Radiology: Artificial Intelligence

EDITORIAL

Machine Learning for COVID-19 Diagnosis and Prognostication: Lessons for Amplifying the Signal While Reducing the Noise

Derek Driggs, MS • Ian Selby, MD* • Michael Roberts, PhD • Effrossyni Gkrania-Klotsas, PhD • James H. F. Rudd, MD, PhD • Guang Yang, PhD • Judith Babar, MD • Evis Sala, MD, PhD • Carola-Bibiane Schönlieb, PhD • on behalf of the AIX-COVNET collaboration*

From the Department of Applied Mathematics and Theoretical Physics (D.D., M.R., C.B.S.) and Division of Cardiovascular Medicine (J.H.F.R.), University of Cambridge, Cambridge, England; Department of Radiology, School of Clinical Medicine, University of Cambridge and CRUK Cambridge Centre, Cambridge Biomedical Campus, Cambridge CB2 0QQ, England (I.S., J.B., E.S.); Oncology R&D, AstraZeneca, Cambridge, England (M.R.); Department of Infectious Diseases, University of Cambridge Hospitals, Cambridge, England (E.G.K.); and National Heart and Lung Institute, Imperial College London, London, England (G.Y.). Received January 11, 2021; revision requested February 1; revision received March 1; accepted March 3. **Address correspondence to** E.S. (e-mail: *es220@medschl.cam.ac.uk*).

*D.D. and I.S. contributed equally to this work.

Supported by the NIHR Cambridge Biomedical Research Centre, the Engineering and Physical Sciences Research Council (EPSRC), the British Heart Foundation, Cancer Research UK National Cancer Imaging Translational Accelerator (NCITA) (C22479/A28667), Intel, and the DRAGON consortium, which received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement no 101005122, and the Cambridge Mathematics of Information in Healthcare (CMIH) Hub EP/T017961/1. C.B.S. further acknowledges support from the Philip Leverhulme Prize, the Royal Society Wolfson Fellowship, the EPSRC grants EP/S026045/1 and EP/T003553/1, EP/N014588/1, the Wellcome Innovator Award RG98755, European Union Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 777826 Nonlocal Methods for Arbitrary Data Sources (NoMADS), the Cantab Capital Institute for the Mathematics of Information, and the Alan Turing Institute.

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

Radiology: Artificial Intelligence 2021; 3(4):e210011 • https://doi.org/10.1148/ryai.2021210011 • Content codes: AI CH

Since the emergence of COVID-19, researchers in machine learning and radiology have rushed to develop algorithms that could assist with diagnosis, triage, and management of the disease (1). As a result, thousands of diagnostic and prognostic models using chest radiographs and CT have been developed. However, with no standardized approach to development or evaluation, it is difficult, even for experts, to determine which models may be of most clinical benefit. Here, we share our main concerns and present some possible solutions.

Systematic Errors in the Literature

In April 2020, during the first wave of the novel coronavirus outbreak in Europe and the United States, Gog published an editorial outlining how researchers could use their skills to help (2). Her article was a call for researchers to proceed cautiously, stating that the priority should be to "amplify the signal" but avoid "adding to the noise" in the literature. In the several months since this appeal to caution, have we, as a research community, followed her guidance?

Our AIX-COVNET collaboration is a multidisciplinary team of radiologists and other clinicians working alongside image-processing and machine learning specialists to develop artificial intelligence tools to support frontline practitioners in the COVID-19 pandemic (3). We set out to quantify common problems in the enormous number of articles that developed machine learning models for COVID-19 diagnosis and prognostication using thoracic imaging. We systematically reviewed every such study published between January 1 and October 3, 2020, and found two predominant sources of error (4). First, an apparent deterioration in standards of research, and second, a lack of collaboration between the machine learning and medical communities leading to inappropriate and redundant efforts.

