ORIGINAL ARTICLE

Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success

SA-CME

James H. Thrall, MD, Xiang Li, PhD, Quanzheng Li, PhD, Cinthia Cruz, MD, Synho Do, PhD, Keith Dreyer, DO, PhD, James Brink, MD

Credits awarded for this enduring activity are designated "SA-CME" by the American Board of Radiology (ABR) and qualify toward fulfilling requirements for Maintenance of Certification (MOC) Part II: Lifelong Learning and Self-assessment. To access the SA-CME activity visit https://3s.acr.org/Presenters/CaseScript/CaseView?CDId=0GSf930sE7s%3d.

Abstract

Worldwide interest in artificial intelligence (AI) applications, including imaging, is high and growing rapidly, fueled by availability of large datasets ("big data"), substantial advances in computing power, and new deep-learning algorithms. Apart from developing new AI methods per se, there are many opportunities and challenges for the imaging community, including the development of a common nomenclature, better ways to share image data, and standards for validating AI program use across different imaging platforms and patient populations. AI surveillance programs may help radiologists prioritize work lists by identifying suspicious or positive cases for early review. AI programs can be used to extract "radiomic" information from images not discernible by visual inspection, potentially increasing the diagnostic and prognostic value derived from image datasets. Predictions have been made that suggest AI will put radiologists out of business. This issue has been overstated, and it is much more likely that radiologists will beneficially incorporate AI methods into their practices. Current limitations in availability of technical expertise and even computing power will be resolved over time and can also be addressed by remote access solutions. Success for AI in imaging will be measured by value created: increased diagnostic certainty, faster turnaround, better outcomes for patients, and better quality of work life for radiologists. AI offers a new and promising set of methods for analyzing image data. Radiologists will explore these new pathways and are likely to play a leading role in medical applications of AI.

Key Words: Artificial intelligence, machine learning, opportunities, challenges, pitfalls

J Am Coll Radiol 2018;15:504-508. Copyright © 2017 American College of Radiology

INTRODUCTION

The invention of the programmable digital computer in the 1940s stimulated mathematicians and philosophers to speculate on the limits of what machines could do. Could machines learn to think? How close could machine capabilities come to those of human beings? A conference at Dartmouth University in 1956 explored these questions and led to the coining of the term *artificial intelligence* (AI) [1]. The race was on.

In the ensuing 60 years, enthusiasm for AI has waxed and waned but has reignited recently with the availability of ever less expense, massively parallel computing systems. The term *deep learning* was added to the AI lexicon to reflect the ability to harness new computing power to develop more powerful AI approaches with more layers of analysis than heretofore possible. The successes of AI programs from IBM (Armonk, New York) in the games of chess and the quiz show *Jeopardy!* (Deep Blue) and from Google (Mountain View, California) in the game of *Go* (DeepMind) [2] were exciting milestones that made people outside of the scientific community aware of AI and its potential.

Department of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts.

Corresponding author and reprints: James H. Thrall, MD, Department of Radiology Massachusetts General Hospital, 25 New Chardon Street, #400, Boston, MA 02109; e-mail: jthrall@mgh.harvard.edu.

Dr Li has an institutional disclosure for the system: "Automatic prescreening method for pneumothorax detection." Dr Thrall has an institutional disclosure on analysis of liver lesions. No patent has been applied for. The other authors have no conflicts of interest related to the material discussed in this article.

1546-1440/17/$36.00 ■ https://doi.org/10.1016/j.jacr.2017.12.026

Major corporations and governments around the world have embraced AI technology as one of the important strategies for dealing with the enormous amounts of digital data being generated in the information age—the age of "big data." AI is also on the doorstep of medical practice. The trickle of publications now appearing in journals will soon turn into a flood.

The radiology community has played a leading role in propelling medicine into its digital age and now has the opportunity to become a leader in exploring medical applications of AI. The tens of millions of radiology reports and billions of images now archived in digital form exemplify the concept of "big data" and constitute the required substrate for AI research.

The fundamental question is whether AI applications in radiology can add value. Adding value includes the discovery of new knowledge and extraction of more and better information from imaging examinations to achieve better outcomes for patients at lower cost. For radiologists, adding value includes establishment of more efficient work processes and improved job satisfaction.

The goal of this perspective is to help create a framework—apart from a discussion of AI technology per se—for developing strategies to explore the potential of AI in radiology and to identify a number of scientific, cultural, educational, and ethical issues that need to be addressed.

OPPORTUNITIES

Two areas of opportunity that can help provide a framework for approaching AI in imaging deserve discussion: the desirability of establishing standards and infrastructure and the opportunity to establish a categorical model for approaching the spectrum of clinical and research applications of AI to help identify and understand their respective value propositions.

