Radiology

Editor's Note: Publication of AI Research in *Radiology*

David A. Bluemke, MD, PhD

Radiology 2018; 289:579–580 • https://doi.org/10.1148/radiol.2018184021 • ©RSNA, 2018

Since the beginning of our discipline, research in radiology has been related to hardware devices. We have been largely concerned with the performance of various imaging modalities (eg, CT, MRI, PET) to detect disease (eg, cancer, cardiovascular disease, trauma). Our research results tell us the sensitivity and specificity of hardware to diagnose disease. Research hardware may be a new type of CT or MRI scanner. Even a new molecular imaging contrast agent operates like hardware: Researchers seek to determine the impact of the contrast agent plus the scanner on diagnostic performance. We expect hardware research to eventually translate to our own imaging practices. If PET detects more metastases than CT for a certain disease, we eventually translate such results into practice guidelines.

However, the nature of imaging research is rapidly changing. Instead of hardware research, leaders in our field are increasingly doing software research—in artificial intelligence, or AI.

In the past 2 years, we have received a rapidly increasing number of research submissions in AI. A prototype, high-quality AI research topic in *Radiology* is by Lakhani and Sundaram from Thomas Jefferson University (1). They studied chest radiograph detection of tuberculosis by AI. They used convolutional neural networks to identify tuberculosis by using a dataset of 1007 radiographs. The authors used a recommended approach (2) for evaluation of the computer algorithm: first training the AI neural network, validated on a separate set of radiographs. The third step is to use independent cases, ideally from different equipment vendors or from another hospital, to test the AI results to show robustness of the AI algorithm. Lakhani and Sundaram took further steps to compare their results to those of radiologists looking at the same radiographs. The best performing computer algorithm had an outstanding receiver area under the curve (AUC) of 0.99 (maximum value, 1.0).

Although elegant, Lakhani and Sundaram have a software result, not a hardware result. In most software research, the only individuals with the algorithm are the researchers. Without the AI algorithm, the results cannot be reproduced. Many AI publications are transient—they are proof-of-concept; they cannot be validated. As a radiologist, you cannot implement the AI research in your clinical practice without the algorithm, and the algorithms are largely discarded. In this setting, there is near zero chance that practice guidelines will be changed.

If this were a "hardware" research topic, the research by Lakhani and Sundaram would stimulate other research groups to try to reproduce the results. After a few papers showing similar results, there would be talks at RSNA in our educational courses, soon followed by high-level society statements. Those statements would likely say that AI algorithms are acceptable for prescreening our radiographs–maybe the AI could even issue final reports in parts of the world with few or no radiologists. Based on the AI algorithm, anti-tuberculosis medications would be started.

As a result, imaging publications in AI may have a fleeting shelf-life of several months. If we were discussing an MRI scanner, this would be unacceptable—hardware needs to last 5–10 years. Algorithms come and go; researchers move on to their next project to detect sarcoidosis rather than tuberculosis. What is our response? Should we stop publishing this research?

New AI research in radiology is amazing. Our discipline has tried for 30 or more years for computers to help us analyze our images. Prior non-AI approaches have mostly not succeeded. In my research lab, technologists and pre- and postdoctoral students analyzed thousands of cardiac MRI cases by drawing circles at the borders of the heart for the last 20 years. Yet in 6 months or less, AI neural networks are now trained to draw those circles better and more consistently than any of our prior efforts. My reaction to seeing new AI developments is equivalent to "shock and awe."

How do I balance shock and awe results from our AI researchers with my assertion that most of the results cannot be reproduced? Recently, the editorial board of *Radiology* has taken some first steps toward improving research standards in this area. To continue my politically inspired analogy, *Radiology* will seek to "trust but verify" AI research in a more robust manner.

Our first policy affecting AI research is regarding preprint servers, such as *arXiv.org.* AI researchers frequently put their latest algorithms on arXiv to claim "I'm first" supremacy. arXiv publications are not peer reviewed. They *do* however look like normal publications—especially to laypersons. Preprint servers are used by AI researchers to rapidly share software, algorithms, and ideas.

However, research placed on preprint servers is generally not accepted by the medical field. Why not? The stakes are higher in medicine: Diagnosing cancer requires more validation than does the code to have Alexa make my grocery list. The best system that we have in medical research is the peer-review process. Despite peer review, we all know that much of our medical research cannot be reproduced. On the other hand, the situation would be much worse without peer review.

1. The policy of *Radiology* is to discourage authors from placing their results on preprint servers. There are two reasons for this. First, if the results are already available, the incremental benefit of publication in *Radiology* is low. Second, the vast majority of submissions for publication undergo substantial changes due to peer review and editorial processes. Reviewers invest considerable time in

improving the publication quality. The RSNA invests money in improving the format of the final manuscript. If an article is on arXiv and then published in *Radiology*, there would be two different papers on the same topic by the same authors—almost certainly with different results and different tables.

2. Our second policy affecting AI research is to *strongly* encourage making the computer algorithms available to other researchers. Authors of AI research should make a git archive of their source code or make it available on the author's web page. Git archive providers such as GitHub, Bitbucket, or Source Forge are already available and in use by some researchers. Authors should place a link to the web page for their code in their Materials and Methods section. They should also provide a unique identifier for the revision of the code used in the publication.

There are two other challenges for reproducing AI research. The first is access to source data. Although the computer code is available, the research results cannot be replicated without the imaging data. The National Institutes of Health (NIH) already requires that researchers make source data available; this needs to be the case for non-NIH research as well. The second deficiency is standardization of AI research in imaging. Lakhani and Sundaram (1) used an independent image dataset to verify their findings. Many others do not, mostly due to lack of standardization in our field. The editorial board for *Radiology* will be addressing publishing standards in AI imaging research in the near future.

I recently heard a thought-leader in our field lecturing that we are now in the era of artificial intelligence 4.0. In that case, I missed the first three stages. For medical imaging, we are just ending an exploratory phase 1 era, in which AI publications are mostly proof of principle without standardization and lacking reproducibility. The above policies hope to ignite phase 2, in which we will achieve both goals of improved standardization and validation of results. What is phase 3 AI? In phase 3, I suspect we will no longer call these developments "AI" at all. Sophisticated AI is built into my iPhone and Alexa device. But to me, they are just iPhones and Alexa devices that learn my shopping habits, driving patterns, and schedule. When AI truly succeeds in medical imaging, we will stop calling it AI. The AI portions will simply be integrated tools in our PACS, scanner, or workstation—not separate features.

References

1. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology 2017;284(2):574–582.
2. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology 2018;286(3):800–809.