To create models quickly, researchers frequently have relaxed standards for developing safe, reliable, and validated algorithms. This laxity is most obvious in the datasets used to train these models. These datasets contain too few examples from patients with COVID-19, their quality is unreliable, and their origins are poorly understood. Many have been developed with access to only a few hundred COVID-19 images, where comparable models before the pandemic were trained using up to half a million examples (5). Few articles address this small-data issue, or the resulting imbalance of class sizes, making it unlikely that their results will generalize to the wider community. For example, because of the prevalence of data from China, many researchers train on small datasets from China when the model is intended for European populations, and recent research suggests such models are ineffective in practice (6). Differences between the training data and the target population, including patient phenotypes and data acquisition procedures, can all affect a model's generalizability (6). Training generalizable models from small amounts of labeled data is a common problem in medical imaging, and techniques such as transfer learning, self- or semisupervised learning, and parameter pruning can ameliorate this issue (7,8).

Although data sharing is critical for the research community to thrive, distributing or using public datasets of poor quality and unknown origins can further damage research efforts. Many public datasets are combinations of images assembled from other public datasets and redistributed under a new name (9,10). This repackaging of data has led to researchers unknowingly validating their models on public datasets that contain their training data as a subset, likely producing an optimistic view of their performance. There are also a surprising number of studies that unknowingly use a public dataset of pediatric patients for

non–COVID-19 cases (9). Additionally, many researchers have not acknowledged that some popular public datasets of patients with COVID-19 are composed of images taken from journal articles with no access to the original DICOM files (11). Whether "pictures of pictures" provide the same quality data as original images is an issue that was discussed before the beginning of this pandemic (12,13) without an established consensus. In this time of crisis, these concerns have been ignored.

Given the prevalence of research quality standards for developing medical models, it is perhaps surprising that such widespread issues exist in the COVID-19 literature. We have determined that disconnects between research standards in the medical and machine learning communities partly explain these issues. For example, the Prediction model Risk Of Bias Assessment Tool (PROBAST) checklist (14) for assessing the risk of bias in medical models requires models to be validated on an external dataset, but in machine learning research, it is common practice to validate a model using an 80:20 training-to-testing split from the same data source. On the other hand, model quality checklists, such as the radiomics quality score (RQS) (15), suggest that to protect against overfitting, a model must train on at least 10 training examples per model parameter. However, deep learning models have been shown to generalize well despite heavy overparameterization (16), so this requirement is often inappropriate for deep learning models. Furthermore, with deep learning models, it is difficult to interpret the extracted features, making it difficult to run standard risk-of-bias assessments from the medical literature (17). These gaps between research standards in medicine and machine learning allow the dissemination of irreproducible research, and they extend far beyond the immediate COVID-19 crisis. Collaboration and communication between these communities to bridge these gaps will be necessary as more machine learning models phase into clinical deployment.

Recommendations for Clinical Model Development during the COVID-19 Pandemic and Beyond

Our collaboration is exemplary as it comprises clinicians, machine learning researchers, mathematicians, and radiologists. Given our own experiences and the findings presented in our discussion of the literature, we propose some guiding principles for developing clinical models in the COVID-19 era and beyond.

Work as a multidisciplinary team.—Many existing studies were performed without any input from clinicians. Because of this, models have been built to solve problems that do not necessarily provide significant clinical benefit (4). For example, in the United Kingdom, chest radiographs have a much more significant role in COVID-19 diagnosis than CT scans, but early models focused mostly on diagnosis from CT (18,19). Adapting to local medical practices is difficult without collaborating with clinicians.

Source original data.—The origins of public datasets are often unknown, so it is difficult to determine their quality or suitability for inclusion in model development. Such datasets are also unlikely to represent a model's target population, making it less likely for a model's performance to generalize upon deployment. Training on high-quality data that are representative of the target community, with validation on data sourced externally, provides the best estimate of a model's performance.

