Published in final edited form as:
J Am Coll Radiol. 2019 September ; 16(9 Pt B): 1309–1317. doi:10.1016/j.jacr.2019.05.036.

Artificial Intelligence in Imaging: The Radiologist's Role

Daniel L. Rubin, MD, MS[1]

[1]Department of Biomedical Data Science, Radiology, and Medicine (Biomedical Informatics Research), Stanford University

Abstract

Rapid technological advancements in artificial intelligence (AI) methods have fueled explosive growth in decision tools being marketed by a rapidly growing number of companies. AI developments are being driven largely by computer scientists, informaticians, engineers, and business people, with much less direct participation by radiologists. Participation by radiologists in AI is largely restricted to educational efforts to familiarize them with the tools and promising results, but techniques to help them decide which AI tools should be used in their practices and to how to quantify their value is not being addressed. This paper discusses the role of radiologists in imaging AI and suggests specific ways they can be engaged by (1) considering the clinical need for AI tools in specific clinical use cases, (2) by undertaking formal evaluation of AI tools they are considering adopting in their practices, and (3) by maintaining their expertise and guarding against the pitfalls of over-reliance on technology.

Sentence summary

This paper describes the role of radiologists in the AI era, suggesting specific ways for them to be engaged as educated consumers and critical evaluators these technologies to ensure the benefits of these tools outweigh a number of potential risks.

Keywords

Artificial Intelligence; Radiology; Imaging; Evaluation

INTRODUCTION: CLINICAL MOTIVATION AND AI TO THE RESCUE

There is tremendous excitement about the potential of applying AI methods—particularly deep learning— to radiology images. Near-human level of performance has been quickly achieved in the ImageNet database, beginning with AlexNet [1] and soon followed by others

Corresponding Author: Daniel Rubin, Medical School Office Building (MSOB) 1265 Welch Road, X335 Stanford, CA 94305-5464, dlrubin@stanford.edu650-427-0025, Twitter: rubinqilab.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

DISCLOSURES
No disclosures to report.

Conflict of interest information: Dr. Rubin has a patent Methods for evaluation pending.

[2–6],. Many papers are appearing about the use of AI for assisting image interpretation [7–10], automating other imaging tasks, such as image enhancement [11, 12], object segmentation [13–18], automated exam protocolling [19], detection of critical findings and worklist prioritization [20, 21], and even clinical prediction [22–29].

The exuberance over the potential of AI in radiology is well founded in terms of clinical need. The rapid rise in the number of images, coupled with increased utilization of cutting edge imaging technologies is putting increasing pressure in radiologists, challenging their ability to deliver optimal care, and physician burnout is an important problem in radiology [30]. In addition, radiologists vary in their ability to recognize and interpret image features [31]. AI tools could potentially reduce variation in practice [32–35] and automate detection of imaging abnormalities by focusing attention on studies and images that are most critical [35–37].

A TIDAL WAVE OF AI APPLICATIONS

Many AI algorithms to tackle the aforementioned challenges are appearing at an explosive pace. Table 1 shows a list of several major current types of AI applications being developed. At RSNA 2019 there were over 200 companies that highlighted AI products in development, and several AI products from a few AI companies have already achieved Food and Drug Administration (FDA) clearance [38]. To deal with the accelerating pace of AI tools seeking clearance, FDA recently released its first-ever guidance on developing a streamlined and timely approval of AI products [39, 40].

The zeal to develop and market AI algorithms is reminiscent to some authors of the California Gold Rush [41]—a suitable analogy given that many of the AI companies are in the Silicon Valley. In the Gold Rush, the value of the gold depended on the market for banks buying gold from the miners. In the AI era, the value of these algorithms will depend on the market for radiologists that decide to purchase them. So as radiologists are confronted with an onslaught of AI algorithms, how are they to decide whether to use any of them and which to use?

Becoming educated consumers of AI algorithms is the role of the radiologist in the AI era. While the current focus and hype is on new AI applications and the data used to train them (the “new oil” of the current era [42]), there is little focus on the consumer’s perspective of these products—the radiologist and patients in whom these methods are used. Our interest in this paper is thus on the radiologist and their patients. We focus however on radiologists, and we presume that the benefits to their practice translate into patient benefits through better image workflow, interpretation, and decision making.

THE ROLE OF THE RADIOLOGIST IN IMAGING AI

As imaging AI products are developed and marketed, the role of the radiologist is to be an educated consumer about these tools. Being such an educated consumer requires the radiologist to (1) consider the clinical need for AI tools in specific clinical use cases, (2) undertake formal evaluation of AI tools before adopting them in practice, and (3) maintain his or her clinical Radiology expertise and guard against the pitfalls of over-reliance on technology.

(1) Clinical needs:

There are many AI products becoming available, and how is the radiologist to choose? Radiologists should first consider whether an AI application is likely to materially benefit their practice. For example, a chest radiologist may find no value in an AI algorithm that detects pneumothorax, however, an algorithm that detects and reports volumetric changes in the size of the pneumothorax or detects subtle features indicating tension could be useful. Clinical scenarios for which AI algorithms are considered to be potentially valuable are referred to as "AI use cases."

The American College of Radiology (ACR) Data Science Institute (DSI) has been developing a comprehensive catalog of AI use case documents [43]. Each document provides a narrative description of a clinical need being addressed by the AI algorithm, technical details about the expected inputs needed by the AI algorithm and the outputs it produces for the radiologist (Figure 1). The development of AI use cases by the ACR DSI is a community-based effort, and this is a good starting point for radiologists to consider which clinical needs benefit from AI.

