https://doi.org/10.1038/s41467-022-34234-4

Developing medical imaging AI for emerging infectious diseases

Shih-Cheng Huang, Akshay S. Chaudhari, Curtis P. Langlotz, Nigam Shah, Serena Yeung & Matthew P. Lungren

Check for updates

Advances in artificial intelligence (AI) and computer vision hold great promise for assisting medical staff, optimizing healthcare workflow, and improving patient outcomes. The COVID-19 pandemic, which caused unprecedented stress on healthcare systems around the world, presented what seems to be a perfect opportunity for AI to demonstrate its usefulness. However, of the several hundred medical imaging AI models developed for COVID-19, very few were fit for deployment in real-world settings, and some were potentially harmful. This review aims to examine the strengths and weaknesses of prior studies and provide recommendations for different stages of building useful AI models for medical imaging, among them: needfinding, dataset curation, model development and evaluation, and post-deployment considerations. In addition, this review summarizes the lessons learned to inform the scientific community about ways to create useful medical imaging AI in a future pandemic.

The COVID-19 pandemic arose during one of the most innovative periods for biomedical data science, chiefly led by achievements in artificial intelligence (AI), which inspired many efforts globally to leverage digital health data and machine learning techniques to address challenges posed by the pandemic. However, of the innumerable COVID-19-related machine learning efforts developed during the pandemic's most critical time, almost all failed to materialize demonstrable value, and worse, some were potentially harmful[1–5]. That these failures occurred despite the AI and data science community's worldwide united attempt to generate, aggregate, share and utilize large volumes of COVID-19-related data, ranging from simple dashboards, to models for populational-wide risk prediction[6,7], early detection and prognostication[8–12], severity scoring[13–15], long-term outcome and mortality predictions[16–19], is alarming.

Given the sustained belief in the promise of AI for medical imaging and the undelivered promises of machine learning to help during the pandemic, it is essential to summarize the lessons that can be learned from this collective experience to prepare for the future. The purpose of this work is to investigate how we could better position ourselves to leverage artificial intelligence for medical imaging to be useful in a future pandemic. Based on current literature, we provide an evidence-based roadmap for how machine learning technologies in medical imaging can be used to battle ongoing and future pandemics. Specifically, we focus in each section on the four most pressing issues, namely: needfinding, dataset curation, model development and subsequent evaluation, and post-deployment considerations (Fig. 1). For each section, we highlight lessons that can be learned from the shortcomings of prior studies and provide recommendations and guidelines to address them.

Needfinding and planning

"At the root of designing useful tools is the concept of fulfilling a true need[20]". However, this needs-first principle may sometimes be lost in the process when developing AI models for medicine and healthcare. Likewise, the lack of alignment and understanding of the end user's needs has contributed significantly to the limited adaptation of COVID-19 AI tools in real-world settings. Instead of consulting clinicians to identify a true need, many COVID-19 models were developed based on the availability of datasets. For instance, many medical imaging AI models were developed specifically for COVID-19 detection using CT and CXR, even though several radiology societies have suggested that imaging tests should not be used independently to diagnose COVID-19[21,22]. Even if these models can achieve high evaluation metrics, their utility is limited in the absence of a genuine clinical need. Therefore, it is crucially important for AI teams to work closely with clinicians and conduct impact analysis on workflows to identify a true need in healthcare settings that can be fulfilled by automation.

Once a clinical need is identified, it is just as essential to determine a specific "action" to pair with the machine learning model's output to address this clinical need. Most medical image AI models are currently evaluated based on how closely the model's output matches recorded diagnoses or outcomes[23], instead of what matters most to healthcare workers and patients - the ability to bring favorable change in care and improve patient outcomes via a specific action. For example, a COVID-19 model capable of identifying patients at high risk of progression to severe COVID-19 can be paired with the action of administering antiviral therapies such as Ritonavir and Remdesevir[24]. Similarly, a model that tracks COVID-19 progression using time-series medical images can be paired with the action of discharging the patient or requiring supplemental oxygen.

