measurements), hence enabling development of advanced MRI reconstruction techniques aimed at reducing scan times. Other databases are more focused on certain diseases. The Multimodal Brain Tumor Image Segmentation Benchmark (BraTS) dataset, for example, which is published and continuously updated as part of the BraTS challenges, is dedicated to automated segmentation of brain glioma (19–21). Additionally, the Open Source Imaging Consortium dataset focuses on pulmonary fibrosis (22), and the Mammographic Image Analysis Society dataset (23) and Digital Database for Screening Mammography (24) offer mammographic images for breast cancer detection.

The recent rise of large language models (LLMs) has drawn interest to datasets that contain medical text. LLMs are advanced AI systems capable of processing and generating humanlike text, which can be immensely valuable in radiology (eg, in assisting radiologists in generating reports). Recently developed models include RadBERT (25) and Radiology-GPT (26); additionally, databases such as MIMIC-CXR (27) that contain chest radiographs paired with free-text radiology reports hold much promise for development of further LLMs. In summary, publicly available databases are leveraged for research and development, such as through community challenges; these databases promote the development of reproducible code for clinical translation (eg, segmentation tools and diagnostic models

Abbreviations

AI = artificial intelligence, LLM = large language model

Summary

If we want artificial intelligence to succeed in radiology, we must share data and learn how to share data.

Key Points

- Data are a key component of AI model development. To succeed, we must learn to share data.
- Sharing datasets is increasingly being required by institutions that provide publicly funded research. Let's do it right.
- Open science is science, and it's in all of our best interest for accelerating patient care to embrace it.

Keywords

Open Science, Data Sharing

for disease prediction). Nevertheless, open-access datasets are still needed for different applications (eg, dynamic MRI), and in many cases, articles are published without sharing data and code.

The common reluctance to share datasets at the time of publication is associated with the substantial challenges related to curation and publication of open-access databases. These challenges come in different forms. First is the issue of patient privacy, which is closely related to a patient's well-being and confidentiality—core principles of ethical practice in medicine. Protecting privacy requires applying data anonymization steps, including removal of patient-specific metadata often stored in Digital Imaging and Communications in Medicine images and skull-stripping from three-dimensional brain scans where face recognition may be a concern. Furthermore, the risk of identification necessitates comprehensive informed consent from individuals, which further complicates database curation. Moreover, with the intention of patient privacy, certain legislative policies, such as the General Data Protection Regulation in Europe and the California Consumer Privacy Act in the United States, may potentially complicate data sharing (28). Second, acquiring sufficient data for training models poses a formidable challenge, particularly in modalities such as MRI, where costs can be prohibitively high, and CT, where radiation exposure is a concern. Third, the process of data annotation, a crucial step for training AI models, necessitates expertise in radiology. However, imaging experts have limited time to generate reference standard annotations and their time can be very expensive. An additional barrier stems from the lack of standardized data formats when it comes to raw MRI data (specifically k-space), with disparate vendors like GE, Siemens, and Philips adopting their own data saving conventions. Such variation not only complicates data sharing but also poses challenges for seamless integration and analysis. For instance, the use of various formats like Digital Imaging and Communications in Medicine, NIfTI, and raw MRI data hinders cross-vendor compatibility and requires extensive preprocessing, consuming valuable time and resources in the medical imaging workflow. Last (and anecdotally), we acknowledge that institutional policies may create barriers for various reasons when it comes to sharing datasets. A potential solution is agreeing on the appropriate license that clearly outlines rights and permissions associated when publishing a dataset. We encourage researchers to advocate for change within their institutions as we have done from personal experience.

Due to the many challenges in data curation, researchers often face scarcity or complete lack of data suitable for training AI models, which causes additional challenges and sometimes methodologic failures. For example, a database curated for one purpose might be used for a different purpose. A dataset published for brain tumor segmentation, for example, could be used mistakenly to develop MRI reconstruction algorithms. A recent study demonstrated that such "off-label" data use could lead to biased results and overly optimistic algorithm evaluation (29) due to subtle, task-specific preprocessing steps that are applied to different datasets. Although preprocessing often cannot be detected by the human eye, it can alter the performance of the trained AI algorithms and lead to biased results. Moreover, such workflows could lead to algorithmic failure in clinical settings, as algorithms developed using preprocessed data have very limited applicability to the clinical real world and might miss clinically important details (29).

