Title

RSNA-MICCAI Panel Discussion: Machine Learning for Radiology from Challenges to Clinical Applications.

Permalink

https://escholarship.org/uc/item/112660b4

Journal

Radiology: Artificial Intelligence, 3(5)

Authors

Mongan, John

Kalpathy-Cramer, Jayashree

Flanders, Adam

et al.

Publication Date

2021-09-01

DOI

10.1148/ryai.2021210118

Peer reviewed

Radiology: Artificial Intelligence

SPECIAL REPORT

RSNA-MICCAI Panel Discussion: Machine Learning for Radiology from Challenges to Clinical Applications

John Mongan, MD, PhD • Jayashree Kalpathy-Cramer, PhD • Adam Flanders, MD • Marius George Linguraru, DPhil

From the Department of Radiology and Biomedical Imaging and Center for Intelligent Imaging, University of California San Francisco, 505 Parnassus Ave, Room M-391, San Francisco, CA 94143 (J.M.); Department of Radiology and MGH and BWH Center for Clinical Data Science, Massachusetts General Hospital, Boston, Mass (J.K.C.); Department of Radiology, Thomas Jefferson University Hospital, Philadelphia, Pa (A.F.); Sheikh Zayed Institute for Pediatric Surgical Innovation, Children's National Hospital, Washington, DC (M.G.L.); and Departments of Pediatrics and Radiology, George Washington University School of Medicine, Washington, DC (M.G.L.). Received May 6, 2021; revision requested June 4; revision received June 30; accepted July 12. **Address correspondence to** J.M. (e-mail: *john.mongan@ucsf.edu*).

Authors declared no funding for this work.

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

Radiology: Artificial Intelligence 2021; 3(5):e210118 • https://doi.org/10.1148/ryai.2021210118 • Content code: AI

On October 5, 2020, the Medical Image Computing and Computer Assisted Intervention Society (MICCAI) 2020 conference hosted a virtual panel discussion with members of the Machine Learning Steering Subcommittee of the Radiological Society of North America. The MICCAI Society brings together scientists, engineers, physicians, educators, and students from around the world. Both societies share a vision to develop radiologic and medical imaging techniques through advanced quantitative imaging biomarkers and artificial intelligence. The panel elaborated on how collaborations between radiologists and machine learning scientists facilitate the creation and clinical success of imaging technology for radiology. This report presents structured highlights of the moderated dialogue at the panel.

Challenges

How do you incentivize annotators to participate?

A.F.: For the RSNA challenges, we have been pleasantly surprised by the overwhelming interest and willingness to donate time to these projects. For example, in 2019, the American Society of Neuroradiology (ASNR) put out an open call for volunteers. Within 1 day, more than 140 radiologists not only volunteered, but also provided supplemental information about what motivated them to take part in the Intracranial Hemorrhage Detection Challenge. Many of the respondents expressed immense interest in artificial intelligence (AI) technologies but had no direct opportunity to participate in their own work environment. Because the call for volunteers went out to the entire ASNR membership, there was a mixture of respondents from academia and private practice. For many of the respondents, this was their first foray into an AI project, particularly one of this scale, which used a large, heterogeneous, multinational dataset and international domain expertise. There were additional incentives for individuals who performed high volumes of annotations. The volunteers who provided the largest number of high-quality annotations were included as authors of the initial manuscript. The remainder were acknowledged as collaborators on the organization website, the challenge website, and in the initial manuscript. We also provided regular informational blasts on our group e-mail to keep all of the annotators apprised of the status of the challenge. We explained key concepts about how the challenge operated and some of the specific issues we were addressing regarding bias, data curation, adjudication of annotations, and division of the dataset into respective training, validation, and test sets. This helped to keep the annotators engaged in the entire process up until closure and presentation of the awards at the RSNA Annual Meeting in December.

J.M.: Another important aspect is to facilitate participation. We try to provide clear instructions and easy-to-use annotation tools to make the process as enjoyable, efficient, and frustration-free as possible.

What are the main advances or additions brought by the series of RSNA challenges to the clinical field?

A.F.: The most important outcomes have been related to engagement of interested individuals in a range of disciplines and promotion of "team science" in addressing the sorts of questions posed by challenges. The challenges have attracted data scientists who have never considered working in the medical imaging space and drawn clinical radiologists to collaborate on projects that they may not have had sufficient background to initiate on their own. The challenges take place on a platform well known to the international data science community that provides a rich resource for collaboration between contestants. It also serves as a singular resource for all things related to machine learning and AI, so even the AI neophyte can find many items to pique their interest.