Standards and Infrastructure

AI imaging research would benefit from the establishment of (1) national and international image sharing networks, (2) reference datasets of proven cases against which AI programs can be tested and compared, (3) criteria for standardization and optimization of imaging protocols for use in AI applications, and (4) a common lexicon for describing and reporting AI applications.

Access to large numbers of proven cases is necessary to test and validate AI programs and, for many applications, to train them. Taking face recognition as an example, the DeepFace system used 4.4 million labeled faces from 4,030 people collected by Facebook (Menlo Park, California) as training data. Its accuracy consistently approaches human-level performance [3]. Thus, for the purpose of analyzing medical images, even for relatively common conditions, it would be advantageous to be able to aggregate material from multiple institutions. The technical basis for accomplishing this already exists through networks created by the RSNA and the ACR as well as work such as The Cancer Imaging Archive at the National Cancer Institute to establish systems for image sharing [4]. The existence of these networks represents a unique opportunity to create shared image repositories for AI research. The new ACR Data Science Institute could serve as a convening organization.

As a corollary to the foregoing, robust methods are needed for quality control of shared images and ensuring the integrity of image data. Standards still need to be developed that address curation of images. If image data are corrupted in transmission or storage or during processing, it will be difficult to duplicate work or confirm its validity. Although several scientific informatics systems to support data transfer, storage, quality control, and query have been applied across institutions, such as those provided by the National Center for Biotechnology Information [4], the complete needs of such a shared informatics system are not yet defined for radiology.

An image sharing network would support and facilitate use of reference datasets of proven cases on which to test and compare new AI programs for accuracy and other measures of performance, such as required processing power. Such datasets also could be used in optimizing programs and selecting the best programs for further development and clinical use. Ideally, a sufficient number of reference datasets would be established in each application area to reflect the demographics of different patient populations. A database conceptually similar to the "ImageNet" [5], which has been used as the benchmark for object recognition tasks, would be valuable for advancing AI in radiology.

The high variability in imaging protocols between institutions and even variability in the execution of a given protocol within an institution are potential impediments to development and use of AI applications in imaging [6]. For example, subjective analysis of a CT scan may be somewhat tolerant of the timing of contrast material administration, and variable timing can create an "apples-to-oranges" problem for AI programs that rely on quantitative factors and may require scaling or a background subtraction step. The impact of protocol variability needs to be more fully studied.

The development of a common lexicon is a corollary to this concept to help improve communication and understanding between users. "AI-RADS" would fit in with BI-RADS and LI-RADS.

Categories of AI Applications

The principle focus of much initial clinical AI research in imaging has been on improving diagnostic accuracy, the importance of which was recently emphasized in the Institute of Medicine report entitled "Improving Diagnosis in Health Care." [7] A number of encouraging results have been reported in AI-enabled computer-aided diagnosis [8-14] with two achieving human-level performance [12,14].

Beyond the core goal of better diagnostic accuracy [15], AI methods can be used to address a number of practical issues faced in daily practice that are worthy of consideration: (1) optimization of work lists to prioritize cases, (2) preanalysis of cases in high-volume applications where observer fatigue may be a factor, (3) extracting information from images that is not apparent to the naked eye, and (4) improving the quality of reconstructed images.

Timeliness in the diagnostic process is important, especially when urgent or life-threatening conditions are present. Long unread case queues in busy departments can delay diagnosis and are particularly problematic if clinical presentation is not typical. One can imagine a suite of AI programs optimized for sensitivity, rather than overall accuracy, and aimed at detecting pulmonary emboli, pneumothoraces, strokes, free air in the abdomen, and other immediate life-threatening conditions. These programs could run automatically and immediately as soon as examinations were completed, alerting the radiologist to prioritize the case for immediate attention. The likelihood of earlier identification of important findings would be increased. Any associated false-positive cases need to be interpreted at some point anyway.

Observer fatigue is an unavoidable aspect of radiology practice and a particular issue in screening examinations where the likelihood of finding a true positive is low. In this setting, an AI program optimized for very high negative predictive value could identify an "enriched" subset of cases likely to harbor any true-positive cases for early review.

AI applications offer an important new way to extract heretofore unavailable information from images and are a new portal for imaging contributions to the era of precision medicine. The age of big data and deep learning has spawned the concept of "radiomics" wherein hundreds of abstract mathematical features of images can be defined or detected and through AI programs correlated with other data on genomics or response to therapy. These radiomic features are not discernible to human observers and represent a kind of *mathematical imaging phenotype* of disease expression [16]. Early work in establishing radiomic "signatures" predictive of response to therapy is promising. In the broad context of precision medicine, the complete imaging phenotype may be thought of as encompassing both imaging manifestations of disease detected by conventional means as well as identifying new features in which AI will play the key role [16].

CHALLENGES

The challenges in exploiting the potential of AI in medicine may be thought of as circumstantial, relating to human societal behaviors, and intrinsic, relating to the capabilities of the underlying science and technology. None of the challenges alone will be a showstopper but all may slow progress and need to be addressed.