Additionally, a lack of data may also lead to the improper construction of training and test datasets which can create a different source of bias. For example, AI algorithms trained on data from healthy individuals or those who have a limited range of conditions may not generalize well to a test dataset that includes a larger range of conditions, as typically seen in a hospital (6). Furthermore, studies have shown that AI algorithms are sensitive to sex and racial bias and could perform poorly in underserved populations, such as Black individuals or Hispanic women (30,31). Another source of bias is highly specific data; AI methods trained on data acquired using a medical imaging device from a single vendor or anatomic region might not generalize to data of other vendors or organ systems (32). Finally, another problem that arises from lack of training data is the shift of research efforts toward research problems where data are abundant. For example, the scarcity of open-access datasets in abdominal medical imaging poses a major obstacle to the advancement of algorithms for diagnosing inflammatory bowel disease and hinders the development of AI techniques for accelerated, motion-robust dynamic body imaging (33).

These substantial challenges necessitate the development of novel approaches to facilitate data sharing and open science. At present, the medical imaging community is dedicating efforts along several lines. First, centralized platforms have been developed that enable online storage of open-access data. A notable example is the Medical Imaging and Data Resource Center (MIDRC), established by the National Institutes of Health; MIDRC serves as a centralized platform for COVID-19–related medical imaging data, facilitating in-depth analysis of virus impact. Additionally, platforms like Zenodo, a well-known open-access repository, also contribute to the accessibility and sharing of medical imaging datasets, fostering collaboration and research endeavors across the field. Zenodo assigns a unique digital object identifier to each repository. This ensures a persistent and citable link, enhancing the discoverability and long-term accessibility of research datasets, and facilitates proper attribution in scholarly publications. Second, community-driven efforts focus

on establishing standard data formats. For example, the International Society of Magnetic Resonance in Medicine (ISMRM) community has established a unified data format, known as ISMRM Raw Data format, which enables storing raw MRI data acquired with scanners of different vendors and facilitates easy sharing of data and code (34). Additionally, scientists from University Medical Center Utrecht recently published a blueprint describing a vendor-neutral AI platform that they developed and implemented that aims to overcome some of the barriers related to clinical deployment of AI methods (35).

Substantial efforts are also dedicated to promoting a culture of open code. GitHub, a platform for version control and collaborative software development, has emerged as a leading venue for this aim. Community-based toolboxes that have been shared on Github, such as BART (36), Gadgetron (37), and PulseSeq (38), have garnered extensive and widespread use in the MRI field. Additionally, initiatives such as Papers with Code *(https://paperswithcode.com/)* have also become popular for code sharing. Furthermore, competitions have emerged as major catalysts for promoting reproducible research and open code practices across various domains. For instance, challenges such as BraTS (19–21), which is organized as part of the Medical Image Computing and Computer-Assisted Intervention conference, and Kaggle competitions have spurred collaborative efforts in medical imaging tasks. Moreover, the fastMRI challenge (39) has pushed the boundaries of MRI reconstruction algorithms, encouraging the development of innovative solutions for faster and more accurate medical image reconstruction. Altogether, these platforms and initiatives promote transparency, collaboration, and accessibility, enabling researchers to share their work, facilitate peer review, and enhance the credibility and replicability of scientific findings.

In addition to the aforementioned initiatives, journals have a pivotal role in endorsing and advancing open science. This journal, *Radiology: Artificial Intelligence*, provides an opportunity for authors to submit a Data Resources manuscript. This manuscript category allows authors to publish a summary of a publicly available dataset intended specifically for AI research, along with appropriate links. The publication of a Data Resources manuscript allows for a centralized location to clearly describe the dataset and provide all potential corresponding links, such as the publicly available dataset link, GitHub code repository, and/or previous publications, that can easily be cited by the authors and potential data users. Furthermore, this journal's Data Resources Collection page *(https://pubs.rsna.org/page/ai/data_resources)* provides a full list of all published Data Resources manuscripts to date and will be continually updated as new such manuscripts are published, further facilitating access and reference to these publicly available datasets. We understand that the perceived notion that there is more to lose when sharing data, but opportunities to submit publications such as Data Resources manuscripts offer a way to give credit to the researchers and the patients—and we believe this outweighs the concerns.

We understand the hesitancy to share data and code openly given the personal apprehension of criticism. However, from our experience, we found that sharing data and code has many benefits, such as facilitating collaborations and follow-up studies. We encourage you to start small. Choose one goal to focus on, whether it is actively participating in an open science initiative within your society (40), committing to publishing open code, reading guidelines for how to prepare medical imaging data for publication (41), releasing an open dataset from a previously published article, or allocating funds in your next grant to support open-science practices. Science has always been an iterative process, and in this dynamic field of AI for medical imaging, we must learn to embrace open science to accelerate the translation of our tools into clinical practice.