J.K.C.: Publication is another tangible output—some challenge datasets have more than 2000 citations. Once a dataset is public, there is frequently an increase in publications on the topic. There also seems to be a correlation between challenges and products becoming commercially available. For instance, the number of vendors offering algorithms for bone age, intracranial hemorrhage, and pneumonia seemed to tick up in the period following each of these challenges.

Abbreviations

AI = artificial intelligence, MICCAI = Medical Image Computing and Computer Assisted Intervention Society, RSNA = Radiological Society of North America

Summary

Collaboration between data scientists and clinical radiologists is essential for advancing the state of the art of clinical radiology through application of artificial intelligence.

Keywords

Back-Propagation, Artificial Neural Network Algorithms, Machine Learning Algorithms

Medical annotations are known to often have high inter-rater variability. This variability could bring rich information to AI methods. There is, of course, a cost in labeling images multiple times, but given the possible benefits, should we be proposing challenges with multiple raters?

J.K.C.: I am in favor! The RSNA challenges typically have used multiple raters, at least for the test set. We definitely have seen variability among our raters and would find a consensus process to be useful.

J.M.: Historically when we've done multiple annotations, we've released only the consensus annotation; there may be useful information in the individual source annotations. We will look to include that in future dataset releases, where possible.

Annotation

What are the advancements in PACS and clinical radiology software that can improve labeling and report generation, so that we reduce the effort needed to create high-quality AI models using extracted clinical radiology data?

J.K.C.: To the extent that we can make the annotations at the time of reading more machine readable and thus usable for AI algorithm development, we ease the task of later curating the datasets for use. Structured reporting templates can help. Some of the new generations of PACS and visualization tools do allow the generation and storage of annotations and markup (eg, bounding boxes, segmentations) in standard formats (DICOM-SR, DICOM-SEG, FHIR). These standards are key for interoperability and can make it easier to mine higher quality data from clinical archives. On the technical side, methods are being developed to use more "weakly labeled" data for training.

Training machine learning models often requires specific annotations that are time-consuming to acquire and generate. Is there a good way to build this process into the design of collaborative projects between AI researchers and radiologists?

J.M.: Annotations typically come from either a human expert interpretation of the data or some independent or semi-independent measure of the outcome or variable in question. The form of the annotation should be driven by consideration of the clinical use case and what kind of annotations are required to build a model that solves the clinical problem. Doing so effectively requires collaboration between AI researchers and radiologists. Because annotations are often time-consuming and expensive to acquire, it's particularly important to define the form or schema of annotations before annotation begins. When data are annotated in the absence of such collaboration, the result is often a dataset that doesn't support creating an AI model that meaningfully addresses the clinical problem.

A.F.: There is a need to balance efficiency, accuracy, and time constraints of annotations created by human experts. The process of labeling, segmenting, or marking up an image for a data science challenge is very different when it is done for clinical reasons in contrast to a research project. A radiologist naturally will provide greater focus to the task when a patient and a provider are depending on the accuracy of the interpretation. When annotating medical images for a data science challenge, the radiologist's cognitive tasks must be intuitive and efficient so that greater volumes of images can be addressed in a single session. All of these factors come into play when designing appropriate annotations for a challenge.

Should we be collecting machine-level "raw" data instead of the reconstructed image data, and base classification and segmentation on that data?

J.M.: It's an interesting idea, and there may be something to be gained by this approach. There are several challenges involved. First, raw data are often an order of magnitude larger than reconstructed images, and are often in less standardized, more vendor-specific formats than the relatively universal DICOM standard we have for reconstructed images. In clinical practice, raw data typically don't leave the scanner, so methods for extraction are less well developed. Raw data aren't typically archived, so projects dependent on raw data would need to assemble datasets prospectively rather than retrospectively. Humans can't directly interpret or annotate raw data, so annotations likely would need to come from a nonimaging source and are likely to be both "weaker" (applied at the study or patient level, rather than the anatomic level) and noisier (ie, imperfectly correlated with imaging findings). These obstacles are major, but not entirely insurmountable—the question would be whether there's sufficient additional information in the raw data beyond what's in reconstructed images to merit taking on these issues.

J.K.C.: There is a lot of interest in "upstream" AI for improving image quality, reducing artifacts, reducing radiation or contrast material dose, or speeding up the scans, where collecting the "raw" data is imperative (1). Computational methods to directly classify and segment based on k-space or sinogram data are still in their infancy as conventional networks such as convolutional neural networks have mostly been applied in image or spatial domains, not in sensor or frequency domains.

Clinical Deployment and Implementation

How important is user experience as a factor for clinical uptake and usage of AI algorithms? Should more research be focused on the usability and understandability of AI and not just the methodology behind it?