Circumstantial Challenges

Radiology, as a specialty, has had to deal with rapid technological change as much as any other discipline in medicine. Radiologists have benefitted greatly from working with digital systems, but concerns exist about machines taking jobs away from humans, reflecting a possible, even likely, cultural barrier to adoption of AI in radiology. Stoking these concerns, Obermeyer and Emanuel [17], writing in the *New England Journal of Medicine*, have already predicted that "machine learning will displace much of the work of radiologists and anatomic pathologists" and "machine accuracy will soon exceed that of humans." Chockley and Emanuel [18], writing in the *JACR*, comment that "machine learning will become a powerful force in the next 5-10 years and could end radiology as a thriving specialty."

Biology is far more complex than chess or *Jeopardy!* or *Go*, and the foregoing predictions go far beyond what has been accomplished with AI in imaging to date. Medical images are highly heterogeneous at both an individual and a population level. Moreover, for a given application, if the number of available labeled images is limited, it will be challenging to train AI systems [19], and there is a risk of "overfitting" the data with loss of generalizability [6,13]. Rather than replacing radiologists, the much more likely prospect is that practice of radiology will beneficially assimilate AI methods, improving quality and efficiency similar to the experience with other digital imaging methods.

The pool of investigators trained in radiology AI methods is limited. This can be addressed by recruiting scientists with backgrounds in AI into radiology and through the educational programs already being undertaken by radiology professional societies. Historically, once an area is recognized as important, capable people quickly populate it, so this is not likely to be a long-term issue. Practicing radiologists will need to learn about AI but will not need to become experts in AI research or design of AI programs to beneficially use AI-based results.

Upfront costs for AI—investigator salaries and hardware and software—are substantial. Forward-looking departments and institutions will bear these costs to support researchers and create AI core laboratories that others can access much as other core laboratories are run.

The FDA is likely to play an important role in approving AI programs for clinical use, but it is unclear what the process will be for validating AI programs or whether and how credentialing of individuals in their use will happen. It is also unclear how the "black box" nature of AI programs will affect liability. A legal issue within institutions will be the need to define ownership of data and to determine who has the right to use it.

Intrinsic Challenges

Among the intrinsic challenges for AI in imaging are (1) how best to establish the source of truth for validating results, (2) whether processing speeds will be fast enough for relevance to clinical practice, (3) whether protocol-tolerant AI programs can be developed, and (4) whether criteria can be established for determining in what patient population or populations a given program is valid.

With artificial intelligence programs that are trained on known or proven cases ("supervised" learning [6]), a robust "source of truth" for each diagnosis is required and is always required for validation, whether learning is supervised or unsupervised. The source of truth can come from patient outcomes or results of other "gold standard" testing methods apart from the imaging method being studied, but the source of truth used must be rigorous and should be explicitly stated for each AI program that is developed and used clinically.

At present, computing systems fast enough to supply results in a clinically relevant time frame for emergency or urgent diagnoses are not generally available in medical institutions. However, this is not likely to be a practical problem going forward because of rapid development of lower-cost graphics processing unit–based computing systems and easy access to cloud computing solutions.

A generic challenge in clinical research is trying to generalize results beyond the patient population in which the research was performed. This is often addressed in drug trials by doing research in a large, diverse, inclusive population. However, as more has been understood about precision medicine, the opposite approach is turning out to be better in many cases, that is, to perform research in smaller, more homogeneous, and better-defined populations, where the results apply more reliably to the patients in that specific population. This approach is used in clinical trials of molecularly targeted drugs. It is not clear how this will play out in imaging applications of AI, but the issue of patient population must be considered with careful definition of population being studied. The trade-off will be between more robust generalizable programs with lower accuracy versus narrow solutions for well-defined patient populations that are less robust [6].

PITFALLS

AI programs typically require substantial numbers of cases for training. Institutional xenophobia and other proprietary interests may restrict access to image data between institutions. Failure to assemble a sufficiently large enough training set is a potential pitfall that could have the effect of making the results less accurate or generalizable [19]. The risk of overfitting was noted previously [6,13].

The tolerance of using AI programs in imaging between different patient populations is not yet known. Failure to recognize that a program is not generalizable, for example, from adults to children or between different ethnic groups, could lead to incorrect results. Among other issues, organ sizes and prevalence of disease vary widely between different patient populations.

Likewise, the tolerance or latitude for variations in image acquisition protocols is unknown for AI programs [6]. Applying AI programs on image data from protocols other than the one used for program training may introduce errors. Simplistically, IBM's Deep Blue program for *Jeopardy!* was trained on questions in the English language and would not have worked without further training had it been given questions in a different language.