Streamline data acquisition and processing.—Collecting high-quality data is always a challenge in machine learning, particularly data on a novel virus, but preparation can make data collection easier. Researchers must be familiar with local guidance surrounding the use and sharing of patient data, and pre-emptive protocols for obtaining, anonymizing, and securely storing data, including for anticipated future pandemics, are essential. The current crisis has demonstrated that without these pre-emptive protocols, data collection can be severely delayed. Equally important is developing efficient and potentially semi-automated data preprocessing pipelines to ensure rapid access to high-quality, well-curated datasets. Making these procedures publicly accessible also ensures that different groups do not need to spend time curating the same data.

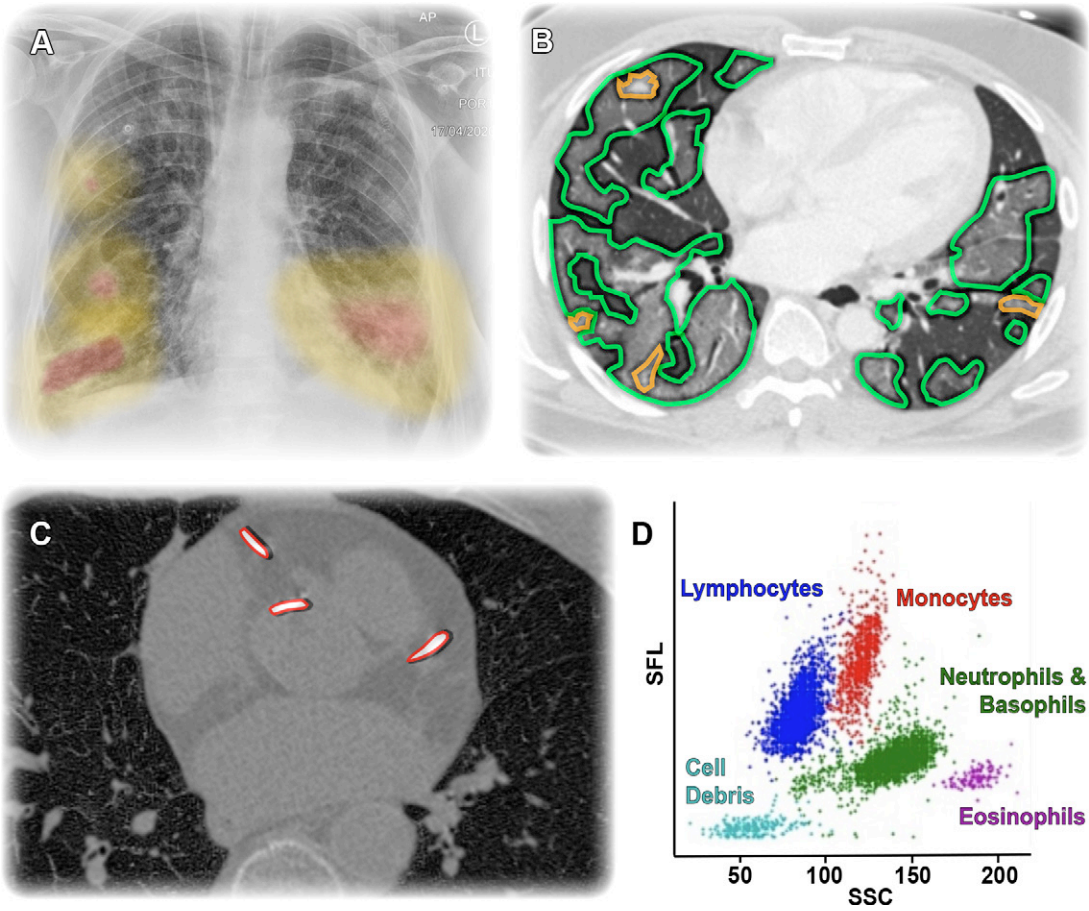
Acknowledge the small-data problem.—Obtaining large amounts of labeled data for medical applications is difficult, especially when they relate to a novel virus. Models should be adjusted to respond to this small-data problem. Although this is an ongoing area of research, several strategies have been shown to boost performance when working with small or sparsely labeled datasets, including semi- and self-supervised learning (7,20), weight transfusion, and limiting the number of trainable parameters (8).

