Importance of Better Human-Computer Interaction in the Era of Deep Learning: Mammography Computer-Aided Diagnosis as a Use Case

Robert M. Nishikawa, PhD, **Kyongtae T. Bae, MD, PhD, MBA**

Department of Radiology, University of Pittsburgh, Pittsburgh, Pennsylvania.

As witnessed at last year's RSNA, the revolution is upon us. Experts tout deep learning as a revolution in radiology. Some radiologists fear for their jobs because some experts suggest deep learning as an infallible replacement for radiologists [1,2]. Radiologists at the forefront of this technology have tried to assure their colleagues that there is nothing to fear and that deep learning will make their lives better by performing tedious repetitive tasks and by providing intelligent decision support tools to enhance radiologists' real expertise and performances. Although deep learning may indeed provide tools to enhance radiologists' performances and efficiency, from our perspective, investigators place too much emphasis on developing the tools and not nearly enough effort on the optimum implementation of these tools.

In this opinion article, we focus on the use of deep learning as related to a specific clinical application of radiology image analysis, namely computer-aided diagnosis (CAD). We restrict our argument to CAD, and specifically CAD for mammography, for a number of reasons. First, CAD is an obvious application of deep learning, and there is much activity both academically and commercially. Second, mammography CAD is the most widely used CAD application, and there are several studies evaluating its impact on screening mammography. Although we take a narrow focus here, we assert that this CAD example is relevant to all other applications where using deep learning assists radiologists. We need more research on the man-machine interface in parallel with developing algorithms.

DEEP LEARNING IN RADIOLOGY AND CAD

Deep learning refers to a class of machine-learning algorithms that are inspired by and simulate the multilayer structure and function of the brain, called *artificial neural networks*. Armed with powerful computing, big data, and smart algorithms, deep learning technology has shown impressive outcomes in a number of IT applications in a wide range of industries and professions. This technology is also expected to make a huge impact in many aspects of radiology practice, including image interpretation, delivery and application of radiology information, clinical decision making, and outcome management. With the success of deep learning technology in the future, the role and the business of radiology will change drastically.

Robert Nishikawa: Department of Radiology, University of Pittsburgh, 3362 Fifth Ave, FARP, Pittsburgh PA 15213; nishikawarm@upmc.edu.

CAD is largely divided into two groups: computer-aided detection (CADe) and computer-aided diagnosis (CADx). CADx represents a class of broader and more sophisticated tools than CADe, but CADx has not been integrated into the clinic for a variety of reasons. On the other hand, CADe is FDA approved and is available for clinical use in radiology practices. Deep learning in the future may allow us to implement more powerful and versatile tools that combine CADe and CADx. With this goal in mind, a number of research groups are working on applying deep learning to CAD, encompassing a wide range of body parts and pathologies.

CAD IN CURRENT CLINICAL BREAST IMAGING

In our opinion, applying deep learning to current breast imaging CAD will not be successful if we just develop more accurate algorithms. Breast imaging CAD has penetrated the market only for screening mammography and, appreciably, only in the United States. Better algorithms alone will not improve CAD clinically. The accuracy of the algorithms is not what is limiting the current success of CAD in breast imaging. We assert that the manner in which CAD has been implemented has dampened its effectiveness. We present two pieces of evidence for this assertion.

First, Nishikawa et al showed that radiologists ignore the vast majority (71%) of correct computer detections of a cancer on a mammogram [3]. When initially developing CAD, the prevailing thought was that if the computer marked a lesion, radiologists, because they are good at detecting cancer, would immediately recognize that they missed a cancer. This is, apparently, not the clinical reality. Visual psychologist Jeremy Wolfe and his team showed that radiologists are more likely to miss a cancer if the prevalence of cancer is low than if the prevalence of cancer is high [4]. This is a fundamental property of human perception [5].

When applied to screening mammography, the positive predictive value (PPV) of a CAD prompt is low. Given that breast cancer prevalence is 0.5%, a radiologist with a sensitivity of 80% will have one missed cancer for CAD to detect in every 1,000 women screened. Assuming CAD sensitivity is 100% and there are two false detections per case (0.5 per image or 2,000 images with a false-positive), then the prevalence of a true-positive, or PPV of CAD detection (assuming CAD detected the cancer in two views), is 0.1% (2 true-positive detections and 2,000 false-positives). Even if we improve the false detection rate of CAD by a factor of 10, the prevalence of a true computer detection is still low. Extrapolating the findings of Wolfe and colleagues, because the prevalence of the true CAD detection is low, radiologists tend to ignore the true computer detections more than they would at high prevalence. Although this would argue for the need of better CAD algorithms, the level of performance at which CAD would have high PPV exceeds the performance of most breast radiologists, at which point CAD as a second reader would be replaced by CAD as the primary reader.