By determining a "decision-action" pair[23], AI developers can instead evaluate the model's utility based on the estimated net benefit in the context of the clinical need (see Model Building and Evaluation section). In a healthcare system, many factors can influence the best course of action, the net incremental value of taking this action, and

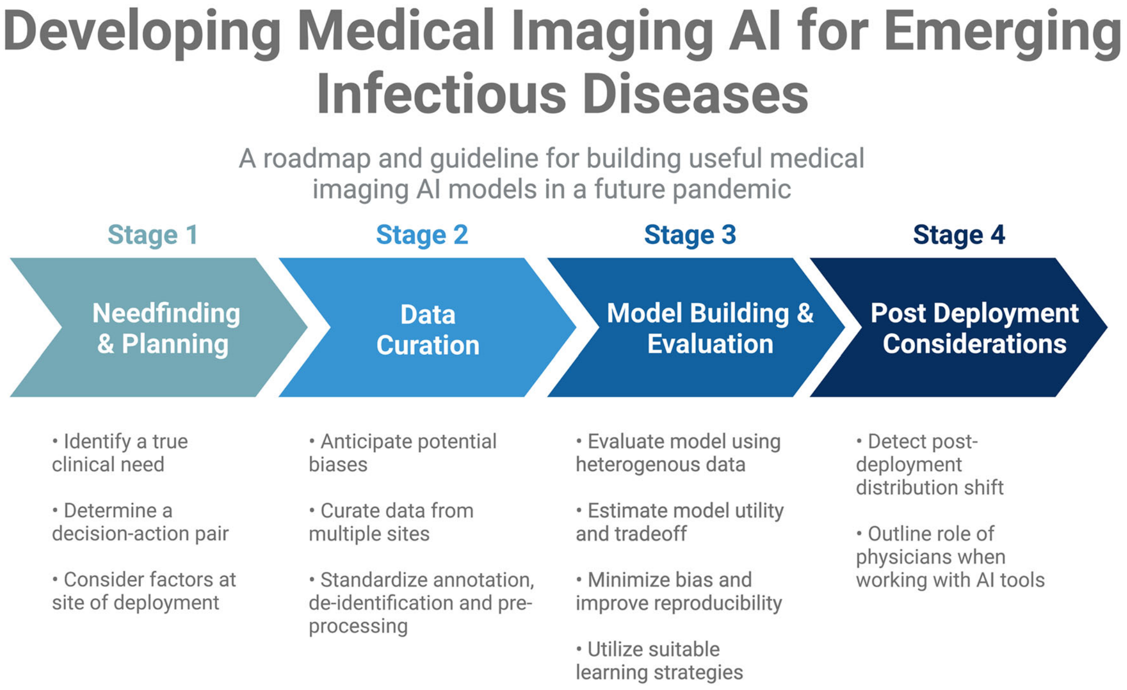
Developing Medical Imaging AI for Emerging Infectious Diseases
A roadmap and guideline for building useful medical imaging AI models in a future pandemic
Stage 1
Needfinding & Planning
Stage 2
Data Curation
Stage 3
Model Building & Evaluation
Stage 4
Post Deployment Considerations
• Identify a true clinical need
• Determine a decision-action pair
• Consider factors at site of deployment
• Anticipate potential biases
• Curate data from multiple sites
• Standardize annotation, de-identification and pre-processing
• Evaluate model using heterogenous data
• Estimate model utility and tradeoff
• Minimize bias and improve reproducibility
• Utilize suitable learning strategies
• Detect post-deployment distribution shift
• Outline role of physicians when working with AI tools

Fig. 1 | The four stages of building useful medical imaging AI models for emerging infectious diseases. We provide recommendations and guidelines for the four stages of medical imaging AI development process in each section of this manuscript, namely: needfinding, dataset curation, model development and subsequent evaluation, and post-deployment considerations. Each stage of the development process should be considered when building medical imaging AI for emerging infectious diseases.

whether the action is taken. Therefore, there is a need to clearly define the care pathway for the patient, including the specifics of who is interpreting the model's output, who is taking action, and how many patients this person can effectively take action for, to determine the optimal decision-action pair. AI developers should continue to consult physicians and healthcare providers to determine the decision-action pair and to estimate the net utility of the action.

It is also crucial for AI developers to thoroughly consider the real-world environment their AI models are intended to deploy at. Many factors at the deployment site, including patient demographics, disease prevalence, clinical workflows, and availability of relevant software or infrastructures, can significantly affect the usability and reliability of the AI model. Take, for example, an AI model that is intended to assist radiologists by highlighting abnormal regions of the medical image. Without infrastructures to integrate this model into the clinicians' workflow or proper software to overlay the model's predicted abnormal regions on the original image, the model's utility is greatly limited. Furthermore, deep learning models are known to be sensitive to distribution shifts, which means that a model's performance can be significantly impacted if factors at the site of deployment cause input data to deviate from the model's training data. For instance, models developed without considering patient position and laterality for chest x-rays or slice thickness and protocol for CT scans at the deployment site might result in an unexpected drop in performance and limited utility. Hence, just as important as identifying a genuine clinical need and determining a prediction action pair, carefully considering aspects at the site of deployment can often mitigate unexpected and undesired outcomes.