Follow and improve medical standards.—There are gaps between research standards in medicine and machine learning, and more research is required to resolve these inconsistencies. Machine learning researchers should be aware of the RQS (15) and the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) (21), standard checklists to evaluate models using radiomic features. It is also imperative to evaluate a model's risk of bias using standards such as PROBAST (14) and to report results following guidelines such as the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) checklist (22). Conversely, medical standards must be updated to support deep learning practices. Indeed, calls for an updated TRIPOD checklist (TRIPOD-ML) (23) and the related reporting guidelines SPIRIT-AI (24) and CONSORT-AI (25) are steps in this direction.

A Multimodal Approach to the Diagnosis and Prognosis of Patients with COVID-19

With these considerations in mind, there remain plenty of opportunities for machine learning models to aid clinicians during the current pandemic and beyond, with much of the knowledge gained applicable to other diseases including future pandemics. Below, we outline several data sources that could be used to develop models that are helpful to clinicians (Figure).

Chest radiographs.—Chest radiographs are a first-line investigation in many countries, including the United Kingdom. Researchers could examine not only the initial imaging findings



Our collaboration has identified five promising applications of machine learning in the COVID-19 pandemic. The AIX-COVNET collaboration's vision for a multistream model incorporates multiple imaging segmentation methods (**A**, **B**, and **C**) with flow cytometry (**D**) and clinical data. **(A)** A saliency map on a radiograph from the NCCID dataset (26). **(B)** Segmented parenchymal disease on a CT scan from the National COVID-19 Chest Imaging Database (NCCID) (26). **(C)** Segmentation of calcified atherosclerotic disease on an image from the NCCID (26). **(D)** A projection of a flow cytometry scatterplot of side-scattered light (SSC) versus side-fluorescence light (SFL), giving insight into cell structures (analysis performed on a Sysmex UK [30] flow cytometer).

and extent of respiratory involvement, but also how radiographic progression in serial studies correlates with patients' clinical phenotypes. Many works have developed deep learning models using chest radiographs of patients with COVID-19, but further research is required to determine if similar models could be clinically viable, especially for prognostic models.

Thoracic CT.—Another promising area of research that has received some attention is in developing segmentation and classification methods to locate lung parenchyma that could be affected by COVID-19 and classify these regions as a symptom of COVID-19 or a result of another disease. High-quality datasets for chest radiographs and CT include the British National COVID-19 Chest Imaging Database (NCCID) (26) and the Medical Imaging and Data Resource Center (MIDRC-RICORD) datasets curated by the RSNA (27).

Comorbidities.—Given that patients with cardiovascular comorbidities are at higher risk of severe disease and mortality (28), it is natural to consider the cardiovascular information that is also contained in thoracic CT. Models that incorporate automated calcium scoring, for example, allow for the burden of atherosclerotic disease to be incorporated into prognostic models, even in those patients with no prior cardiovascular diagnosis. The effects of COVID-19 on the heart have received little attention.

Flow cytometry.—Many diseases cause irregularities in the physical and chemical properties of blood cells, affecting distinct cell types differently. COVID-19 might cause a specific and unique set of changes that can be rapidly detected with flow cytometry.

Electronic health records.—This often-untapped plethora of granular and longitudinal data has recently shown promising results when used in models for COVID-19 prognostication (29). Multiple centers collect data in different formats, consider different features, and store data in potentially many different systems. One significant challenge is to design an algorithm robust to these factors.

Ideally, a model would use more than one data source. An especially promising direction for investigation is how to optimally combine clinical and radiomic features (4).

Many clinicians welcome helpful and appropriately validated models into the clinic. By making these projects open source, interested hospitals can integrate these models into their clinical workflow.