(2) Evaluation of AI tools:

The ability to quickly produce AI tools is outpacing the thoroughness with which they are being evaluated or validated on independent data in new settings. In a recent review of 516 published papers on AI tools, only 6% (31 studies) performed external validation, and none had design features that are recommended for robust validation of clinical performance [44]. Although vendor products undergoing FDA review for regulatory clearance undergo more thorough validation, the generalizability of these products in clinical practice is not generally performed (it could be considered part of post-marketing surveillance). Most vendors offer the possibility of allowing radiology practices to try out their AI products before purchasing them. This "try before you buy" paradigm is not a robust approach to assessing whether an AI algorithm will be beneficial to a practice, since it is based on anecdotal experience and the cases used to train the algorithm may not reflect the actual mix of patients seen by that practice. Some radiologists may be tempted to assume that such data collection and evaluation of metrics is not needed, presuming the AI will work well in their patients if the algorithm has been FDA cleared. This is not necessarily a good assumption because the datasets used for FDA clearance may not be sufficiently representative of every radiology practice. This inability of AI algorithms to "generalize" (work well on new, unseen data) is a known potential weakness of all AI methods. For example, a recent study that evaluated different combinations of large collections of data for training and testing an AI algorithm (both single institution and multi-institution) to detect pneumonia for testing found that the performance of the AI algorithm varied substantially [45], and decrease in performance of AI algorithms on independently collected data has been observed by other workers as well [46].

Failure of AI algorithms to generalize well to new data arises because everything that AI algorithms that are trained solely on data (deep learning) "know" is based on the data that were used to train them. If the training data do not include certain types of cases that a radiology practice may encounter (e.g., different diseases, different image types, artifacts, etc), then the algorithm may provide unexpected results. Bias in training data is a common cause of AI algorithms to fail to generalize, e.g., due to differences in patient populations, types of equipment, imaging parameters used, lack of representation of rare diseases. For example, facial recognition algorithms trained on Caucasians were not deployed by police departments due to their failure to accurately characterize the faces of persons of color [47]. Consequently, it is imperative for radiologists to evaluate the performance of AI algorithms on data in their local practice. If the results of such evaluations are provided to regulatory agencies, it can provide the basis for post-marketing surveillance of AI.

Table 2 shows a list of the steps needed for the radiologist to evaluate an AI algorithm on their local institutional data. The first step is to determine which output(s) from the AI algorithm are important to the clinical use case. For example, if the radiologist is interested in an AI tool to help with pneumothorax detection, she should first decide which information from the AI will be important to her practice (e.g., is a simple yes/no answer for pneumothorax sufficient? Is it important to provide the size of the pneumothorax?) For this step, reviewing the outputs specified in the ACR DSI use cases can be helpful (Figure 1). The radiologist should also consider other outputs that may be important to her besides those listed; for example, it may be that the radiologist desires to have help determining quantitative change in the size of the pneumothorax, and that is currently not listed in the ACR DSI use case, and the radiologist may wish to request that feature from the AI vendor.

The second step in AI evaluation is to collect representative patient cases that will be used to test the AI algorithm ("test cases"). It is critical that these test cases reflect the actual patient population seen by a radiology practice in order to adequately address ability of the AI to generalize, as described earlier. This could be done by taking a random sample of cases from the medical record and then looking at the cases to make sure that the case mix appears reasonable. As a diagnostic check, it would be helpful to compute the frequency of each diagnosis in the test case to determine how well that lines up with the expected disease prevalence in the radiologist's patient population. If there are rare cases that are not represented, then a small number of those should be added to the test set to enrich it. The size of the test set needed will depend on how many different conditions (output classes) the AI provides and the frequency of various conditions in the population. For current AI applications coming to market that focus on lesion detection, 50–100 cases should be sufficient, and for AI focusing on diagnosis, approximately 200 is likely to be sufficient. Once the cases are identified, the images need to be extracted from the PACS. The radiology reports (and potentially medical record data such as pathology) should also be collected as described below to establish the ground truth.

The third step in AI evaluation is to establish the ground truth for each test case. If the AI algorithm provides a diagnosis, then data from the medical record or radiology report will be needed (see above). If the AI detects or segments lesions or organs, then the radiologist will need to review each image and provide ROIs to establish the ground truth. When the ground truth depends on the radiologist only (e.g., for diagnosis or for ROI), this is not a true ground truth, since there is inter-reader variability and such cases are better referred to as a “reference standard.” If one wanted a better ground truth, more than one radiologist could be engaged to review the images, but this is usually not practical and is generally not important when trying to show the AI is performing comparably to the radiologist.