Data curation

Curating representative and high-quality data is crucial for building useful machine-learning models. Many medical image datasets were publicly released to encourage other researchers to build COVID-19 AI models. However, a recent systematic review revealed that the majority of more than a hundred datasets the authors identified did not pass their assessments for risk of bias[25]. These biases can cause the machine learning models to learn spurious correlations between predictors and outcomes, and fail to generalize when deployed in hospitals and clinics. For instance, patients with more severe cases of COVID-19 typically get their chest X-rays acquired supine or recumbent in anterior-posterior [AP] projection, whereas healthier patients get imaged while upright (posterior-anterior [PA] projection). Building a model using a dataset that includes images of patients in both AP and PA projections can cause spurious correlations in predicting COVID-19 severity, long-term outcomes, or mortality based on the image projection rather than on semantic image features. These unexpected correlations led to "shortcut learning[26]", which undermined the model's performance and might have been anticipated if a medical imaging specialist had been part of the multidisciplinary team designing the model.

One of the most important yet challenging aspects of building medical imaging AI models is curating large-scale datasets with images from various healthcare settings. Without large-scale data from multiple heterogeneous data sources, models can be biased towards specific patient demographics or overfit certain imaging devices' characteristics, leading to lower performance when deployed at a new institution. However, annotating medical images is time-consuming and cost-prohibitive at scale. Thus, most publicly available COVID-19 medical image datasets that met the criteria for proper risk and bias assessment are relatively small, with only several hundred labeled images[25]. Furthermore, datasets that contain more than a few thousand imaging studies tend to be limited to a particular healthcare system or country[27,28], likely due to the lack of a standardized strategy to preserve patient privacy when sharing data. To improve model performance and generalizability, many studies splice together data from multiple sources to create "remix datasets". While this might seem like a promising solution, models that appear promising in the course of development can have lower performance when deployed. For example, the machine learning model might learn to identify the hospital instead of COVID-19 based on pixel distribution of the hospital's imaging hardware or the hospital's watermark and may predict COVID-19 solely based on identifying images from hospitals with high COVID-19 prevalence. Additionally, different datasets might have different standards for annotation (i.e., PCR test results vs. expert opinion) or different tolerance for inter-rater variability, which can increase uncertainty for the model. Therefore, ensuring label integrity and standardization between different sites and datasets is essential.

Data-sharing frameworks that standardize or account for image acquisition, de-identification, pre-processing, annotation, and quality assurance from different hospitals while preserving patient privacy are critical for developing useful models. One notable example is RSNA's RICORD[29], a public COVID-19 database with medical images assembled from sites worldwide. This effort clearly defined data inclusion criteria, developed a data de-identification protocol (RSNA anonymizer), and utilized a cloud-based data annotation tool (MD.ai) to standardize data sharing while preserving patient privacy. Furthermore, multiple radiologists were involved in the annotation process to ensure label integrity and ascertain that labels were chosen with clinically-driven goals. Additionally, RICORD requires all sites to provide a detailed description of the data, imaging metadata for each imaging exam, and relevant patient medical data. RICORD is now hosted on the Medical Imaging and Data Resource Center (MIDRC), a linked-data commons explicitly built with AI in mind. Efforts such as RICORD and MIDRC enable easy access to large-scale multi-institutional medical images along with standardized and high-quality annotations, thereby providing more diverse and better representations of patient demographic and imaging acquisition methodology, which may help mitigate potential confounders.

Model building and evaluation

Thorough and proper evaluation of AI models before deployment is crucial. Without evaluations in different settings and populations, irrational confidence in model's performance can follow, which can have deleterious consequences when the model is deployed. On top of evaluating models using heterogeneous data from multiple sites and representative data that the models are expected to ingest during deployment, it is also vital to evaluate AI models across different demographic groups. Without representative data or the right training objective, AI models can be biased against specific protected attributes such as age, gender, and race[30]. For example, in CheXclusion, the authors showed that state-of-the-art deep learning models for Chest X-rays are biased to demographic attributes[31]. In addition, Banerjee et al. have shown that medical image AI models can be explicitly trained to predict self-reported races with high accuracy—something even medical experts cannot[32]. This indicates that medical imaging models have the potential to use spurious correlations between patient demographics and the outcome of interest as shortcuts for prediction, which will render the prediction unreliable or even outright harmful. Therefore, methods such as true positive rates, statistical parity, group fairness, and equalized odds should always be considered for detecting algorithmic bias before deployment.