The Impact of the Pandemic on the Future of Artificial Intelligence Development in Radiology

The COVID-19 pandemic presents an opportunity to accelerate cooperation between image scientists, data scientists, radiologists, and other clinicians; our collaboration is but one example. Researchers are close to realizing the potential of machine learning in health care, but there are still many barriers to deployment. To overcome many of these, we do not necessarily need more powerful machine learning models, but a better understanding of how to develop these tools responsibly. Bridging disconnects between machine learning and medical communities is an important step forward, and the current pandemic will forge vital collaborations with potential benefits beyond COVID-19.

Disclosures of Conflicts of Interest: D.D. disclosed no relevant relationships. **I.S.** Activities related to the present article: institution received grant from Innovative Medicines Initiative (Innovative Medicines Initiative grant funding was paid to the University of Cambridge as part of the DRAGON consortium. Author's salary is paid using a portion of this funding, but author can confirm that no specific or additional payment was made relating to this article and the source of funding had no influence on the content of this opinion piece. (The DRAGON consortium is a group of high-tech SMEs, academic research institutes, biotech and pharma partners, affiliated patient-centered organizations and professional societies aiming to apply artificial intelligence for improved and more rapid diagnosis and prognosis in COVID-19.) Further details may be found at *https://www.imi.europa.eu/projects-results/project-factsheets/dragon*. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. **M.R.** disclosed no relevant relationships. **E.G.K.** disclosed no relevant relationships. **J.H.F.R.** disclosed no relevant relationships. **G.Y.** Activities related to the present article: institution received grant from ERC H2020 and ERC IMI (European Research Council Innovative Medicines Initiative on Development of Therapeutics and Diagnostics Combatting Coronavirus Infections Award 'DRAGON: rapiD and secuRe AI imaging based diaGnosis, stratification, fOllowup, and preparedness for coronavirus paNdemics' [H2020- JTI-IMI2 101005122]) (AI for Health Imaging Award 'CHAIMELEON: Accelerating the Lab to Market Transition of AI Tools for Cancer Management' [H2020-SC1-FADTS-2019-1952172]). Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. **J.B.** disclosed no relevant relationships. **E.S.** Activities related to the present article: serves as senior consulting editor for *Radiology: Artificial Intelligence*. Activities not related to the present article: consultant for Amazon; payment for lectures from Siemens and GSK; stock/stock options in Co/counter Lucida Medical. Other relationships: disclosed no relevant relationships. **C.B.S.** Activities related to the present article: institution received grant form European Union funding and RCUK funding. Activities not related to the present article: employed by University of Cambridge. Other relationships: disclosed no relevant relationships.