The fourth step is to choose the appropriate evaluation metric for the AI algorithm. There are many potential metrics for evaluation [48–51], and we briefly discuss the most common metrics applicable to most current AI algorithms. For any AI task, the AI makes a determination (e.g., detects a lesion, makes a diagnosis, or includes certain pixels in a segmentation). That determination is compared with the ground truth for each case. If the AI makes a positive call that is correct, it is a true positive (TP). If that call is incorrect, it is a false positive (FP). If the AI makes a negative call (e.g.. no abnormality) that is correct, it is a true negative (TN) and if it is incorrect, it is a false negative (FN). Performance metrics are computed based on different combinations of TP, FP, TN, and FN (Figure 2). For detection, one often measures the precision and recall of the AI algorithm. Precision is the percentage of all positives that are true positives, or TP/(TP+FP), and recall is the percentage of all actual positives that are retrieved, or TP/(TP+FN). For diagnosis, one often measures sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). Sensitivity (also referred to as the true positive rate or recall) is the percentage of TP cases that are correctly identified (e.g., the percentage of disease cases that are correctly identified). Specificity (also referred to as the true negative rate) is the percentage of TN that are correctly identified (e.g., the percentage of normal cases that are correctly identified). A final metric that is often measured is the area under the ROC curve, which captures in single number how well an AI algorithm trades off sensitivity and specificity (Figure 3) The ROC curve is a compact summary of how well an AI algorithm performs overall, and it is commonly superimposed on those from other algorithms to compare AI methods. However, the ROC curve doesn’t give the entire story about how well an AI algorithm performs in clinical practice, it represents all possible operating points (tradeoffs between TP, FP, TN, and FN) of an AI algorithm. An AI algorithm that is deployed in practice operates at a particular point on the ROC curve, defining the values of the AI performance metrics (Figure 4). In general, an AI algorithm designed to detect abnormalities will have high sensitivity but poorer specificity (false positives). An algorithm designed to detect normal conditions (rule out disease) generally is tuned for specificity. Depending on the AI use case, sensitivity or specificity may be much more important, and the impact on the rate of FP and FN can be seen only by looking at the tails of the ROC curve (Figure 4). A high rate of FP, for example, is a major reason why the value of computer assisted diagnosis (CAD) algorithms to assist in evaluation of mammography imaging has been questioned [52–57]. However, more recent studies using newer AI models suggest potential value of CAD in breast imaging [35, 58]. Ultimately the value of a method needs to be assessed in a local practice, highlighting the importance of evaluation on local practice data.

The fifth step is to define a performance threshold for the metric that a radiologist wishes to achieve in order to decide that an AI algorithm will be useful in practice. Because of tradeoffs in performance metrics mentioned above, one generally does not aim for perfection in any one metric (e.g., 99% sensitivity in detecting cancer) since this may make other metrics poor clinically (e.g., unacceptably high false positive rate). A better approach may be for the radiologist to first assess her performance and then evaluate if the AI performs similarly well.

The sixth step is to evaluate the test cases against the metric. This is done by submitting the test cases to the AI algorithm and comparing its output to the ground truth (established third step) and to create a "confusion matrix." The confusion matrix is a table summarizing the number of TP, FP, TN, and FN, from which a number of AI performance metrics can be computed, such as precision, recall, accuracy, and others (Figure 2). As noted above, it could be useful to compute similar metrics for the radiologist (though this would require establishing a more robust ground truth, e.g., consensus read by several radiologists or pathology confirmation).

The seventh step is optional, entailing implementing a strategy to monitor AI performance, usually undertaken by intermittently repeating the preceding steps with a new (more recently acquired) test set. The rational for considering doing this is because radiology practices evolve over time, the patient populations change, imaging equipment and imaging protocols change, and all these as well as other factors may change how well an AI algorithm performs on the images (unless the AI algorithm is continually evolving as well). In addition, data registries that capture both AI performance metrics and metadata can help to determine specific situations where algorithm performance may be less than expected, as well as provide a way identify ways to improve AI algorithms.

(3) Maintaining radiologist expertise:

In the best case scenario, AI algorithms will be a supplemental resource to the radiologist, akin to a "second pair of eyes" rendering an opinion on cases, improving efficiency and diagnostic accuracy. This is similar to the radiologist showing a case to a colleague she trusts for a second opinion.

There are some dangers however, of unexpected negative consequences of AI on radiology practice, even if these algorithms perform well according to metrics on local practice data as described earlier. The first negative consequence is blind acceptance of the AI output. The AI algorithms are generally expected to be used to supplement, and not replace, the radiologist, who is presumed to have formulated an independent judgement before considering the output from the AI algorithm. In some cases, especially high volume and time-pressured practices, there may be a temptation to simply accept the AI reading and not formulate an independent judgement. In that case, radiologist performance will be no better than that of the AI algorithm (of course the same applies to showing a case to a colleague). The danger in the case of the AI algorithm, however, is that if it does not generalize well to unusual cases, it may lead the radiologist astray.

The second danger is adverse effects on over-reliance on technology. Reliance on technology in general reduces human resilience and can result in diminished human abilities [59]. A trivial example is the fact that few grocery clerks can make change without the assistance of the cash register. Fortunately, calculator technology is robust with failures being exceedingly rare. The potential dangers of diminished human ability in autonomous vehicles is substantially larger, where over-reliance in technology likely resulted in reduced human attention to monitoring of the technology with catastrophic results [60]. As humans rely more and more on technology that assists them in their tasks, they become less vigilant, which can compromise safety [61, 62]. Similar major issues have been encountered in medicine with reliance on auto-complete features of electronic pharmacy ordering systems [63]. There is likewise danger to radiologists relying too heavily on new AI technologies, which may reduce their attention and perceptive skills. In fact, at least one study has documented possible automation bias effects in CAD that degrade radiologist decision-making [64]. On the other hand, a recent study suggests that patients are less sure about the skills of computers, and they value the experience of the radiologist [65]. Patients have concerns that AI tools could produce restricted views with wrong diagnoses, and they believe such automated systems should remain secondary to the opinion of the radiologist. It will thus be beneficial for radiologists to keep these patient perspectives in mind as well as the pitfalls of assistive technologies as AI algorithms enter the market. Finally, over-reliance on technology and temptation to blindly accept AI outputs could adversely affect the training of future radiologists, who may not learn the critical observation and interpretative skills that make radiology a unique discipline.