Second, radiologists are not using screening mammography CAD as intended, which is as a second reader. Table 1 shows published data from three different clinical studies of CAD. All studies used data from the Breast Cancer Screening Consortium. The data show a clear trend for decline in incremental improvement in sensitivity and recall rate with using CAD.

When used correctly as a second reader, investigators designed CAD to help radiologists avoid overlooking a cancer. CAD is not designed for improving specificity. Therefore, sensitivity and recall rate should both increase. Table 1 shows that, over time, the benefit of CAD to improve sensitivity lessened and the detriment of CAD to increase recall rate also declined, to the point where both the sensitivity and the recall rate actually reduced with using CAD. This is an unexpected trend, particularly for recall rates, and is not possible under the intended second reader paradigm. One possible explanation is that radiologists consult CAD to avoid overcalling. Radiologists concerned about a high recall rate may not recall suspicious lesions that are not detected by CAD, contributing to reduced sensitivity. Another possibility is that radiologists use CAD as a concurrent, or first, reader paradigm. Zheng et al have shown that when used as a concurrent, or first, reader, CAD can reduce sensitivity because radiologists can miss cancers not detected by CAD compared with reading without CAD [6].

RADIOLOGISTS AND CAD GO ON A BLIND DATE

Just like a radiologist, CAD is subject to a finite diagnostic performance with its own sensitivity and specificity profile. If radiologists are overconfident of CAD's input, they may suppress their own sound diagnostic judgment to blindly follow CAD's choice. On the other hand, underappreciating CAD's input may result in rejecting an accurate CAD diagnosis. Currently, there is no mechanism for a radiologist or CAD to know the consequences of each other's action. Radiologists and CAD operate independently with little feedback and interaction. Investigators train and validate CAD at a laboratory with a set of data, and radiologist users receive their training with their own set of clinical experiences and are subject to their own heurist bias. Radiologist users have limited understanding of CAD performance characteristics and the underlying decision mechanisms. There is no systemic feedback to narrow the mutual biases between the radiologist and CAD. Radiologists and CAD are essentially in a blind date with limited understanding of each other. We consider this decision architecture of CAD as an inadequate and inefficient clinical diagnostic or radiologic decision support tool.

To foster and promote CAD as a well-balanced partner with radiologists, we should understand the behavioral aspect of the radiologist's clinical practice and consider this behavior in the clinical decision architecture when implementing CAD. With deep learning technology and big data, we may be able to design CAD to guide and protect radiologists by anticipating potential misuse of CAD by radiologists and algorithmically discouraging them from taking an undesirable shortcut that may negatively affect diagnostic performance. A smart decision support tool could provide checks, balances, and feedback and actively improve a radiologist's performance.

As we consider developing CAD as a well-informed decision support tool, we cite research from Daniel Kahneman, a psychologist, a Nobel laureate in economics, and the founder of behavior economics. He modeled the workings of the mind as an uneasy interaction between two fictitious, but insightful, characters: the automatic system 1 and the effortful system 2 [7]. System 1 uses association and metaphor to produce a quick and dirty draft of reality, which system 2 draws on to arrive at explicit beliefs and reasoned choices. Ideally, we

should use system 2 to evaluate information and make logical and rational decisions. But, as Kahneman describes it, system 2 can be lazy and just endorse system 1, even if system 1's choice is not logical. This may explain the findings of Wolfe et al [4, 5] and Zheng et al. [6]

When using CAD as the first or concurrent reader, it likely enforces a radiologist's fast automatic system 1 thinking and turns off the controlled process of system 2 thinking. As a consequence, a radiologist may gain productivity but lose an opportunity to reason and come up with an alternative decision. When using CAD appropriately as the second reader, it could encourage a radiologist's system 2 thinking. However, a simple listing of CAD detection may not be sufficient to stimulate a radiologist's system 2 thinking to reconsider the initial choice. Radiologists would simply write off CAD's suggestion even if it is accurate. A more effective approach would be presenting a CAD choice designed with the knowledge of an individual radiologist's reaction to particular patterns of images.

Future CAD may have its decision architecture tuned to engage and aid a radiologist's system 2 thinking when needed. Turning system 2 on constantly when system 1 is making a routine decision is inefficient and creates limited capacity. So, CAD, equipped with deep learning capability, learns and analyzes a radiologist's previous interpretation data to heed when mistakes are likely and intervenes effectively with system 2 to avoid significant mistakes. This may be possible if we provide the radiologist with appropriate feedback on why the CAD detected something that the radiologist disagrees with. Kahneman's team also observed that research subjects' pupils dilated significantly more when involved in system 2 cognitive activity than when involved in system 1 activity [7]. Future CAD may be able to take advantage of multisensory capability to actively gauge a radiologist's attentiveness and to align system 2 operation mode when needed.

