COMMENTARY

Artificial Intelligence in Radiology: A Call for Thoughtful Application

C. Noelle Driver[1,2], Bradley S. Bowles[1], Brian J. Bartholmai[3] and Alexandra J. Greenberg-Worisek[4,*]

Artificial intelligence (AI) is the most recent development in a long series of disruptive technological innovations in radiology. Medical imaging began in the late 1800s after the discovery of the X-ray, but exploded in the late 1900s with the availability of computers to create, analyze, and store digital images. Future speculation about radiology includes widespread AI involvement; however, thus far, translation of AI to clinical radiology has been limited.

HISTORY OF COMPUTERIZED RADIOLOGY

The historical context for artificial intelligence (AI) applications in radiology falls at the intersection of radiographic methods and modern computers (**Figure 1**). Radiology uses radiation to image the insides of bodies for diagnosing and treating disease, and includes X-rays, tomography, fluoroscopy, and many other techniques. Radiography was a relatively slow, analog process until the development of microprocessor chips in the 1970s, which enabled the creation of small-scale computerized devices. The impact on radiology was immediately apparent, as the first computerized tomography (CT) and magnetic resonance imaging (MRI) prototypes were developed soon after. While microcomputer technologies evolved into personal computers, industry interest and involvement in CT and MRI rapidly advanced hardware and imaging techniques and reduced scan times. Early images suffered from poor resolution and artifacts of acquisition, but refinements in software processing over time greatly improved image quality, and, consequently, clinical utility.

It is possible that developing AI technology will be integrated with radiology in a way that similarly enhances clinical utility; however, it is necessary to consider what clinical problems the technology is appropriate for and apply it selectively to ensure successful translation. Diagnostic imaging revolutionized modern clinical practice and research, but such rapid growth also introduced various challenges, including low-value utilization, higher-cost imaging services, and high volumes; this creates a burden radiologists who aim to read images both quickly and accurately. Although AI was first applied in radiology to detect microcalcifications in mammography in 1992, it has gained much more attention recently. We propose that thoughtful application of AI tools to narrow tasks that address important challenges in radiology is key for successful translation. AI will be most effectively disruptive to radiology in areas where unmet needs exist and technology can relieve challenges instead of imposing new ones.

CHALLENGES IN RADIOLOGY

One challenge in the field of radiology is related to high utilization and cost. Increasing clinical utility of rapid imaging services from the 1970s onward brought increasing clinical utilization as well; this, in turn, led to increasing of total costs. In the United States, the number of CT scans per year quadrupled from about 15 million to nearly 90 million from 1990 to 2010;[1] other modalities have followed similar trends. Although imaging is only a small portion of the Medicare budget, the spotlight on low-value imaging is, in part, because images are one of the most costly medical tests. In a 2012 study of Medicare spending, laboratory tests accounted for 9% of services and 4% of payments, whereas imaging accounted for 11% of services and 17% of payments.[2] Several legislative efforts have aimed to curb imaging utilization and, after decades of exponential growth, there is evidence that utilization slowed in the late 2000s.[3] As reimbursement models in America continue to incentivize utilization over value,[4] efforts are ongoing to reduce low-value use. We hypothesize that algorithms that align with the priorities of reducing costs and curbing utilization or unnecessary follow-up have a greater chance of successful translation.

Second, volume burden to radiologists increased in tandem with utilization. In 2015, it was estimated that the average radiologist interprets one CT or MRI image every 3–4 seconds in an 8-hour workday to satisfy demand.[5] With technological advances in image resolution, radiologists scan more images per study. These increasing numbers of images requiring interpretation means that turnaround time has greatly increased. Improving turnaround time is of great interest to the field, and was key to successful translation of several AI-based algorithms detailed below.

Increasing utilization, cost, and volume, coupled with the need for reduced turnaround time, are directly related to the third challenge we believe AI may be able to assist: the essential components of reliability and accuracy. When