If algorithmic bias is detected for a model, AI developers should collaborate with clinicians to identify the sources of bias and make adjustments to improve the model[33]. Leveraging large-scale data from multiple sources has been demonstrated as an effective strategy for combating the risk of algorithmic bias[31]. However, sharing data across different hospitals while preserving patient privacy might not always be feasible. Alternatively, Obermeyer et al. found that biases are sometimes attributed to label choice since the labels are often measured with errors that reflect structural inequalities. Specifically, the model in their study uses health cost as a proxy for health severity, and since historical data have exhibited reduced healthcare spending on Black patients, the model learned to be biased against Black patients who are as sick as white patients. Instead of modifying the model or the input data, changing the labels used to train the model could address algorithmic bias in certain situations[34]. Other strategies to address model bias[35] include (1) pre-processing approaches such as oversampling of minority groups[36], (2) in-processing approaches that add explicit constraints in the loss function to minimize performance difference between subgroups[37,38], and (3) post-processing approaches, such as equalized odds to correct the outputs based on the individual's group[39–41].

In the "Needfinding and planning" section, we emphasized the importance of determining utility via the model's decision–action pair. Knowing the specific actions associated with a model's outputs allows us to determine the model's utility and tradeoff. Unfortunately, most studies chose models based on evaluation metrics such as AUROC and accuracy, which are statistical abstractions that do not directly relate to improvements in medical care[42]. Instead, models should also be evaluated based on the estimated net incremental value or utility of taking specific actions due to the model's prediction. This allows AI developers to determine directly if a particular action based on a model's output would bring more benefit or harm to the patient. One way to define utility is by estimating the net dollar value of taking a specific action based on the model's prediction, including costs of intervention, allocation of resources, and future patient health expenses. Once the utility is defined, models can be evaluated using the indifference curve—a line on the ROC plot that shows combinations of sensitivity and specificity that result in the same utility[23]. Not only should the indifference curve be used to choose the operating point of a model, it should also be used to choose the model to deploy in hospitals and clinics. As stated by Irwin and Irwin, "ROC curves show what is feasible, indifference curves show what is desirable. Together they show what should be chosen[43]". It is worth noting that a model can have lower AUROC, and yet an operating point with higher utility. Other studies have shown that it is possible to incorporate utility directly into model building and thus achieve tighter alignment to patient outcomes[44]. While a study of net incremental value is not necessary for AI publications, it is crucial to consider the net utility when comparing models for deployment in real-world settings.

Although very few COVID-19 AI models failed to provide clinical value solely due to their learning strategy, it is still important to highlight learning strategies that could significantly benefit model developments for future efforts. In situations where frameworks have been set up to share labeled data across sites, we advocate that data providers also share pertinent medical information to encourage the development of multimodal fusion models. It has been repeatedly shown that patient medical history and laboratory data are typically required to enable physicians to interpret medical images in the appropriate clinical context[45,46]. For instance, many medical characteristics have been linked to COVID-19 severity and long-term outcomes[47]. Thus, building machine learning models without these patient medical records could limit the capabilities of an AI model. In fact, a systematic review has shown that medical image models that fuse data from multiple modalities generally lead to increased performance compared to the performance of single modality models[48–51]. Similarly, several COVID-19 medical imaging models incorporating clinical variables have observed improvement in performances over imaging-only models[8,11].

If the appropriate data-sharing framework is unavailable, models can still be trained across multiple institutions without explicitly sharing data using federated learning - a collaborative machine learning paradigm. Instead of bringing data to the machine learning model, the model is distributed and trained within the firewalls of each institution[52]. Existing studies have already demonstrated that a federated learning framework can lead to nearly the same performance as that of models that use centralized data or ensembling strategies[53,54]. More recently, blockchain-based frameworks allowed decentralized learning without a central coordinator, promoting equality in training multicentric models[55]. Several software frameworks for enabling federated learning in real-world settings already exist, including NVIDIA CLARA[56], Intel OpenFL[57], and Flower[58]. Furthermore, federated evaluation platforms, such as MLPerf[59], allow developers to evaluate the generalizability of their models across multiple sites while preserving patient privacy.

While an abundance of medical images exists in most healthcare institutions, training machine learning models using the traditional supervised paradigm requires annotations by medical experts and thus is cost-prohibitive at scale. Self-supervised pre-training strategies allow machine learning models to obtain supervisory signals from the data without explicit labels, thus allowing models to learn from all available data even if some are not annotated[60]. Furthermore, after the self-supervised pre-training stage, these models can be fine-tuned for many downstream tasks with limited number of labels. Studies have shown that medical image models pre-trained using self-supervised learning are robust to distribution shift and require fewer labels during supervised fine-tuning[61,62]. Recent studies have also demonstrated the possibility of multimodal self-supervised learning for medical images by leveraging the corresponding radiology reports, and have demonstrated superior performance using only 1% of the training labels compared to supervised models[63,64]. Yan et al. have further proposed training self-supervised models with data from multiple healthcare centers using Federated Learning and have shown improvement in robustness and performance over models trained using data from a single institution[65].