CONCLUSION

The continual expansion of radiology in the healthcare process, the advances in imaging methods, and the volume of images they are producing, combined with the pressures of efficient workflow all create great demand for technologies that improve radiologist efficiency and reduce variation in practice without reducing accuracy. The AI tools coming to market offer potentially exciting opportunities to meet the needs of radiologist, but exuberance about their commercial prospects and the competitive business imperatives may push a flood of tools into the hands of radiologists with little understanding of whether and how to adopt them into their practices. The fact that many AI algorithms may not generalize to new data, combined with the regulatory pressures for rapid review and clearance, could cause unanticipated deleterious outcomes in clinical practice, particularly if the tradeoff of sensitivity and specificity of the AI tools is not optimal. Radiologists currently have little formal role in the development of AI tools other than being the targeted consumer. It is advisable for radiologists to become educated consumers, by (1) considering what clinical needs matter their practices, and whether the AI tools meet those needs, (2) evaluating AI tools they are considering adopting using case data from their own practices, (3) being cognizant of the potential hazards of over-reliance on technology and maintaining their clinical skills. The radiologist is ultimately responsible for the care of the patient, not the technology, and they will be well served by monitoring the benefit of these tools in their practice on an ongoing basis.

ACKNOWLEDGMENTS

(Blinded)

Funding support: This work was supported in part by grants from the National Cancer Institute, National Institutes of Health, U01CA142555, 1U01CA190214, and 1U01CA187947

REFERENCES

1. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks In:Advances in Neural Information Processing Systems 25: Curran Associates, Inc., 2012; 1097–1105.
2. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. Communications of the Acm 2017; 60:84–90.
3. Zeiler MD, Fergus R. Visualizing and Understanding Convolutional Networks. Computer Vision - Eccv 2014, Pt I 2014; 8689:818–833.
4. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. https://arxiv.org/abs/1409.1556; Accessed: 3/15/2019.
5. Szegedy C, Liu W, Jia YQ, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going Deeper with Convolutions. 2015 Ieee Conference on Computer Vision and Pattern Recognition (Cvpr) 2015:1–9.
6. He KM, Zhang XY, Ren SQ, Sun J. Deep Residual Learning for Image Recognition. 2016 Ieee Conference on Computer Vision and Pattern Recognition (Cvpr) 2016:770–778.
7. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz CP, Patel BN, Yeom KW, Shpanskaya K, Blankenberg FG, Seekins J, Amrhein TJ, Mong DA, Halabi SS, Zucker EJ, Ng AY, Lungren MP. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 2018; 15:e1002686. [PubMed: 30457988]
8. Nishio M, Sugiyama O, Yakami M, Ueno S, Kubo T, Kuroda T, Togashi K. Computer-aided diagnosis of lung nodule classification between benign nodule, primary lung cancer, and metastatic lung cancer at different image size using deep convolutional neural network with transfer learning. PLoS One 2018; 13:e0200721. [PubMed: 30052644]
9. Le MH, Chen J, Wang L, Wang Z, Liu W, Cheng KT, Yang X. Automated diagnosis of prostate cancer in multi-parametric MRI based on multimodal convolutional neural networks. Phys Med Biol 2017; 62:6497–6514. [PubMed: 28582269]
10. Qiu Y, Yan S, Gundreddy RR, Wang Y, Cheng S, Liu H, Zheng B. A new approach to develop computer-aided diagnosis scheme of breast mass classification using deep learning technology. J Xray Sci Technol 2017; 25:751–763. [PubMed: 28436410]
11. Higaki T, Nakamura Y, Tatsugami F, Nakaura T, Awai K. Improvement of image quality at CT and MRI using deep learning. Jpn J Radiol 2019; 37:73–80. [PubMed: 30498876]
12. Chaudhari AS, Fang Z, Kogan F, Wood J, Stevens KJ, Gibbons EK, Lee JH, Gold GE, Hargreaves BA. Super-resolution musculoskeletal MRI using deep learning. Magn Reson Med 2018; 80:2139–2154. [PubMed: 29582464]
13. Ma X, Hadjiiski LM, Wei J, Chan HP, Cha KH, Cohan RH, Caoili EM, Samala R, Zhou C, Lu Y. U-Net based deep learning bladder segmentation in CT urography. Med Phys 2019.
14. Candemir S, Antani S. A review on lung boundary detection in chest X-rays. Int J Comput Assist Radiol Surg 2019.
15. Sadda P, Imamoglu M, Dombrowski M, Papademetris X, Bahtiyar MO, Onofrey J. Deep-learned placental vessel segmentation for intraoperative video enhancement in fetoscopic surgery. Int J Comput Assist Radiol Surg 2019; 14:227–235. [PubMed: 30484115]
16. Sahiner B, Pezeshk A, Hadjiiski LM, Wang X, Drukker K, Cha KH, Summers RM, Giger ML. Deep learning in medical imaging and radiation therapy. Med Phys 2019; 46:e1–e36. [PubMed: 30367497]
17. Shen D, Wu G, Suk HI. Deep Learning in Medical Image Analysis. Annu Rev Biomed Eng 2017; 19:221–248. [PubMed: 28301734]
18. Tong N, Gou S, Yang S, Ruan D, Sheng K. Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks. Med Phys 2018; 45:4558–4567. [PubMed: 30136285]
19. Tudor J, Klochko C, Patel M, Siegal D. Order Entry Protocols Are an Amenable Target for Workflow Automation. J Am Coll Radiol 2018; 15:854–858. [PubMed: 29691135]

