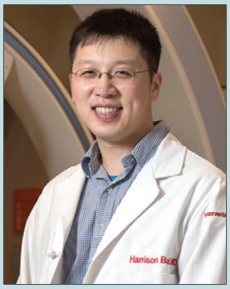
Radiology

RICORD: A Precedent for Open AI in COVID-19 Image Analytics

Harrison X. Bai, MD • Nicole M. Thomasian, BS

Dr Harrison Bai is an assistant professor of diagnostic imaging at the Warren Alpert Medical School of Brown University in Providence, Rhode Island. His research interests focus on artificial intelligence, machine learning, and computer vision as applied to medical image analysis. Dr Bai is an associate editor for the journal *Radiology: Artificial Intelligence* and is currently a principal investigator for an RSNA Research Scholar grant and an NIH grant.



Nicole Thomasian, BS, is a dual medical degree and master of public policy (MD/MPP) candidate with Brown University and the Harvard Kennedy School. Her research interests focus on optimizing technology at the systems interface to promote health security. Nicole is a prior Fulbright fellow and currently serves as a student fellow at the Belfer Center for Science and International Affairs.

The coronavirus disease 2019 (COVID-19) pandemic continues to surge across the globe, recently eclipsing prior peak records in daily cases (1). The finding that a novel pathogen known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) causes COVID-19 was the result of an open science endeavor to provide rapid sequencing of the viral genome (2). Data sharing initiatives to foster accelerated COVID-19 research have since followed, running the gamut from open access academic journal resource hubs and bioinformatics consortiums to open partnerships in therapeutics discovery (3). In recognizing the role of open data curation in supporting the global pandemic response, the RSNA, with partners at four international sourcing institutions and the National Institutes of Health (NIH), have developed the RSNA International COVID-19 Open Annotated Radiology Database (RICORD), presented by Tsai and Simpson et al in this issue of *Radiology* (4). The RICORD data set is the first public, expertly annotated data set for COVID-19 thoracic imaging that encompasses both multinational and multimodal data.

The artificial intelligence (AI) community can leverage the robustness of the RICORD imaging data to accelerate advances in clinical diagnostics, prognostics, and management of SARS-CoV-2. Characterization of SARS-CoV-2 on chest images is ongoing, but evidence to date suggests that the syndrome possesses the radiographic qualities of an organizing pneumonia. Stereotyped imaging features on chest CT scans include ground-glass opacities in a peripheral distribution, often rounded with bi- or multilobar involvement (5,6). The field has already moved beyond early efforts in COVID-19 computer vision techniques that centered around obtaining an early diagnosis and indexing the severity of SARS-CoV-2 disease. At this stage, the primary utility for AI-based COVID-19 chest imaging applications is forecasting disease and monitoring therapy in correlation with clinical data (7).

RICORD is a groundbreaking undertaking in the promotion of quality and accessible imaging data. First, a lack of heterogeneous and high-volume data complicated early efforts to use machine learning to characterize SARS-CoV-2 on chest images. The diversity in the RICORD images answers this need for better data generalizability with its 240 chest CT and 1000 chest radiograph images across four international sites. RICORD also circumvents another critical barrier to the advancement of AI-driven bioinformatics research: a paucity of large, labeled data sets with open data use privileges. In particular, the lack of access to large volume data undermines the development of deep learning applications, but the recent resurgence of this access has enabled key advances in image interpretation. Until workaround data manipulation solutions for overfitting in low-volume settings like pretraining or data augmentation become routine, small cohort size will continue to impede deep learning model use. We suspect RICORD will be helpful to many different stakeholders within the machine learning community, as these stakeholders can tailor its use depending on their individual needs. For clinical radiology groups, RICORD will likely serve as a pristine external validation set to test AI algorithms developed on their respective multi-institutional data set. Other nonaffiliated clinicians or academics, such as computer scientists or engineers without easy access to clinical data repositories, can use RICORD thoracic imaging data for primary algorithm development.

Beyond advancing AI data quality, RICORD also promotes consensus methodologies in image preprocessing and database curation. For example, RICORD features a harmonized annotation schema collated by the RSNA, the Society of Thoracic Radiology, and the European Society

From the Department of Diagnostic Imaging, Rhode Island Hospital, Warren Alpert Medical School of Brown University, 1 Eddy St, Providence, RI 02906 (H.X.B., N.M.T.) and the Harvard Kennedy School of Government, Cambridge, Mass (N.M.T.). Received November 5, 2020; revision requested November 10; revision received and accepted November 11. Address correspondence to H.X.B. (e-mail: *harrison_bai@brown.edu*).

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

See also the article by Tsai and Simpson et al in this issue.