Explainable artificial intelligence (AI) techniques, including saliency maps, generative adversarial networks (GANs), and counterfactual explanations (CE), can be used to identify spurious correlations and confounders the model relies on to make predictions. For example, by using saliency maps and GANs, DeGrave et al. found that models rely on regions outside of the lung fields, such as laterality markers, to make predictions on COVID-19 datasets. While prior work has argued that saliency should not be used for explanation[66], counterfactual explanation techniques have been shown to reveal learned correlations to the model's prediction. By applying minimal but meaningful perturbations of an input image to change the original prediction of a model, CEs can provide explanations that are understandable by humans. In addition to detecting spurious correlations, explainable AI techniques can also reveal human-subliminal signals about the disease and give us a better chance of addressing challenges posed by the pandemic. Lastly, slice discovery methods[67,68], such as Domino[69], can identify and describe semantically meaningful subsets of data on which the model performs poorly, thereby revealing spurious correlations.

Post deployment considerations

Even after a model is thoroughly evaluated and deemed suitable for deployment, several challenges can still impact the model's ability to bring favorable change for patients post-deployment. One such challenge is post-deployment distribution shift. Most medical image models are trained on a static dataset curated from a specific label definition, patient population, timeframe, scanner type, and protocol. However, real-time changes in hospital workflow or disease prevalence may alter the model's input data distribution, causing models' performance to degrade. There are two major types of distribution shift: concept shift and data shift[21]. Concept shift includes the arrival of a novel class or class evolution, such as bacterial pneumonia vs COVID-19 pneumonia. Data shift can have many causes, including new imaging modalities, new versions of software, or adjustments to data acquisition procedures. For example, if a COVID-19 model is trained to predict patient mortality with pre-vaccination COVID-19 patient data, it would dangerously overestimate the risk of death in real-world data during deployment after the vaccine was widely distributed. One way to detect distribution shifts is by monitoring the model's performance longitudinally and checking for statistically significant performance drops. While this method effectively detects label shifts, it is challenging to implement in practice as it requires periodic data annotation. A more tangible approach is based on detecting shifts in the input data distribution. This can be done by measuring differences between real-world and training data using two-sampled-test-based statistical approaches[70]. Alternatively, methods for modeling predictive uncertainty, such as Prior Networks[71], can be used to detect data shifts in real time. Hospitals should establish protocols and procedures for adjusting the model as soon as a distribution shift is detected to maintain optimal model performance and patient safety.

Another challenge is clearly outlining the role of physicians when working with AI tools. Some existing literature has argued that the role of radiologists in the AI era is to become educated consumers of AI tools by identifying clinical needs for AI, evaluating AI tools before deployment, and maintaining their expertise by avoiding overreliance on technology[72–74]. Furthermore, it is crucial to educate clinicians about the limitations of AI. This allows clinicians to help identify potential distribution shifts, either by reporting conflicting diagnoses between that made by the model and that made by themselves or by noting changes in their medical practice[75]. Hospitals should also make sure concerns about an AI tool's functionality or factors that can potentially change data or label distributions can be easily reported by clinicians. In addition, educating radiologists on the limitations of AI can prevent them from blind acceptance of the AI's output. Overreliance on AI tools can diminish the physician's perspective and diagnostic skills. Furthermore, a machine learning tool is typically limited to the few diagnoses it is intended for. It is, therefore, still important for radiologists to look for other abnormalities in the image that the model cannot detect. In other words, clinicians should still make their independent diagnoses and use AI tools only as supplements or aids.

Conclusions

While advancements in AI for medical imaging hold great promise in improving healthcare, many overlooked aspects of the model development process and use lifecycle have hindered large-scale deployment during the most critical time of the COVID-19 pandemic. Independent of the choice of learning strategies or model architectures, many COVID-19 AI models were unfitting for use due to shortcomings in needfinding, data curation, model evaluation, and post-deployment considerations. We urge researchers to be mindful of the aforementioned potential biases and limitations currently hindering the deployment of medical imaging AI models. In addition, considerable planning is necessary to prepare for future health crises, including developing data-sharing frameworks and implementing AI deployment infrastructures in hospitals. We hope this review will inform the data science and healthcare community about strategies for building useful medical imaging AI models for future infectious diseases.