[1]Graduate School of Biomedical Sciences, Mayo Clinic, Rochester, Minnesota, USA; [2]Alix School of Medicine, Mayo Clinic, Rochester, Minnesota, USA; [3]Department of Radiology, Mayo Clinic, Rochester, Minnesota, USA; [4]Center for Clinical and Translational Sciences, Mayo Clinic, Rochester, Minnesota, USA. *Correspondence: Alexandra J. Greenberg-Worisek (Greenberg.alexandra@mayo.edu)
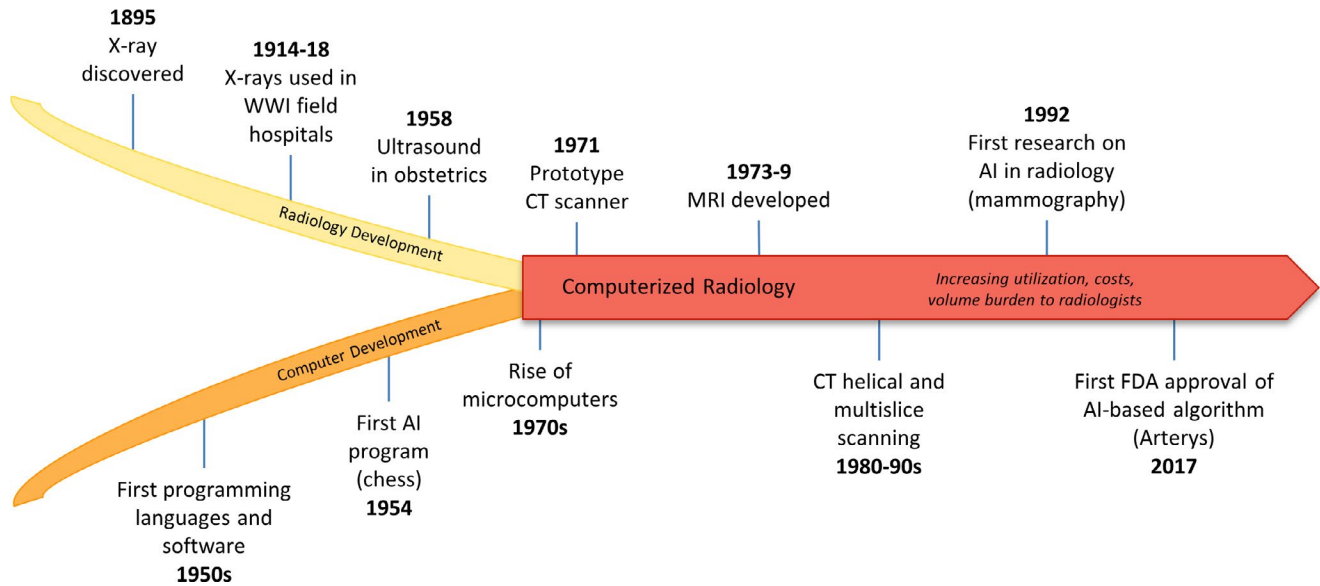
Figure 1 Historical context for artificial intelligence in radiology. The history of computerized radiology falls at the developmental intersection of radiography and modern computers. Once the field transitioned from analog to digital acquisition, advances in imaging techniques led to rapid growth in utilization, volume, and costs. AI represents a pivotal development that will continue to drive the evolution of this field. AI, artificial intelligence; FDA, US Food and Drug Administration; CT, computed tomography; MRI, magnetic resonance imaging; WWI, World War I.

interpreting quickly, radiologists use more heuristics, subject to error due to a number of cognitive biases that have been detailed elsewhere.[6] Errors caused by such cognitive biases increase rates of interobserver and intraobserver variation, even among trained specialists; one study reports discrepancy rates up to 32% in the identification and interpretation of findings on abdominal and pelvic CTs,[7] and another study found interobserver lung nodule screening disagreement up to 42%.[8]

AI tools are not immune to bias depending on the data sets used to train them. Once deployed, algorithms may change and improve with exposure to more cases. Thus, AI tools have the potential to automate specific detection and classification tasks both faster and more reliably than humans; however, this ability depends heavily on how algorithms are developed and trained.

CHALLENGES IN AI ALGORITHM DEVELOPMENT

The US Food and Drug Administration (FDA) has recently suggested several components of "good machine learning practice" (GMLP) for consideration when developing AI tools.[9] Although these points are not official FDA regulation or even draft guidance, they are useful recommendations for developers. The first two recommendations concern access to high-quality and focused data sets, calling for use of clinically relevant data "acquired in a consistent, clinically relevant and generalizable manner that aligns with… intended use and modification plans." In short, data sets must be carefully selected and curated, and have direct relevance to the clinical problem at hand. We note that creating quality data sets is an opportunity for clinicians and investigators wishing to become involved in AI research who otherwise lack the necessary programming expertise.

The AI tool must also be well-trained. GMLP point 3 states that there must be "appropriate separation between training, tuning, and test datasets" to ensure rigorous quality control of the final AI product. The fourth and last recommended component is to maintain an "appropriate level of transparency of output and of the algorithm aimed at users." In short, regulators and developers must be able to monitor the algorithm's behavior. AI tools are inherently difficult to regulate, particularly for those that are continuously learning; the easier it is for regulators and developers to assess the relevance of the tool's output to the clinical problem, the easier AI tools will progress through approval processes. Following GMLP standards and demonstrating stringent quality control will enhance clinical and analytical validity and increase successful advancement to clinical implementation.