The biggest limitation for AI in imaging may be inherent limitations in defining normal versus abnormal in continuously variable biologic data. Ranges for normal are set as a certain number of standard deviations from the mean of a supposedly "normal" population. This means for any test or measurement, a given percentage of truly normal people will have "abnormal" results. Investigators in AI will face this conundrum where nominal criteria for normal versus abnormal can be difficult to define when, for example, setting limits for organ sizes.

CRITERIA FOR SUCCESS

The most important criteria for success follow directly from the discussion of opportunities. The name of the game is to create value in the delivery of medical care and delivery of radiology services—increased diagnostic certainty, decreased time on task for radiologists, faster availability of results, and reduced costs of care with better outcomes for patients [15]. As with any new technology, substantial time and experience will be required to establish whether these benefits apply and understand their relative magnitudes.

Achieving these outcomes will be enhanced if AI programs can be developed that are tolerant of different data acquisition protocols and work in diverse patient populations. Absent that, success will require full understanding of the circumstances under which a given program is valid.

CONCLUSION

It is not yet clear what the full or final role of AI methods will be in imaging or their impact on radiologists. What is clear is that AI provides a promising new set of tools for interrogating image data that should be explored with vigor. The growing interest in AI in the imaging community bodes well for its potential leadership role [6].

TAKE-HOME POINTS

- Worldwide interest in AI applications, including imaging, is high and growing rapidly.
- The large amount of image and report data now in digital form ("big data") provides a substrate for development of AI applications.
- Development of AI applications in imaging would benefit from the development of standards and infrastructure—acquisition protocols, validation criteria, lexicon for communication.

REFERENCES

1. The Dartmouth Artificial Intelligence Conference. The next fifty years. Available at: https//www.dartmouth.edu/∼ai50/homepage.html. Accessed June 13, 2017.
2. McAfee A, Brynolfsson E. Where computers can defeat humans and where they can't. Available at https://www.nytimes.com/2016/03/16/opinion/where-computers-defeat-humans-and-where-they-cant.html. Accessed June 12, 2107.
3. Taigman Y, Yang M, Ranzato M, Wolf L. 2014. DeepFace: closing the gap to human-level performance in face verification. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society; 2014:1701-8.
4. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 2016;26:1045-57.
5. Russakovsky O, Deng J, Su H, et al. 2015. ImageNet large scale visual recognition challenge. Int J Comput Vis 2015;115:211-52.
6. Ravi D, Wong C, Deligianni F, et al. Deep learning for health informatics. IEEE J Biomed Health Inform 2017;21:4-21.
7. Balogh EP, Miller BT, Ball Jr eds. Improving diagnosis in health care. Washington DC, USA: National Academies Press; 2015.
8. Anthimopoulos M, Christodoulidis S, Ebner L, Christe A, Mougiakakou S. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. IEEE Trans Med Imaging 2016;35:1207-16.
9. Cireşan DC, Giusti A, Gambardella LM, Schmidhuber J. Mitosis detection in breast cancer histology images with deep neural networks. In: Mori K Sakuma I, Sato Y, Barillot C, Navab N, eds. Medical image computing and computer-assisted intervention—MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II. Berlin, Germany: Springer; 2013: 411-8.
10. Shen D, Wu G, Suk HI. Deep learning in medical image analysis. Annual Review of Biomedical Engineering:19. Available at:. https://doi.org/10.1146/annurev-bioeng-071516-044442. Published March 9, 2017. Accessed June 13, 2017.
11. Shen W, Zhou M, Yang F, Yang C, Tian J. Multi-scale convolutional neural networks for lung nodule classification. In: Ourselin S, Alexander DC, Westin CF, Cardoso MJ, eds. Information processing in medical imaging: 24th International Conference, IPMI 2015, Sabhal Mor Ostaig, Isle of Skye, UK, June 28–July 3, 2015, Proceedings. Cham, Switzerland: Springer International; 2015:588-99.
12. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115-8.
13. Setio AAA, Ciompi F, Litjens G, et al. Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. IEEE Trans Med Imaging 2016;35:1160-9.
14. Lee H, Tajmir S, Lee J, et al. Fully automated deep learning system for bone age assessment. J Digit Imaging 2017;30:427-41.
15. Fryback DG, Johnbury JR. The efficacy of diagnostic imaging. Med Decis Making 1991;11:88-94.
16. Aerts HJWL. The potential of radiomic-based phenotyping in precision medicine: a review. JAMA Oncol 2016;2:1636-42.
17. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. N Engl J Med 2016;375:1216-9.
18. Chockley K, Emanuel E. The end of radiology? Three threats to the future practice of radiology. J Am Coll Radiol 2016;13:1415-20.
19. Cho J, Lee K, Shin E, et al. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? arXiv preprint arXiv 2015. 1511.06348. Available at: https://arxiv.org/abs/1511.06348. Accessed January 18, 2018.