Shih-Cheng Huang[1,2] ✉, **Akshay S. Chaudhari**[1,2,3], **Curtis P. Langlotz**[1,2,3], **Nigam Shah**[1], **Serena Yeung**[1,2,4,5,6,7] & **Matthew P. Lungren**[1,2,3,7]

[1]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [2]Center for Artificial Intelligence in Medicine & Imaging, Stanford University, Stanford, CA, USA. [3]Department of Radiology, Stanford University, Stanford, CA, USA. [4]Department of Computer Science, Stanford University, Stanford, CA, USA. [5]Department of Electrical Engineering, Stanford University, Stanford, CA, USA. [6]Clinical Excellence Research Center, Stanford University School of Medicine, Stanford, CA, USA. [7]These authors contributed equally: Serena Yeung, Mattew P. Lungren.
✉e-mail: mschuang@stanford.edu

Received: 30 April 2022; Accepted: 19 October 2022;
Published online: 18 November 2022

References

1. Data science and AI in the age of COVID-19—report. *The Alan Turing Institute* https://www.turing.ac.uk/research/publications/data-science-and-ai-age-covid-19-report (2021).
2. Roberts, 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.* **3**, 199–217 (2021).
3. Li, M. D. et al. Radiology implementation considerations for artificial intelligence (AI) applied to COVID-19, from the AJR special series on AI applications. *AJR Am. J. Roentgenol.* **219**, 15–23 (2022).
4. Wynants, L. et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* **369**, m1328 (2020).
5. Born, J. et al. On the role of artificial intelligence in medical imaging of COVID-19. *Patterns* **2**, 100269 (2021).
6. Jiang, Z. et al. Combining visible light and infrared imaging for efficient detection of respiratory infections such as COVID-19 on portable device. Preprint at https://arxiv.org/abs/2004.06912 (2020).
7. Liu, Y. et al. A COVID-19 risk assessment decision support system for general practitioners: design and development study. *J. Med. Internet Res.* **22**, e19786 (2020).
8. Mei, X. et al. Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. *Nat. Med.* **26**, 1224–1228 (2020).
9. Barbosa, E. J. M. et al. Machine learning automatically detects COVID-19 using chest CTs in a large multicenter cohort. *Eur. Radiol.* **31**, 8775–8785 (2021).
10. Guiot, J. et al. Development and validation of an automated radiomic CT signature for detecting COVID-19. *Diagnostics* **11**, 41 (2020).
11. Chen, X. et al. A diagnostic model for coronavirus disease 2019 (COVID-19) based on radiological semantic and clinical features: a multi-center study. *Eur. Radiol.* **30**, 4893–4902 (2020).
12. Xu, M. et al. Accurately differentiating COVID-19, other viral infection, and healthy individuals using multimodal features via late fusion learning. Preprint at *bioRxiv* https://doi.org/10.1101/2020.08.18.20176776 (2020).
13. Frid-Adar, M., Amer, R., Gozes, O., Nassar, J. & Greenspan, H. COVID-19 in CXR: from detection and severity scoring to patient disease monitoring. *IEEE J. Biomed. Health Inf.* **25**, 1892–1903 (2021).
14. Li, M. D. et al. Multi-population generalizability of a deep learning-based chest radiograph severity score for COVID-19. *Medicine* **101**, e29587 (2022).
15. Li, M. D. et al. Automated assessment and tracking of COVID-19 pulmonary disease severity on chest radiographs using convolutional siamese neural networks. *Radiology: Artif. Intell.* **2**, e200079 (2020).
16. Schalekamp, S. et al. Model-based prediction of critical illness in hospitalized patients with COVID-19. *Radiology* **298**, E46–E54 (2021).
17. Ramtohul, T. et al. Quantitative CT extent of lung damage in COVID-19 pneumonia is an independent risk factor for inpatient mortality in a population of cancer patients: a prospective study. *Front. Oncol.* **10**, 1560 (2020).