WHAT'S NEXT?

How deep learning will impact radiology is unknown. There are no large-scale implementations to evaluate. One aspect of deep learning is CAD, and we do have several large studies of CAD (that do not implement deep learning). We need to couple the development of deep learning to improve CAD performance with the development of better methods of implementing CAD. We should investigate at least three approaches. We are hopeful that there are others. (1) Implementing CAD in an interactive format in which radiologists only request assistance from CAD when they would like help with a case [8]. (2) Providing radiologists with more information upon request, such as providing similar imaging cases with known outcome [9,10]. (3) Alerting radiologists to cases that they identify as normal and CAD flags as a high probability that an abnormality exists. We should personalize this approach to the radiologist. For example, we train the deep learning algorithm on imaging data interpreted previously by individual radiologists to offer a personalized CAD program that, as an informed partner, would help predict and appropriately modify the radiologist's behavior. We suggest giving no location information to the radiologist, only that he or she should reexamine the case more carefully. The overarching theme is to have radiologists carefully interact with the CAD output (system 2) only when necessary and not to blindly and needlessly offer radiologists the CAD output. In that way, radiologists will make the best informed decision efficiently.

We expect investigators to spend a tremendous amount of effort in terms of manpower, money, and ideas for developing diagnostic imaging tools based on deep learning. However, without significant advances in how to best implement these tools clinically, we are in for another round of disappointments, and the revolution, like a mismatched couple on a blind date, will end in failure.

Acknowledgments

REFERENCES

1. Chockley K, Emanuel E. The end of radiology? Three threats to the future practice of radiology. J Am Coll Radiol 2016;13: 1415–20. [PubMed: 27652572]

2. Ng A. Automation and anxiety. The Economist. June 25, 2016 Available at: https://www.economist.com/news/special-report/21700758-will-smarter-machines-cause-mass-unemployment-automation-and-anxiety. Accessed October 20, 2017.

3. Nishikawa RM, Schmidt RA, Linver MN, Edwards AV, Papaioannou J, Stull MA. Clinically missed cancer: how effectively can radiologists use computer-aided detection? AJR Am J Roentgenol 2012;198:708–16. [PubMed: 22358014]

4. Evans KK, Birdwell RL, Wolfe JM. If you don't find it often, you often don't find it: why some cancers are missed in breast cancer screening. PLoS ONE 2013;8: e64366. [PubMed: 23737980]

5. Wolfe JM, Horowitz TS, Kenner NM. Rare items often missed in visual searches. Nature 2005;435:439–40. [PubMed: 15917795]

6. Zheng B, Ganott MA, Britton CA, et al. Soft-copy mammographic readings with different computer-assisted detection cuing environments: preliminary findings. Radiology 2001;221:633–40. [PubMed: 11719657]

7. Kahneman D. Thinking, Fast and Slow. New York, NY: Farrar, Straus and Giroux; 2013.

8. Hupse R, Samulski M, Lobbes MB, et al. Computer-aided detection of masses at mammography: interactive decision support versus prompts. Radiology 2012;266(1):123–9. [PubMed: 23091171]

9. Zheng B, Lu A, Hardesty LA, et al. A method to improve visual similarity of breast masses for an interactive computer-aided diagnosis environment. Med Phys 2006;33:111–7. [PubMed: 16485416]

10. Giger ML, Huo Z, Vyborny CJ, et al. Intelligent CAD workstation for breast imaging using similarity to known lesions and multiple visual prompt aids. Proc SPIE Med Imag 2002;4684:768–73.

11. Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. N Engl J Med 2007;356:1399–1409. [PubMed: 17409321]

12. Fenton JJ, Abraham L, Taplin SH, et al. Effectiveness of computer-aided detection in community mammography practice. J Natl Cancer Inst 2011;103:1152–61. [PubMed: 21795668]

13. Lehman CD, Wellman RD, Buist DM, et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Intern Med 2015;175:1828–37. [PubMed: 26414882]

Table 1.

Trend of changes in sensitivity and recall rate with the use of CAD

Study	Sensitivity			Recall Rate (per 100 Women Screened)		
	Before CAD	With CAD	% Change	Before CAD	With CAD	% Change
Fenton et al (2007) [11]	0.804	0.840	4.5	10.1	13.2	30.7
Fenton et al (2011) [12]	0.797	0.811	1.8	8.4	8.9	5.6
Lehman et al (2015) [13]	0.873	0.853	−2.3	9.1	8.7	−4.4

The trend in specificity (and recall rate) was statistically significant ($P < .0001$, logistic regression, proc logistic, version 9.4, SAS, Cary, North Carolina, USA). The trend in sensitivity was not statistically significant ($P = .45$). CAD = computer-aided diagnosis.