20. Winkel DJ, Heye T, Weikert TJ, Boll DT, Stieltjes B. Evaluation of an AI-Based Detection Software for Acute Findings in Abdominal Computed Tomography Scans: Toward an Automated Work List Prioritization of Routine CT Examinations. Invest Radiol 2019; 54:55–59. [PubMed: 30199417]

21. Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau NG, Venugopal VK, Mahajan V, Rao P, Warier P. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet 2018; 392:2388–2396. [PubMed: 30318264]

22. Nie D, Lu J, Zhang H, Adeli E, Wang J, Yu Z, Liu L, Wang Q, Wu J, Shen D. Multi-Channel 3D Deep Feature Learning for Survival Time Prediction of Brain Tumor Patients Using Multi-Modal Neuroimages. Sci Rep 2019; 9:1103. [PubMed: 30705340]

23. Shaish H, Mutasa S, Makkar J, Chang P, Schwartz L, Ahmed F. Prediction of Lymph Node Maximum Standardized Uptake Value in Patients With Cancer Using a 3D Convolutional Neural Network: A Proof-of-Concept Study. AJR Am J Roentgenol 2019; 212:238–244. [PubMed: 30540209]

24. Lin W, Tong T, Gao Q, Guo D, Du X, Yang Y, Guo G, Xiao M, Du M, Qu X, Alzheimer's Disease Neuroimaging I. Convolutional Neural Networks-Based MRI Image Analysis for the Alzheimer's Disease Prediction From Mild Cognitive Impairment. Front Neurosci 2018; 12:777. [PubMed: 30455622]

25. Cha KH, Hadjiiski Ph DL, Cohan Md RH, Chan Ph DH, Caoili Md EM, Davenport Md M, Samala Ph DR, Weizer Md AZ, Alva Md A, Kirova-Nedyalkova Md Ph DG, Shampain Md K, Meyer Md N, Barkmeier Md Ph DD, Woolen Md S, Shankar Md PR, Francis Md IR, Palmbos Md P. Diagnostic Accuracy of CT for Prediction of Bladder Cancer Treatment Response with and without Computerized Decision Support. Acad Radiol 2018.

26. Liang S, Zhang R, Liang D, Song T, Ai T, Xia C, Xia L, Wang Y. Multimodal 3D DenseNet for IDH Genotype Prediction in Gliomas. Genes (Basel) 2018; 9.

27. Causey JL, Zhang J, Ma S, Jiang B, Qualls JA, Politte DG, Prior F, Zhang S, Huang X. Highly accurate model for prediction of lung nodule malignancy with CT scans. Sci Rep 2018; 8:9286. [PubMed: 29915334]

28. Betancur J, Commandeur F, Motlagh M, Sharir T, Einstein AJ, Bokhari S, Fish MB, Ruddy TD, Kaufmann P, Sinusas AJ, Miller EJ, Bateman TM, Dorbala S, Di Carli M, Germano G, Otaki Y, Tamarappoo BK, Dey D, Berman DS, Slomka PJ. Deep Learning for Prediction of Obstructive Disease From Fast Myocardial Perfusion SPECT: A Multicenter Study. JACC Cardiovasc Imaging 2018; 11:1654–1663. [PubMed: 29550305]

29. Shi B, Grimm LJ, Mazurowski MA, Baker JA, Marks JR, King LM, Maley CC, Hwang ES, Lo JY. Prediction of Occult Invasive Disease in Ductal Carcinoma in Situ Using Deep Learning Features. J Am Coll Radiol 2018; 15:527–534. [PubMed: 29398498]

30. Chetlen AL, Chan TL, Ballard DH, Frigini LA, Hildebrand A, Kim S, Brian JM, Krupinski EA, Ganeshan D. Addressing Burnout in Radiologists. Acad Radiol 2018.

31. Barlow WE, Chi C, Carney PA, Taplin SH, D'Orsi C, Cutter G, Hendrick RE, Elmore JG. Accuracy of screening mammography interpretation by characteristics of radiologists. J Natl Cancer Inst 2004; 96:1840–1850. [PubMed: 15601640]

32. Burt JR, Torosdagli N, Khosravan N, RaviPrakash H, Mortazi A, Tissavirasingham F, Hussein S, Bagci U. Deep learning beyond cats and dogs: recent advances in diagnosing breast cancer with deep neural networks. Br J Radiol 2018; 91:20170545. [PubMed: 29565644]

33. Li W, Cao P, Zhao D, Wang J. Pulmonary Nodule Classification with Deep Convolutional Neural Networks on Computed Tomography Images. Comput Math Methods Med 2016; 2016:6215085. [PubMed: 28070212]

34. Bharti P, Mittal D, Ananthasivan R. Computer-aided Characterization and Diagnosis of Diffuse Liver Diseases Based on Ultrasound Imaging: A Review. Ultrason Imaging 2017; 39:33–61. [PubMed: 27097589]

35. Rodriguez-Ruiz A, Krupinski E, Mordang JJ, Schilling K, Heywang-Kobrunner SH, Sechopoulos I, Mann RM. Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System. Radiology 2019; 290:305–314. [PubMed: 30457482]

36. Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, Mehrtash A, Allison T, Arnaout O, Abbosh C, Dunn IF, Mak RH, Tamimi RM, Tempany CM, Swanton C, Hoffmann U, Schwartz LH, Gillies RJ, Huang RY, Aerts H. Artificial intelligence in cancer imaging: Clinical challenges and applications. CA Cancer J Clin 2019; 69:127–157. [PubMed: 30720861]

37. Yoo YJ, Ha EJ, Cho YJ, Kim HL, Han M, Kang SY. Computer-Aided Diagnosis of Thyroid Nodules via Ultrasonography: Initial Clinical Experience. Korean J Radiol 2018; 19:665–672. [PubMed: 29962872]

38. Tech & Telecom Law News. FDA Signals Fast-Track Approval for AI-Based Medical Devices. https://news.bloomberglaw.com/tech-and-telecom-law/fda-signals-fast-track-approval-for-ai-based-medical-devices-1; Accessed: 3/17/2019.