THE EVOLVING REGULATORY FRAMEWORK: A MOVING TARGET

Although several companies have received FDA approval through traditional De Novo and 510(k) medical device pathways, the FDA has recognized the inadequacy of these pathways for software as a medical device (SaMD). SaMD has different considerations than traditional hardware-based medical devices with regard to patient safety. The software is deployed to different operating systems in different institutions, increasing the chances of malfunction and challenges with interoperability. Developers need to be able to roll out updates efficiently across all software to all licensed users. Finally, data privacy must be maintained. The FDA is piloting an alternative pathway called Pre-Cert in 2019, which will likely lay the foundation for how SaMD products will be reviewed in the future. The Pre-Cert program is intended to cultivate relationships with

trusted manufacturers and allow them to quickly implement software updates and monitor software learning without continuous FDA oversight.

LESSONS LEARNED FROM SUCCESSFUL AND UNSUCCESSFUL AI APPLICATIONS

Although many AI-based algorithms are being developed, a few dozen have been approved by the FDA thus far with purposes ranging from wrist fracture to diabetic retinopathy detection. Three companies, Viz.AI, Aidoc, and MaxQ AI, have received FDA clearance for triaging support tools for potentially life-threatening diagnoses. Viz.AI developed an algorithm that identifies a stroke in angiograms and alerts the on-call physician and was approved in 2018 under the De Novo pathway. Aidoc developed deep-learning algorithms to assess CT scans for a number of acute pathologies, including intracranial hemorrhage (approved in 2018 as via 510(k) citing Viz.AI as predicate), pulmonary embolism (2019), and spine fractures (2019). Aidoc reviews a radiologist's imaging queue and flags those with the suspected pathology for further review. Similarly, MaxQ AI received approval in 2018 for a similar triaging algorithm for intracranial hemorrhage, citing Aidoc's tool as a predicate device.

Conversely, IBM's Watson Health was an unsuccessful attempt to disrupt a whole field: oncology. IBM partnered with Memorial Sloan Kettering to create a machine learning algorithm for cancer treatment. In development stages, it performed phenomenally; however, when brought to clinic, it gave erroneous recommendations. Watson's attempt at overhauling such a complicated and vast management challenge (cancer treatment) was too broad a focus for the state of machine learning. We are still far from the ability to capture all the nuances of real-world patient care with our machines—AI-ready tasks are instead more narrow challenges that address some of the key value metrics in radiology, such as utilization and turnaround time.

Many view AI as a panacea for all challenges in radiology (and, more broadly, medicine), and the debate around AI "replacing" radiologists is ongoing. However, as detailed above, few AI-based algorithms are currently approved by the FDA for use in clinical practice. This is a common issue in clinical translation: out of excitement about a "hot" technology, the technology is applied broadly to see what it solves, rather than identifying a need and looking backward to determine if the technology would be an appropriate solution. We propose that successful clinical translation of products using emerging technologies like AI depends on the application of the technology to a carefully selected, narrow clinical challenge. The creators of Aidoc, for example, identified a specific problem, developed an appropriate tool that addressed a key value metric, and seamlessly integrated that tool into the current workflow. Other FDA-approved algorithms have had a similarly narrow focus, contributing to demonstrable analytical and clinical validity and utility. AI will only do what it is trained to do, so clinical translation, dissemination, and implementation will only be successful with very specific indications and uses, akin to the specific indications given for pharmaceutical approval. When it comes to products that use new technologies, attempts to change the world with one tool may fall short, but a series of small advances applied thoughtfully to the right challenges can eventually change the world.

Conflicts of Interest. The authors declared no competing interests for this work.

1. Brenner, D.J. What do we really know about cancer risks at doses pertinent to CT scans? Communicating our Value, Improving our Future, The American Association of Physicists in Medicine Conference, July 31 August 4, 2016, Washington, DC.
2. Medicare Payment Advisory Commission. Changes in service use consistent with reports of decreases outside of Medicare: Small imaging decrease after decade of rapid growth. In Report to the Congress: Medicare Payment Policy. Washington DC: MedPAC, March 2012, p. 101–103.
3. Levin, D.C. *et al*. Bending the curve: The recent marked slowdown in growth of noninvasive diagnostic imaging. *Am. J. Roentgenol.* **196**, W25–W29 (2011).
4. Committee on Quality Health Care in America, Institute of Medicine. Crossing the Quality Chasm: a new health system for the 21st century. (National Academy Press, Washington, D.C., 2001).
5. McDonald, R.J. *et al*. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Acad. Radiol.* **22**, 1191–1198 (2015).
6. Busby, L.P. *et al*. Bias in radiology: the how and why of misses and misinterpretations. *RadioGraphics* **38**, 236–247 (2018).
7. Abujudeh, H.H. *et al*. Abdominal and pelvic computed tomography (CT) interpretation: discrepancy rates among experienced radiologists. *Eur. Radiol.* **20**, 1952–1957 (2010).
8. Gierada, D.S. *et al*. Lung cancer: interobserver agreement on interpretation of pulmonary findings at low-dose CT screening. *Radiology* **246**, 265–272 (2008).
9. US Food and Drug Administration. Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD): Discussion Paper and Request for Feedback. Retrieved from June 4, 2019, <https://www.fda.gov/media/122535/download>.