Radiology 2021; 299:E219–E220 • https://doi.org/10.1148/radiol.2020204214 • Content codes: CH CT IN • ©RSNA, 2021

of Medical Imaging Informatics. Future AI endeavors can leverage this new common syntax for machine learning, thereby reducing interinstitutional variability in local pipelines. Further, for uniform image acquisition and de-identification, all thoracic imaging data in RICORD is in the Digital Imaging and Communications in Medicine format, which is the international standard. Taken together, these consensus methodologies enhance data interoperability to increase the total pool of available data for easy extraction by the research community.

The rapid deployment of safe and effective machine learning solutions is necessary to keep pace with the ever-shifting clinical and therapeutic needs of the evolving COVID-19 pandemic. One of the most time-consuming and resource-intensive aspects of machine learning algorithm development is image data preprocessing. Open data curation can ease some of the burden of preprocessing techniques by preventing the duplication of efforts within the machine learning community as it relates to data cleaning. RICORD is again distinguished here by its commitment to the utmost rigor in annotation practices. CT scans were annotated by six thoracic subspecialist radiologists, and chest radiographs were triple-annotated, with final adjudication by an experienced thoracic subspecialist (average 15 years of experience) in cases without a majority consensus. In this way, with its ready-to-use data that can be effortlessly siphoned into AI development pipelines, RICORD slashes the workload for researchers without sacrificing quality.

Another strategy for optimizing open machine learning data curation is through cloud-based infrastructures that can provide ease of scaling and intuitive coupling to analytic pipelines. Poor data interoperability associated with conventional servers can manifest as poor image elasticity and compatibility. Migration to the cloud can circumvent this barrier to unlock earlier siloed data for use by the machine learning community. With the launch of the NIH's Imaging Data Commons in October 2020, efforts to converge The Cancer Imaging Archive repository with the cloud are already underway (8). The prospect of featuring RICORD on the NIH's emerging cloud-based Imaging Data Commons infrastructure would further improve the data archiving, exchange, viewing, latency, and distribution experience for developers.

Moving forward, we expect that RICORD will enable powerful advances in computer vision applications for COVID-19 through links to clinical metrics, such as laboratory and outcome data. Machine learning algorithms can leverage the inclusion of longitudinal follow-up data to forecast SARS-CoV-2 disease progression and to support clinical trial monitoring of therapeutic candidates for COVID-19 (9). The addition of more detailed clinical data to RICORD is coming soon in the form of the Medical Imaging and Data Resource Center, which forms a larger RSNA COVID-19 collaboration with partners at the American College of Radiology and the American Association of Physicists in Medicine. This next iteration would also be further strengthened by the addition of more data about the clinical distributions and characteristics of the SARS-CoV-2–negative cohort.

RICORD lays the groundwork for future AI data sharing initiatives via a delineation of consensus methodologies in a superb public data set curation. The RICORD imaging initiative also fosters an ethos of collaboration and transparency in medicine that highlights the importance of open bioinformatics as a path to ethical AI. We envision that RICORD will not only power the development of machine learning algorithms in the context of COVID-19 but will also act as a future catalyst for the rapid deployment of AI solutions to meet future global health needs.

Disclosures of Conflicts of Interest: **H.X.B.** disclosed no relevant relationships. **N.M.T.** disclosed no relevant relationships.

References

1. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020;20(5):533–534 https://doi.org/10.1016/S1473-3099(20)30120-1.
2. Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. Nature 2020;579(7798):265–269 [Published correction appears in Nature 2020;580(7803):E7.].
3. COVID-19 Resources. Radiological Society of North America. https://www.rsna.org/covid-19. Updated October 13, 2020. Accessed November 4, 2020.
4. Tsai EB, Simpson S, Lungren MP, et al. The RSNA International COVID-19 Open Radiology Database (RICORD). Radiology 2021;299(1):E204–E213.
5. Bai HX, Hsieh B, Xiong Z, et al. Performance of Radiologists in Differentiating COVID-19 from Non-COVID-19 Viral Pneumonia at Chest CT. Radiology 2020;296(2):E46–E54.
6. Chung M, Bernheim A, Mei X, et al. CT Imaging Features of 2019 Novel Coronavirus (2019-nCoV). Radiology 2020;295(1):202–207.
7. Liu F, Zhang Q, Huang C, et al. CT quantification of pneumonia lesions in early days predicts progression to severe illness in a cohort of COVID-19 patients. Theranostics 2020;10(12):5613–5622.
8. Imaging Data Commons. National Institutes of Health. https://datacommons.cancer.gov/repository/imaging-data-commons. Accessed November 4, 2020.
9. Gieraerts C, Dangis A, Janssen L, et al. Prognostic Value and Reproducibility of AI-assisted Analysis of Lung Involvement in COVID-19 on Low-Dose Submillisievert Chest CT: Sample Size Implications for Clinical Trials. Radiol Cardiothorac Imaging 2020;2(5):e200441.