39. Manos D FDA upgrades approach to certifying new AI products. https://www.healthcareitnews.com/news/fda-upgrades-approach-certifying-new-ai-products; Accessed: 3/17//2019.

40. U. S. Food and Drug Administration. Digital Health Innovation Action Plan. https://www.fda.gov/downloads/medicaldevices/digitalhealth/ucm568735.pdf; Accessed: 3/17/2019.

41. Teerlink M The AI Gold Rush: Artificial Intelligence And Machine Learning. https://www.digitalistmag.com/future-of-work/2018/10/08/ai-gold-rush-artificial-intelligence-machine-learning-06188069; Accessed: 3/17/19.

42. Sejnowski TJ. The deep learning revolution. Cambridge, Massachusetts: The MIT Press, 2018.

43. American College of Radiology Data Science Institute. TOUCH-AI use cases. https://www.acrdsi.org/DSI-Services/TOUCH-AI; Accessed: 3/17/2019.

44. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images: Results from Recently Published Papers. Korean J Radiol 2019; 20:405–410. [PubMed: 30799571]

45. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med 2018; 15:e1002683 PMID:30399157. PMCID:PMC6219764 following competing interests: MAB and ML are currently employees at Verily Life Sciences, which played no role in the research and has no commercial interest in it. EKO and ABC receive funding from Intel for unrelated work. [PubMed: 30399157]

46. Pan I, Agarwal S, Merck D. Generalizable Inter-Institutional Classification of Abnormal Chest Radiographs Using Efficient Convolutional Neural Networks. 10.1007/s10278-019-00180-9; Accessed: 3/17/2019.

47. Lohr S Facial Recognition Is Accurate, if You're a White Guy. https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html, The New York Times; Accessed: 3/18/2019.

48. Momeni A Assessing Diagnostic Tests In:Introduction to statistical methods in pathology. New York, NY: Springer Science/Business Media, LLC, 2017; 7–37.

49. Simundic AM. Measures of Diagnostic Accuracy: Basic Definitions. EJIFCC 2009; 19:203–211. [PubMed: 27683318]

50. Park SH, Han K. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. Radiology 2018; 286:800–809. [PubMed: 29309734]

51. Handelman GS, Kok HK, Chandra RV, Razavi AH, Huang SW, Brooks M, Lee MJ, Asadi H. Peering Into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods. American Journal of Roentgenology 2019; 212:38–43. [PubMed: 30332290]

52. Fenton JJ, Xing G, Elmore JG, Bang H, Chen SL, Lindfors KK, Baldwin LM. Short-term outcomes of screening mammography using computer-aided detection: a population-based study of medicare enrollees. Ann Intern Med 2013; 158:580–587. [PubMed: 23588746]

53. Fenton JJ, Taplin SH, Carney PA, Abraham L, Sickles EA, D'Orsi C, Berns EA, Cutter G, Hendrick RE, Barlow WE, Elmore JG. Influence of computer-aided detection on performance of screening mammography. N Engl J Med 2007; 356:1399–1409. [PubMed: 17409321]

54. Fenton JJ, Abraham L, Taplin SH, Geller BM, Carney PA, D'Orsi C, Elmore JG, Barlow WE, Breast Cancer Surveillance C. Effectiveness of computer-aided detection in community mammography practice. J Natl Cancer Inst 2011; 103:1152–1161. [PubMed: 21795668]

55. Tchou PM, Haygood TM, Atkinson EN, Stephens TW, Davis PL, Arribas EM, Geiser WR, Whitman GJ. Interpretation time of computer-aided detection at screening mammography. Radiology 2010; 257:40–46. [PubMed: 20679448]

56. Lehman CD, Wellman RD, Buist DS, Kerlikowske K, Tosteson AN, Miglioretti DL, Breast Cancer Surveillance C. Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. JAMA Intern Med 2015; 175:1828–1837. [PubMed: 26414882]

57. Kaunitz AM Computer-Aided Detection for Mammography: Time to Say Goodbye? https://www.jwatch.org/na39183/2015/10/06/computer-aided-detection-mammography-time-say-goodbye, NEJM; Accessed: 3/17/2019.

58. Rodriguez-Ruiz A, Lang K, Gubern-Merida A, Broeders M, Gennaro G, Clauser P, Helbich TH, Chevalier M, Tan T, Mertelmeier T, Wallis MG, Andersson I, Zackrisson S, Mann RM, Sechopoulos I. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. J Natl Cancer Inst 2019.

59. Osoba OA, Welser W. The Risks of Artificial Intelligence to Security and the Future of Work. https://www.rand.org/pubs/perspectives/PE237.html, RAND Corporation; Accessed: 3/17/2019.