18. Zheng, Y. et al. Development and validation of a prognostic nomogram based on clinical and CT features for adverse outcome prediction in patients with COVID-19. *Korean J. Radiol.* **21**, 1007–1017 (2020).
19. Chen, Y. et al. A Quantitative and Radiomics approach to monitoring ARDS in COVID-19 patients based on chest CT: a retrospective cohort study. *Int. J. Med. Sci.* **17**, 1773–1782 (2020).
20. Gong, J., Currano, R., Sirkin, D., Yeung, S. & Holsinger, F. C. NICE: four human-centered AI principles for bridging the AI-to-clinic translational gap. (2021).
21. Lekadir, K. et al. FUTURE-AI: guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging. Preprint at https://arxiv.org/abs/2109.09658 (2021).
22. Rubin, G. D. et al. The role of chest imaging in patient management during the COVID-19 pandemic: a multinational consensus statement from the Fleischner Society. *Radiology* **296**, 172–180 (2020).
23. Shah, N. H., Milstein, A. & Bagley, S. C. Making machine learning models clinically useful. *J. Am. Med. Assoc.* **322**, 1351–1352 (2019).
24. COVID-19 Treatment Guidelines Panel. *Coronavirus Disease 2019 (COVID-19) Treatment Guidelines*. National Institutes of Health https://www.covid19treatmentguidelines.nih.gov/ (2022).
25. Garcia Santa Cruz, B., Bossa, M. N., Sölter, J. & Husch, A. D. Public Covid-19 X-ray datasets and their impact on model bias—a systematic review of a significant problem. *Med. Image Anal.* **74**, 102225 (2021).
26. DeGrave, A. J., Janizek, J. & Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat. Mach. Intell.* **3**, 610–619 (2021).
27. de la Iglesia Vayá, M. et al. BIMCV COVID-19+: a large annotated dataset of RX and CT images from COVID-19 patients. Preprint at https://arxiv.org/abs/2006.01174 (2020).
28. Signoroni, A. et al. BS-Net: learning COVID-19 pneumonia severity on a large chest X-ray dataset. *Med. Image Anal.* **71**, 102046 (2021).
29. Bai, H. X. & Thomasian, N. M. RICORD: a precedent for open AI in COVID-19 image analytics. *Radiology* **299**, E219–E220 (2021).
30. Zhou, Y. et al. RadFusion: benchmarking performance and fairness for multimodal pulmonary embolism detection from CT and EHR. Preprint at https://arxiv.org/abs/2111.11665 (2021).
31. Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I. Y. & Ghassemi, M. CheXclusion: fairness gaps in deep chest X-ray classifiers. *Pac. Symp. Biocomput.* **26**, 232–243 (2021).
32. Gichoya, J. W. et al. AI recognition of patient race in medical imaging: a modelling study. *Lancet Digit Health* **4**, e406–e414 (2022).
33. Chen, I. Y., Szolovits, P. & Ghassemi, M. Can AI help reduce disparities in general medical and mental health care? *AMA J. Ethics* **21**, E167–E179 (2019).
34. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
35. d'Alessandro, B., O'Neil, C. & LaGatta, T. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big data* **5**, 120–134 (2017).
36. Bellamy, R. K. E. et al. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development 63.4/5 (2019): 4-1.
37. Berk, R. et al. A convex framework for fair regression. Preprint at https://arxiv.org/abs/1706.02409 (2017).
38. Chohlas-Wood, A., Coots, M., Zhu, H., Brunskill, E. & Goel, S. Learning to be fair: a consequentialist approach to equitable decision-making. Preprint at https://arxiv.org/abs/2109.08792 (2021).
39. Chouldechova, A. & Roth, A. The frontiers of fairness in machine learning. Preprint at https://arxiv.org/abs/1810.08810 (2018).
40. Corbett-Davies, S. & Goel, S. The measure and mismeasure of fairness: a critical review of fair machine learning. Preprint at https://arxiv.org/abs/1808.00023 (2018).
41. Barocas, S, Hardt, M & Narayanan, A. Fairness and machine learning. https://fairmlbook.org (2019).
42. Vickers, A. J., Van Calster, B. & Steyerberg, E. W. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* **352**, i6 (2016).