References

1. Kundu S, Elhalawani H, Gichoya JW, Kahn CE Jr. How might AI and chest imaging help unravel COVID-19's mysteries? Radiol Artif Intell 2020;2(3):e200053.
2. Gog JR. How you can help with COVID-19 modelling. Nat Rev Phys 2020;2(6):274–275.
3. AIX-COVNET Collaboration Website. University of Cambridge. https://covid19ai.maths.cam.ac.uk. Accessed December 2020.
4. Summers RM. Artificial Intelligence of COVID-19 Imaging: A Hammer in Search of a Nail. Radiology 2021;298(3):E162–E164.
5. Irvin J, Rajpurkar P, Ko M, et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. Proc AAAI Conf Artif Intell 2019;33(01):590–597.
6. Goncalves J, Yan L, Zhang HT, et al. Li Yan et al. reply. Nat Mach Intell 2021;3(1):28–32.
7. He K, Fan H, Wu Y, et al. Momentum Contrast for Unsupervised Visual Representation Learning. ArXiv 1911.05722. [preprint] https://arxiv.org/abs/1911.05722. Posted November 13, 2019. Accessed December 2020.
8. Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: Understanding transfer learning for medical imaging. In NeurIPS. 2019; 3342–3352. https://papers.nips.cc/paper/2019/hash/eb1e78328c46506b46a4ac4a1e378b91-Abstract.html.
9. Kermany DS, Goldbaum M, Cai W, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell 2018;172(5):1122–1131.e9.
10. Shih G, Wu CC, Halabi SS, et al. Augmenting the National Institutes of Health chest radiograph dataset with expert annotations of possible pneumonia. Radiol Artif Intell 2019;1(1):e180041.
11. Zhao J, Zhang Y, He X, Xie P. COVID-CT-Dataset: A CT Scan Dataset about COVID-19. ArXiv 2003.13865 [preprint] https://arxiv.org/abs/2003.13865. Posted March 30, 2020. Accessed December 2020.
12. Kurakin A, Goodfellow IJ, Bengio S. Adversarial examples in the physical world. 5th Int Conf Learn Represent ICLR 2017 - Work Track Proc. ArXiv [preprint] https://arxiv.org/abs/1607.02533. Posted July 8, 2016. Updated February 11, 2017. Accessed December 20, 2020.
13. Rajpurkar P, Chen P, Kiani A, et al. CheXpedition: Investigating Generalization Challenges for Translation of Chest X-Ray Algorithms to the Clinical Setting. ArXiv [preprint] https://arxiv.org/abs/2002.11379. Posted February 26, 2020. Updated March 11, 2020. Accessed December 20, 2020.
14. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med 2019;170(1):51–58.
15. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 2017;14(12):749–762.
16. Zhang C, Recht B, Bengio S, Hardt M, Vinyals O. Understanding deep learning requires rethinking generalization. 5th Int Conf Learn Represent ICLR 2017 - Conf Track Proc. ArXiv [preprint] https://arxiv.org/abs/1611.03530. Posted November 10, 2016. Updated February 26, 2017. Accessed December 21, 2020.
17. Roberts M, Driggs D, Thorpe M, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell. ArXiv [preprint] https://arxiv.org/abs/2008.06388. Posted August 14, 2020. Accessed December 2020.
18. Royal College of Radiologists. The role of CT in patients suspected with COVID-19 infection. https://www.rcr.ac.uk/college/coronavirus-covid-19-what-rcr-doing/clinical-information/role-ct-chest/role-ct-patients. Published March 12, 2020. Accessed December 20, 2020.
19. British Society of Thoracic Imaging. Radiology decision tool for suspected COVID-19. https://www.bsti.org.uk/standards-clinical-guidelines/clinical-guidelines/bsti-nhse-covid-19-radiology-decision-support-tool. Published 2020. Accessed December 20, 2020.
20. Sriram A, Muckley M, Sinha K, et al. COVID-19 Prognosis via Self-Supervised Representation Learning and Multi-Image Prediction. ArXiv 2101.04909. [preprint] https://arxiv.org/abs/2101.04909. Posted January 13, 2021. Accessed December 2020.
21. Mongan J, Moy L, Kahn CE. Checklist for Artificial Intelligence and Medical Imaging (CLAIM). Radiol Artif Intell 2020;2(2):e200029.
22. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015;350:g7594.
23. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet 2019;393(10181):1577–1579.
24. Cruz Rivera S, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med 2020;26(9):1351–1363.
25. Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 2020;26(9):1364–1374.
26. Jacob J, Alexander D, Baillie JK, et al. Using imaging to combat a pandemic: rationale for developing the UK National COVID-19 Chest Imaging Database. Eur Respir J 2020;56(2):2001809.
27. Tsai EB, Simpson S, Lungren M, et al. The RSNA International COVID-19 Open Annotated Radiology Database (RICORD). Radiology 2021. 10.1148/radiol.2021203957. Published online January 5, 2021.
28. Fisher M. Cardiovascular disease and cardiovascular outcomes in COVID-19. Pract Diabetes 2020;37(5):191–193a.
29. Soltan AAS, Kouchaki S, Zhu T, et al. Rapid triage for COVID-19 using routine clinical data for patients attending hospital: development and prospective validation of an artificial intelligence screening test. Lancet Digit Health 2021;3(2):e78–e87.
30. Sysmex UK Ltd Website. Sysmex Europe GmbH. https://www.sysmex.co.uk. Accessed December 2020.