60. Smith S NTSB: Fatal Crash Involving Tesla Autopilot Resulted from Driver Errors, Overreliance on Automation. https://www.ehstoday.com/safety/ntsb-fatal-crash-involving-tesla-autopilot-resulted-driver-errors-overreliance-automation; Accessed: 3/17/2019.

61. Thakkar VG. Forget Self-Driving Cars. Bring Back the Stick Shift: Technology meant to save us from distraction is making us less attentive. https://www.nytimes.com/2019/03/23/opinion/sunday/stick-shift-cars.html, The New York Times; Accessed: 3/24/2019.

62. Newton C Reliance on autopilot is now the biggest threat to flight safety, study says. https://www.theverge.com/2013/11/18/5120270/reliance-on-autopilot-is-now-the-biggest-threat-to-flight-safety, The Verge; Accessed: 3/24/2019.

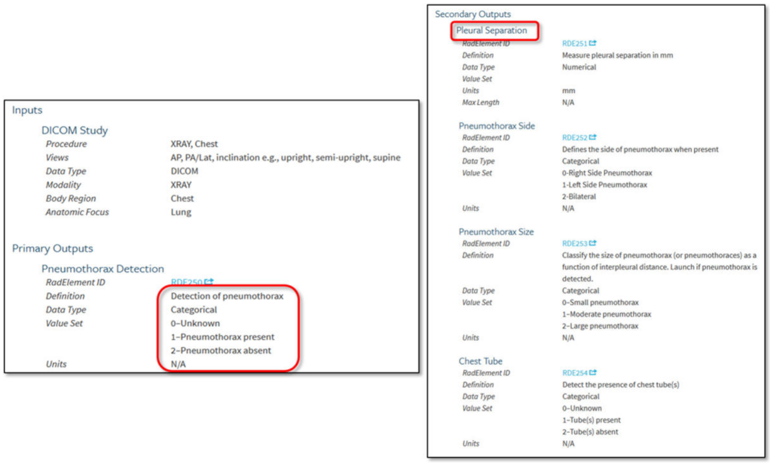
63. Institute for Safe Medication Practices Canada. Understanding Human Over-reliance on Technology. Report Medication Incidents 2016; 16:1–6.

64. Alberdi E, Povykalo A, Strigini L, Ayton P. Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography. Acad Radiol 2004; 11:909–918. [PubMed: 15354301]

65. Haan M, Ongena YP, Hommes S, Kwee TC, Yakar D. A Qualitative Study to Understand Patient Perspective on the Use of Artificial Intelligence in Radiology. J Am Coll Radiol 2019.

Take Home Points

1. The pace of AI development is exploding, and the number of AI tools being marketed to radiologists is accelerating, posing challenges for radiologists to decide which tools to adopt.
2. The role of radiologists in imaging AI is to identify important clinical use cases where these tools are needed and to evaluate their effectiveness in clinical practice.
3. AI tools are expected to improve radiologist practice, but they must guard against over-reliance on these technologies and the accompanying loss of clinical expertise.



Inputs

DICOM Study

Procedure	XRAY, Chest
Views	AP, PA/Lat, inclination e.g., upright, semi-upright, supine
Data Type	DICOM
Modality	XRAY
Body Region	Chest
Anatomic Focus	Lung

Primary Outputs

Pneumothorax Detection

RadElement ID	RDE250
Definition	Detection of pneumothorax
Data Type	Categorical
Value Set	0–Unknown 1–Pneumothorax present 2–Pneumothorax absent
Units	N/A

Secondary Outputs

Pleural Separation

RadElement ID	RDE251
Definition	Measure pleural separation in mm
Data Type	Numerical
Value Set	
Units	mm
Max Length	N/A

Pneumothorax Side

RadElement ID	RDE252
Definition	Defines the side of pneumothorax when present
Data Type	Categorical
Value Set	0-Right Side Pneumothorax 1-Left Side Pneumothorax 2-Bilateral
Units	N/A

Pneumothorax Size

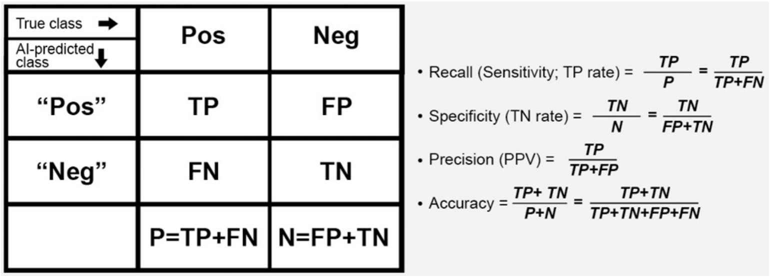
RadElement ID	RDE253
Definition	Classify the size of pneumothorax (or pneumothoraces) as a function of interpleural distance. Launch if pneumothorax is detected.
Data Type	Categorical
Value Set	0–Small pneumothorax 1–Moderate pneumothorax 2–Large pneumothorax
Units	N/A

Chest Tube

RadElement ID	RDE254
Definition	Detect the presence of chest tube(s)
Data Type	Categorical
Value Set	0–Unknown 1–Tube(s) present 2–Tube(s) absent
Units	N/A

Figure 1.

Example ACR DSI use case for AI algorithms (only the sub-section of the use case specifying inputs/outputs of the AI algorithm for this use case is shown, cited from https://www.acrdsi.org/DSI-Services/TOUCH-AI/Use-Cases/Pneumothorax). In evaluating an AI use case, the radiologist should read the text description to understand the clinical goal of the use case and then examine the outputs from the AI algorithm to determine if those will be helpful to the radiologist's practice, e.g., detection of pneumothorax, pleural separation, laterality, size, and presence of chest tube.