J.M.: User experience is key. The extent to which your tool or algorithm will be used is affected by both the extent to which it's an improvement over current practice and the user experience. If it's difficult, confusing, and time-consuming to use your algorithm, you might need to be 1000% better than current practice to get people to use it. If it fits seamlessly into what they're already doing, a 10% improvement might be sufficient. Fitting seamlessly into a clinical workflow doesn't happen by accident; it has to be the goal from the beginning of the project to be achieved. The requirements for this fit drive the form of integration, the structure of the model, and often even the annotation and requirements of the training dataset. I think there should be more focus on studying the usability and integration of AI in the clinical workflow. Any project that intends to develop AI for clinical use should have at least one person who is familiar with the relevant clinical workflow involved in planning and executing the project from its earliest stages.

Is it possible for academics and clinicians to move something into clinical practice alone or is the involvement of industry essential?

J.K.C.: Technically, I would argue it is possible. Some institutions have deployed homegrown tools. However, the risk profile and the regulatory issues need to be considered carefully. Institutions seem to be taking a wide range of approaches in terms of what can be deployed.

J.M.: It's definitely the case that academics and clinicians have moved AI models into clinical practice without industry involvement, but the scale is almost always limited to a few institutions, usually the home institutions of the investigators who developed the AI. In my opinion, you can't have any significant impact on patient care without industry partnership. Put simply, if your model can't be made into a commercial product, it won't ever have impact beyond a few academic centers. There are very few nonacademic medical centers that are ever going to download a model from the Internet and stand it up in clinical practice. Ideally, industry perspectives are part of the conversation of defining the clinical use case and point of integration into clinical workflow that happens before AI model development. Just as it can be nearly impossible to effectively integrate a model into a clinical workflow that it wasn't designed for, the chances that a model will be clinically viable as a product are low unless that was part of the design considerations from the beginning.

Are there lessons to be learned from the airline industry in terms of the relationship between experts, regulators, and technical innovators?

J.M.: Yes, I think so. A lot of the quality and safety innovations in medicine have originated in aviation. I think there's a lot that can be learned from the successes and failures of the airline industry, particularly because they have much more lengthy experience with deployment and use of AI than medicine does, in the form of autopilots. I recently wrote an editorial on this topic, looking specifically at the 737 MAX disasters and what lessons medicine and radiology can take from them to try to avoid repeating the same mistakes (2). One area for concern is that the delegated regulatory models being proposed for AI in medicine are very similar to the regulatory model the Federal Aviation Administration used for the 737 MAX.

Use Cases and Clinical Acceptability

Is prognosis a good use case to target with AI? It seems like AI algorithms can potentially do better than humans at prognosis.

J.M.: Yes, humans have a lot of cognitive biases that limit our ability to accurately prognosticate, and AI models often can do better. The challenge is finding a specific clinical use case where improvement in accuracy of prognosis is clinically relevant. For instance, in many scenarios the clinician may know that the prognosis is very bad—quantitatively refining exactly how bad that is may not change anything for the clinician or patient. These use cases would not be good for AI; no matter how well the algorithm performs, it will have little or no clinical impact. In other scenarios where key resources are very limited, such as prioritization of patients for organ transplant, refinements in the ability to prognosticate may be highly clinically relevant.

Is it important that human physicians always make the final call, with AI acting in an assisting or advisory role? What about the features an AI algorithm detects that are not perceptible to humans? Do these features become part of the clinical assessment, and how much can they be trusted?

J.M.: At present, the vast majority of commercially available radiology AI acts in a decision support capacity, with radiologists responsible for the final call, but I don't think we should limit ourselves to those use cases. There's a lot of potential in use cases beyond decision support. Once you start to move humans out of the loop, the standards for validation and regulatory approval are, and should be, substantially higher, but I think we should look for these use cases as well and work to meet these high standards of performance and validation. This is particularly important for algorithms that detect human-imperceptible features in imaging. Such algorithms are particularly interesting because they don't just improve efficiency, they expand the capabilities of what we can do with imaging. At the same time, they present increased risk of failure, as it will generally be difficult or impossible for humans to directly verify their outputs, since they're based on imperceptible features.

It's particularly important that clinical testing of these algorithms include a broad range of data (eg, patient demographics, geographic locations, scanner makes and models, practice settings) and that radiologists putting them into practice verify that their patients' characteristics are represented in the training and testing data.

How important are uncertainty, model interpretability, and transparency on clinical implementation? Will physicians trust models that are "black boxes"?