43. Irwin, R. J. & Irwin, T. C. A principled approach to setting optimal diagnostic thresholds: where ROC and indifference curves meet. *Eur. J. Intern. Med.* **22**, 230–234 (2011).
44. Ko, M. et al. Improving hospital readmission prediction using individualized utility analysis. *J. Biomed. Inform.* **119**, 103826 (2021).
45. Leslie, A., Jones, A. J. & Goddard, P. R. The influence of clinical information on the reporting of CT by radiologists. *Br. J. Radiol.* **73**, 1052–1055 (2000).
46. Cohen, M. D. Accuracy of information on imaging requisitions: does it matter? *J. Am. Coll. Radiol.* **4**, 617–621 (2007).
47. Li, X. et al. Clinical determinants of the severity of COVID-19: a systematic review and meta-analysis. *PLoS One* **16**, e0250602 (2021).
48. Huang, S.-C., Pareek, A., Zamanian, R., Banerjee, I. & Lungren, M. P. Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. *Sci. Rep.* **10**, 22147 (2020).
49. Huang, S.-C., Pareek, A., Seyyedi, S., Banerjee, I. & Lungren, M. P. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digital Med.* **3**, 136 (2020).
50. Esteva, A. et al. Prostate cancer therapy personalization via multi-modal deep learning on randomized phase III clinical trials. *npj Digital Med.* **5**, 71 (2022).
51. Esteva, A. et al. Development and validation of a prognostic AI biomarker using multi-modal deep learning with digital histopathology in localized prostate cancer on NRG Oncology phase III clinical trials. *J. Clin. Orthod.* **40**, 222–222 (2022).
52. Larson, D. B., Magnus, D. C., Lungren, M. P., Shah, N. H. & Langlotz, C. P. Ethics of using and sharing clinical imaging data for artificial intelligence: a proposed framework. *Radiology* **295**, 675–682 (2020).
53. Sheller, M. J. et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci. Rep.* **10**, https://doi.org/10.1038/s41598-020-69250-1 (2020).
54. Chang, K. et al. Distributed deep learning networks among institutions for medical imaging. *J. Am. Med. Inform. Assoc.* **25**, 945–954 (2018).
55. Saldanha, O. L. et al. Swarm learning for decentralized artificial intelligence in cancer histopathology. *Nat. Med.* https://doi.org/10.1038/s41591-022-01768-5 (2022).
56. NVIDIA Clara. *NVIDIA Developer* https://developer.nvidia.com/clara (2019).
57. Anthony Reina, G. et al. OpenFL: an open-source framework for federated learning. Preprint at https://arxiv.org/abs/2105.06413 (2021).
58. Beutel, D. J. et al. Flower: a friendly federated learning research framework. Preprint at https://arxiv.org/abs/2007.14390 (2020).
59. Karargyris, A. et al. MedPerf: open benchmarking platform for medical artificial intelligence using federated evaluation. Preprint at https://arxiv.org/abs/2110.01406 (2021).
60. LeCun, Y & Misra, I. Self-supervised learning: the dark matter of intelligence. https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/ (2021).
61. Sowrirajan, H. et al. Moco pretraining improves representation and transferability of chest x-ray models. Medical Imaging with Deep Learning. (PMLR, 2021).
62. Azizi, S. et al. Big self-supervised models advance medical image classification. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
63. Zhang, Y., Jiang, H., Miura, Y., Manning, C. D. & Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. Preprint at https://arxiv.org/abs/2010.00747 (2020).
64. Huang, S.-C., Shen, L., Lungren, M. P. & Yeung, S. GLoRIA: a multimodal global-local representation learning framework for label-efficient medical image recognition. in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE, 2021).
65. Yan, R. et al. Label-efficient self-supervised federated learning for tackling data heterogeneity in medical imaging. Preprint at https://arxiv.org/abs/2205.08576 (2022).
66. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
67. d'Eon, G., d'Eon, J., Wright, J. R. & Leyton-Brown, K. The spotlight: a general method for discovering systematic errors in deep learning models. in *2022 ACM Conference on Fairness, Accountability, and Transparency* 1962–1981 (Association for Computing Machinery, 2022).
68. Yeh, C.-K. et al. On completeness-aware concept-based explanations in deep neural networks. *Advances in Neural Information Processing Systems* **33**, 20554–20565 (2020).
69. Eyuboglu, S. et al. Domino: discovering systematic errors with cross-modal embeddings. Preprint at https://arxiv.org/abs/2203.14960 (2022).
70. Rabanser, S., Günnemann, S. & Lipton, Z. Failing loudly: An empirical study of methods for detecting dataset shift. *Advances in Neural Information Processing Systems* **32** (2019).
71. Malinin, A. & Gales, M. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems* **31** (2018).
72. Rubin, D. L. Artificial intelligence in imaging: the radiologist's role. *J. Am. Coll. Radiol.* **16**, 1309–1317 (2019).
73. Richardson, M. L. et al. Review of artificial intelligence training tools and courses for radiologists. *Acad. Radiol.* **28**, 1238–1252 (2021).
74. Allen, B. et al. Evaluation and real-world performance monitoring of artificial intelligence models in clinical practice: try it, buy it, check it. *J. Am. Coll. Radiol.* **18**, 1489–1496 (2021).
75. Finlayson, S. G. et al. The clinician and dataset shift in artificial intelligence. *N. Engl. J. Med.* **385**, 283–286 (2021).

Acknowledgements

Research reported in this publication was supported by NIH grants R01 HL155410 and R01 LM012966. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

Concept and design: S.-C.H. and M.P.L. Drafting of the manuscript: S.-C.H. Critical revision of the manuscript for important intellectual content: S.-C.H., A.S.C., C.P.L., N.S., S.Y., and M.P.L. Supervision: A.S.C., C.P.L., N.S., S.Y., and M.P.L.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Shih-Cheng Huang.

Peer review information *Nature Communications* thanks Synho Do, Andrea Laghi, Michael Richardson and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.