True class → AI-predicted class ↓	Pos	Neg
“Pos”	TP	FP
“Neg”	FN	TN
	P=TP+FN	N=FP+TN

- Recall (Sensitivity; TP rate) = $\frac{TP}{P} = \frac{TP}{TP+FN}$
- Specificity (TN rate) = $\frac{TN}{N} = \frac{TN}{FP+TN}$
- Precision (PPV) = $\frac{TP}{TP+FP}$
- Accuracy = $\frac{TP+TN}{P+N} = \frac{TP+TN}{TP+TN+FP+FN}$

Figure 2.

Confusion matrix and definition of the common metrics for evaluating AI algorithms. Note that a reliable gold standard (ground truth) is critical to establish the “True class”)

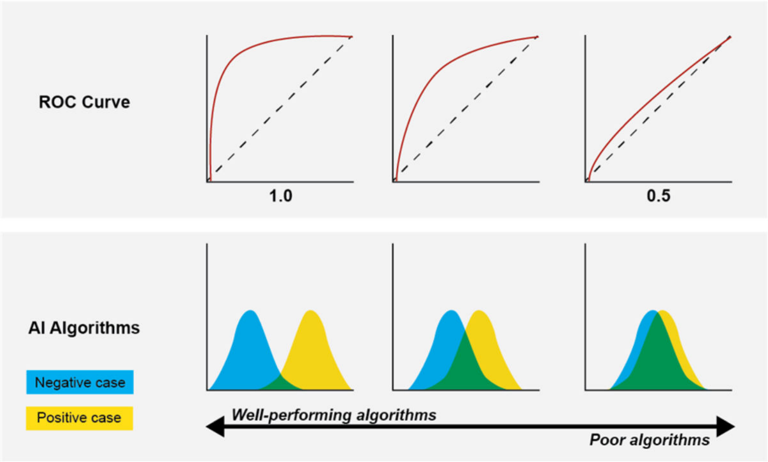


Figure 3.

ROC curves for different AI algorithms. Well-performing AI algorithms separate positive and negative cases well (lower left) and thus the ROC curve has an area near 1 (upper left), while poor AI algorithms do not separate the cases well (lower right) and have an ROC curve area of 0.5 (upper right).

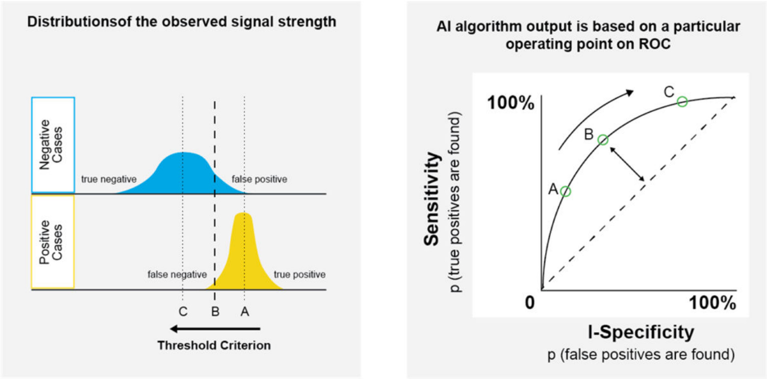

Distributionsof the observed signal strength
Negative Cases
Positive Cases
true negative
false positive
false negative
true positive
C
B
A
Threshold Criterion
AI algorithm output is based on a particular operating point on ROC
100%
C
B
A
Sensitivity
p (true positives are found)
0
100%
l-Specificity
p (false positives are found)

Figure 4.

The importance of the operating point on the ROC curve. A given AI algorithm has a particular ROC curve (see Figure 3), and given that curve, a particular point is selected by the AI algorithm developer to determine the output of the AI algorithm (e.g., when to call an image “normal” or “abnormal.”) Depending on that point chosen, the AI performance metrics may vary greatly. Point “A” provides maximum specificity (right panel), since it calls nearly all the negative cases correctly and produces very few false positives (left panel). Point C, on the other hand, has maximum sensitivity (right panel), since it picks up all of the true positive cases, but produces many false positives (left panel). Point B balances sensitivity and specificity, picking up nearly all of the true positive, but having a reasonably large number of false positives. Whether high sensitivity (cancer detection) or specificity (ruling out an abnormality) is most important depends on the use case, and the tolerance for the number of false positives and false negatives also depends on the use case.

Table 1.

List of primary areas of clinical application for AI methods and current status of development and availability

APPLICATION AREA	CURRENT STATUS
Image enhancement	In market or soon to market
Disease detection	
Lesion segmentation	
Diagnosis	
Treatment selection	In development
Response assessment	
Clinical prediction (of treatment response or future disease)	
Image enhancement	

Table 2.

Steps for the radiologist to undertake an evaluation of an AI algorithm in their practice.

1.	Understand the key outputs of the AI algorithm (e.g., what is it predicting or producing?) and decide which is/are clinically relevant to the radiologist clinical needs
2.	Collect representative patient samples (test cases)
3.	Establish ground truth for each test case
4.	Choose appropriate evaluation metric (e.g., sensitivity, specificity, PPV)
5.	Define performance threshold for the metric (e.g., 99% sensitivity in detecting cancer; this sets a threshold on false positives)
6.	Evaluate the test cases against the metric
7.	Implement monitoring strategy