Author contributions: Guarantors of integrity of entire study, **L.C.B., E.S.**; study concepts/study design or data acquisition or data analysis/interpretation, **L.C.B., E.S.**; manuscript drafting or manuscript revision for important intellectual content, **L.C.B., E.S.**; approval of final version of submitted manuscript, **L.C.B., E.S.**; agrees to ensure any questions related to the work are appropriately resolved, **L.C.B., E.S.**; literature research, **L.C.B., E.S.**; and manuscript editing, **L.C.B.**

Disclosures of conflicts of interest: L.C.B. Employee of Genentech and shareholder of F. Hoffmann La Roche; employer pays for author to attend conferences yearly (eg, ISMRM, ATS this year in 2023); secretary of the ISMRM Reproducibility and Research Study Group (unpaid). **E.S.** Support from UC Berkeley, Technion -Israel Institute of Technology, and Weizmann Institute of Science.

References

1. The White House. FACT SHEET: Biden-Harris Administration Announces New Actions to Advance Open and Equitable Research. https://www.whitehouse.gov/ostp/news-updates/2023/01/11/fact-sheet-biden-harris-administration-announces-new-actions-to-advance-open-and-equitable-research/. Published January 11, 2023. Accessed August 17, 2023.
2. Stupple A, Singerman D, Celi LA. The reproducibility crisis in the age of digital medicine. NPJ Digit Med 2019;2:2. [Published correction appears in NPJ Digit Med 2019;2:19.]
3. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 2020;369:m1328 [Published correction appears in BMJ 2020;369:m2204. [Published update appears in BMJ 2021;372:n236.]
4. National Library of Medicine. NIH-Supported Data Sharing Resources: Domain-Specific Repositories. https://www.nlm.nih.gov/NIHbmic/domain_specific_repositories.html. Updated April 20, 2023. Accessed November 14, 2023.
5. Sperrin M, Grant SW, Peek N. Prediction models for diagnosis and prognosis in Covid-19. BMJ 2020;369:m1464.
6. Varoquaux G, Cheplygina V. Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ Digit Med 2022;5(1):48.
7. Levy MA, Freymann JB, Kirby JS, et al. Informatics methods to enable sharing of quantitative imaging research data. Magn Reson Imaging 2012;30(9):1249–1256.
8. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. J Digit Imaging 2013;26(6):1045–1057.
9. Ollier W, Sprosen T, Peakman T. UK Biobank: from concept to reality. Pharmacogenomics 2005;6(6):639–646.
10. Jack CR Jr, Bernstein MA, Fox NC, et al. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. J Magn Reson Imaging 2008;27(4):685–691.
11. Kurdi B, Lozano S, Banaji MR. Introducing the open affective standardized image set (OASIS). Behav Res Methods 2017;49(2):457–470.
12. Van Essen DC, Smith SM, Barch DM, Behrens TE, Yacoub E, Ugurbil K; WU-Minn HCP Consortium. The WU-Minn human connectome project: an overview. Neuroimage 2013;80:62–79.
13. Stanford University Center for Artificial Intelligence in Medicine & Imaging. Shared Datasets. https://aimi.stanford.edu/shared-datasets. Accessed November 14, 2023.
14. Knoll F, Zbontar J, Sriram A, et al. fastMRI: A Publicly Available Raw k-Space and DICOM Dataset of Knee Images for Accelerated MR Image Reconstruction Using Machine Learning. Radiol Artif Intell 2020;2(1):e190007.
15. Ong F, Amin S, Vasanawala S, Lustig M. Mridata.org: An open archive for sharing MRI raw data. ISMRM-ESMRMB June 2018. https://archive.ismrm.org/2018/3425.html.