J.M.: This is more about the human psychology of trust than about data science or statistics, so there are multiple factors at play and no single answer that will be universally true across all physicians. More transparency and interpretability are generally better, but the extent to which this is important or necessary depends on the clinical use case. The world of human science, technology, and medicine is sufficiently large that no one person can fully understand all of it. Each of us has to decide which boxes we leave black, which we study and delve into to make clear, and which we compromise on some shade of gray. So, at some point, each of us has to trust black boxes; the challenge is to appropriately earn that trust.

In general, there is likely to be greater acceptance of black boxes in situations where there's strong scientific evidence that the models substantially outperform humans, where the models typically produce outputs that are concordant with human clinical judgment, and where the algorithms are performing a task that humans typically don't or can't perform. On the other hand, algorithms that take over a core diagnostic task currently performed by humans probably need to have substantial explainability that can be consulted when the human user disagrees with the model output, or they won't be trusted.

J.K.C.: This is definitely a question that elicits strong responses on both sides. I think that trust is one component of acceptance by radiologists. One way to build trust is for the AI model outputs to be more explainable. For instance, if you have an algorithm that says that a particular study is positive (eg, for a tumor or a fracture) but provides no more information, such an algorithm may end up slowing down the clinician if they don't locate the finding and have to look through multiple additional images to establish that it was a false positive. On the other hand, if the algorithm provides some sense of localization, they can evaluate it and accept or reject the algorithm's output. However, some posthoc methods such as saliency maps have been shown to have limitations, especially for clinical tasks (3–7). When the goal is explainability, it may be best to start by using inherently more explainable machine learning methods.

Given the expense and difficulty of obtaining large quantities of annotated data, what is the clinical acceptability of data augmentation in AI work?

J.M.: I don't think clinical radiologists generally have any objection to the use of augmentation techniques for training data, and given the typical limitations in dataset size, it's more the rule than the exception. Even for more extreme methods of augmentation, such as generation of synthetic training data with GANs (generative adversarial networks), I think this would generally be viewed as an implementation detail, and the principal clinical concern would be the performance of the model in clinical testing. On the other hand, reporting results of testing against augmented or synthetic data rather than real clinical data is likely to be justifiably viewed with skepticism. But when training data augmentation methods can improve performance on clinical test data without introducing biases, that would be seen as a win by radiologists.

Should AI algorithms be required to outperform radiologists in terms of sensitivity and specificity?

J.M.: This depends entirely on the clinical use case. In some use cases, such as triage and prioritization, where every image will eventually be reviewed by a radiologist, it may be possible to be clinically useful with performance that is less than that of the average radiologist. For instance, my team and I have developed pneumothorax detection algorithms (8) designed for triage use cases where performance below the level of an average radiologist may still provide useful prioritization. For other use cases, such as one where an AI algorithm would identify images that have no findings and don't need review by a radiologist, the requirement for performance would likely be much higher, and the radiologist psychology and perception of reality may be more important than trying to directly compare radiologist and algorithm miss rates. The radiologist will make the decision about whether he or she is willing to use the algorithm and what miss rate he or she is willing to tolerate. That could be a threshold of performance that substantially exceeds what the radiologist is actually capable of. This may not be "fair," but it's part of the reality of developing tools for people to use. Determining the minimum threshold of performance for a use case involves consideration of science, psychology of the physicians and patients, and legal liability—there's no single, simple answer and it requires partnership between multiple experts, each of whom has a piece of the puzzle.

Do you have examples of problems receiving a lot of attention in technologically oriented communities such as MICCAI, but that are clinically not relevant or feasible? Are there things that we should not be working on?

J.M.: In my opinion, it's generally not the broad categories where people go wrong, but the details. There's no particular disease or modality that I think is inaccessible to AI, but it's possible to create technically impressive but clinically useless machine learning models in any area of medicine. I would say that you should not work on projects that have the goal of impacting patient care but don't involve clinical partners. I also want to be clear that clinicians often struggle to fully define a clinical problem and use case on their own. For instance, clinical radiologists often see AI as some form of magic and are frequently interested in using it to address diagnosis of findings that are highly subjective with substantial disagreement between experts. Machine learning experts will recognize that if you can't establish ground truth, it's difficult to be successful. It's the conversation and partnership between machine learning and clinical experts that is key.

What's next for AI in radiology?

A.F.: From my perspective, one of the most exciting potentials of medical imaging machine learning is the derivation of nonvisual features from the pixel data (eg, radiomic features) that either more precisely categorize disease or provide greater accuracy in forecasting prognosis or treatment response. In neuroimaging, for example, multiple investigators have been able to uncover invisible features "hiding" in MRI pixel data that are not perceptible to the expert human observer and have the capability to predict the genomic composition of brain tumors with better accuracy than expert neuroradiologists. Tumor genetics predict the phenotypic expression of the disease and provide the foundation of "precision medicine"—therapies specifically designed for the patient. Developing a capability to map and classify tumor genetics without surgery is a very compelling potential application of machine learning.

J.K.C.: Although publications have largely focused on tasks in diagnostic radiology, there is great potential for machine learning to improve the entire pipeline of patient care. "Smart" scanners that allow us to reduce acquisition times and contrast material/radiation dose without sacrificing image quality, improved workflows, and operations planning are all areas where AI has the ability to substantially improve health care delivery. Secondary use of imaging data for purposes other than the original reason for image acquisition is another area that I am excited about. For instance, when we image a patient with cancer, in addition to information about the tumor, we also have access to, but typically ignore, so much more information about the patient that might impact their outcomes such as low muscle mass, increased visceral fat, or low bone density. The ability to make imaging more quantitative is a great opportunity for large-scale epidemiologic studies in a manner that we have not seen yet due to the qualitative nature of radiologic interpretation that is typically present in today's report.

J.M.: We'll see continuing improvement in AI methods. Most of the off-the-shelf network architectures used today are designed for relatively low-resolution color images that have substantial differences from typically high-resolution grayscale radiologic images. I'm particularly excited about the possibilities of neural architecture search to automatically address some of this mismatch. In neural architecture search, network architectures can be learned rather than just learning weights for a fixed network architecture; this opens the door to network architectures that may be better suited to radiologic images. I also think there is still a lot of low-hanging fruit in application of AI to radiologic data other than images, particularly natural language processing of text. I'm looking forward to development and adoption of standards that will catalyze clinical deployment of AI, just as DICOM did for digital radiology. The real key to advancing AI in radiology is data scientists and clinical radiologists working in concert to identify use cases in the overlap between what's technically feasible and what's clinically important and useful.

Disclosures of Conflicts of Interest: **J.M.** institution received grant from GE Healthcare; author receives royalties from GE Healthcare for licensing of pneumothorax detection algorithm; author paid for development of educational presentations from UCSF Postgraduate Medical Education; author's spouse has been employed by AbbVie and Annexon Biosciences; author is associate editor of *Radiology: Artificial Intelligence*. **J.K.C.** institution received grants from GE, NIH, NSF, and Genentech; author received travel accommodations from IBM; author is deputy editor of *Radiology: Artificial Intelligence*. **A.F.** disclosed no relevant relationships. **M.G.L.** author is consultant to the National Institutes of Health for grant review; author is co-founder of PediaMetrix; author's institution received grants from National Institutes of Health, National Science Foundation, Department of Defense, and Philips Healthcare; author has stock/stock options in PediaMetrix.

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

References

1. Zhu G, Jiang B, Tong L, Xie Y, Zaharchuk G, Wintermark M. Applications of deep learning to neuro-imaging techniques. Front Neurol 2019;10:869.
2 Mongan J, Kohli M. Artificial Intelligence and Human Life: Five Lessons for Radiology from the 737 MAX Disasters. Radiol Artif Intell 2020;2(2):e190111.
3. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. arXiv 2018;(NeurIPS). [preprint] https://arxiv.org/abs/1810.03292. Posted October 8, 2018. Accessed May 2021.
4. Arun NT, Gaw N, Singh P, et al. Assessing the validity of saliency maps for abnormality localization in medical imaging. arXiv 2020. [preprint] https://arxiv.org/abs/2006.00063. Posted May 29, 2020. Accessed May 2021.
5. DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. medRxiv. [preprint] https://www.medrxiv.org/content/10.1101/2020.09.13.20193565v2. Posted October 8, 2020. Accessed May 2021.
6. Saporta A, Gui X, Agrawal A, et al. Deep learning saliency maps do not accurately highlight diagnostically relevant regions for medical image interpretation. medRxiv [preprint] https://www.medrxiv.org/content/10.1101/2021.02.28.21252634v1. Posted March 2, 2021. Accessed DATE.
7. Singh N, Lee K, Coz D, et al. Agreement Between Saliency Maps and Human-Labeled Regions of Interest: Applications to Skin Disease Classification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, June 14–19, 2020. Piscataway, NJ: IEEE, 2020; 3172–3181.
8. Taylor AG, Mielke C, Mongan J. Automated detection of moderate and large pneumothorax on frontal chest X-rays using deep convolutional neural networks: A retrospective study. PLoS Med 2018;15(11):e1002697.