16. Desai AD, Schmidt AM, Rubin EB, et al. SKM-TEA: A Dataset for Accelerated MRI Reconstruction with Dense Image Labels for Quantitative Clinical Evaluation. arXiv 2203.06823 [preprint] https://arxiv.org/abs/2203.06823. Published March 14, 2022. Accessed August 2023.
17. Lyu M, Mei L, Huang S, et al. M4Raw: A multi-contrast, multi-repetition, multi-channel MRI k-space dataset for low-field MRI research. Sci Data 2023;10(1):264.
18. Lim Y, Toutios A, Bliesener Y, et al. A multispeaker dataset of raw and reconstructed speech production real-time MRI video and 3D volumetric images. Sci Data 2021;8(1):187.
19. Menze BH, Jakab A, Bauer S, et al. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging 2015;34(10):1993–2024.
20. Ghaffari M, Sowmya A, Oliver R. Automated Brain Tumor Segmentation Using Multimodal Brain Scans: A Survey Based on Models Submitted to the BraTS 2012-2018 Challenges. IEEE Rev Biomed Eng 2020;13:156–168.
21. Adewole M, Rudie JD, Gbadamosi A, et al. The Brain Tumor Segmentation (BraTS) Challenge 2023: Glioma Segmentation in Sub-Saharan Africa Patient Population (BraTS-Africa). arXiv 2305.19369 [preprint] https://arxiv.org/abs/2305.19369. Published May 30, 2023. Accessed November 14, 2023.
22. OSICild. Open Source Imaging Consortium Web site. https://www.osicild.org/. Accessed November 14, 2023.
23. Yoon WB, Oh JE, Chae EY, et al. Automatic Detection of Pectoral Muscle Region for Computer-Aided Diagnosis Using MIAS Mammograms. Biomed Res Int 2016;2016:5967580.
24. Heath M, Bowyer K, Kopans D, et al. Current Status of the Digital Database for Screening Mammography. In: Karssemeijer N, Thijssen M, Hendriks J, van Erning L, eds. Digital Mammography. Computational Imaging and Vision, vol 13. Springer, 1998; 457–460.
25. Yan A, McAuley J, Lu X, et al. RadBERT: Adapting Transformer-based Language Models to Radiology. Radiol Artif Intell 2022;4(4):e210258.
26. Liu Z, Zhong A, Li Y, et al. Radiology-GPT: A Large Language Model for Radiology. arXiv 2306.08666 [preprint] https://arxiv.org/abs/2306.08666. Published June 14, 2023. Accessed November 14, 2023.
27. Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 2019;6(1):317.
28. de Kok JWTM, de la Hoz MÁA, de Jong Y, et al. A guide to sharing open healthcare data under the General Data Protection Regulation. Sci Data 2023;10(1):404.
29. Shimron E, Tamir JI, Wang K, Lustig M. Implicit data crimes: Machine learning bias arising from misuse of public data. Proc Natl Acad Sci U S A 2022;119(13):e2117203119.
30. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 2021;27(12):2176–2182.
31. Chen RJ, Wang JJ, Williamson DFK, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. Nat Biomed Eng 2023;7(6):719–742.
32. Darestani MZ, Chaudhari AS, Heckel R. Measuring Robustness in Deep Learning Based Compressive Sensing. arXiv 2102.06103 [preprint] https://arxiv.org/abs/2102.06103. Published February 11, 2021. Accessed November 14, 2023.
33. Spieker V, Eichhorn H, Hammernik K, et al. Deep Learning for Retrospective Motion Correction in MRI: A Comprehensive Review. IEEE Trans Med Imaging 2023. 10.1109/TMI.2023.3323215. Published online October 13, 2023.
34. Inati SJ, Naegele JD, Zwart NR, et al. ISMRM Raw data format: A proposed standard for MRI raw datasets. Magn Reson Med 2017;77(1):411–421.
35. Leiner T, Bennink E, Mol CP, Kuijf HJ, Veldhuis WB. Bringing AI to the clinic: blueprint for a vendor-neutral AI deployment infrastructure. Insights Imaging 2021;12(1):11.
36. Uecker M, Ong F, Tamir JI, et al. Berkeley Advanced Reconstruction Toolbox. ISMRM 2015. https://archive.ismrm.org/2015/2486.html.
37. Hansen MS, Sørensen TS. Gadgetron: an open source framework for medical image reconstruction. Magn Reson Med 2013;69(6):1768–1776.
38. Layton KJ, Kroboth S, Jia F, et al. Pulseq: A rapid and hardware-independent pulse sequence prototyping framework. Magn Reson Med 2017;77(4):1544–1552.
39. Knoll F, Murrell T, Sriram A, et al. Advancing machine learning for MR image reconstruction with an open competition: Overview of the 2019 fastMRI challenge. Magn Reson Med 2020;84(6):3054–3070.
40. Bell LC, Suzuki Y, van Houdt PJ, Sourbron S, Mutsaerts HJMM. The road to the ISMRM OSIPI: A community-led initiative for reproducible perfusion MRI. Magn Reson Med 2023. 10.1002/mrm.29736. Published online June 6, 2023.
41. Willemink MJ, Koszek WA, Hardell C, et al. Preparing Medical Imaging Data for Machine Learning. Radiology 2020;295(1):4–15.