Why CAD Failed in Mammography

Ajay Kohli, MD, Saurabh Jha, MD

OVERVIEW

Many, skeptical that deep learning, computer vision, or, broadly speaking, artificial intelligence (AI), will change health care in general and radiology in particular, cite the failure of computer-aided diagnosis in mammography. We will review the history of CAD, analyze why CAD failed in mammography, and conjecture how future CAD platforms can succeed. To understand how AI can work, we must understand why CAD failed.

WHY CAD FAILED IN MAMMOGRAPHY

The potential for computers to interpret medical images has been conjectured, especially over the last decade. Although there have been pivotal moments, perhaps one of the more defining moments of this vision was when AlphaGo (DeepMind, Alphabet, London, UK), a computer program using deep neural networks, beat the professional champions of *Go* [1]. The machine's victory over humans instigated an exploration of utility of deep learning platforms in data-rich fields such as radiology. To understand how neural networks can successfully interpret medical images, we must understand why their primitive versions failed. The case we will discuss is the computer-aided detection of cancer on mammograms.

History

Automated computer-aided diagnosis (CAD) comprises CADe (computer-aided detection) and CADx (computer-aided diagnosis). CAD is a multistep process. The data input is the mammogram image where abnormalities in breast tissue, such as clustered microcalcifications and masses, are detected using CADe. Then CADx classifies these abnormalities using quantitative tools assessing features such as spatial density and regularity of cluster. The likelihood of malignancy is estimated using an automated trained classifier program. Finally, the areas of concern on the mammogram are flagged by the computer for the attention of the interpreter [2]. The computer, like the interpreter, must make a choice between overcalling false-positives and undercalling false-negatives. The original intention was that CAD discriminate benign from malignant lesions with a high accuracy and limit false-positives [3].

The FDA approved CAD in mammography in 1998 [3]. The CMS approved reimbursement for CAD in 2002, when PACSs were being rapidly adopted. The convergence of reimbursement and new technology facilitated the implementation of CAD, and by 2010, approximately 74% of mammography interpretations utilized CAD [4].

The optimism of CAD was justified by initial studies comparing CAD to double readers, which concluded that CAD can increase cancer detection rates, even if only by small amounts ranging from 2% to 10%, and accelerate detection of cancer by 2 months. The only issue was its cost-effectiveness [5-7]. However, as data emerged, the utility of CAD was questioned [8,9].

Failure of CAD

From a radiologist's perspective, CAD was a "second reader," or a "spell-checker," which compelled radiologists to look at every region flagged [10]. Subsequent studies found that CAD neither encouraged radiologists to change their decisions nor improved their accuracy [11].

A limitation of earlier studies that examined the utility of CAD was the radiologist's learning curve. The utility of CAD arguably depended on the skill of the radiologist. This confounder was adjusted for in the most exhaustive review of impact of CAD by Lehman et al, who examined more than 495,000 mammograms interpreted by CAD [12].

The researchers incorporated studies of, and adjusted for, a variety of subgroups, including but not limited to age, breast density, menopausal status, and time since last mammogram. The researchers were unable to find improved outcomes with CAD. However, for every true-positive cancer detected by CAD there were more false-positives, which had to be ignored by radiologists. CAD did not help radiologists who already had a high sensitivity. Far from being helpful, CAD was like a monkey on the shoulder of radiologists, who

1546-1440/17/$36.00 ■ https://doi.org/10.1016/j.jacr.2017.12.029

subsequently had to silence the several false predictions suggested by CAD.

Not only did CAD increase the recalls without improving cancer detection [13], but, in some cases, even *decreased* sensitivity by missing some cancers, particularly non-calcified lesions. CAD could lull the novice reader into a false sense of security [7,12]. Thus, CAD had both lower sensitivity and lower specificity, a nonredeeming quality for an imaging test.

There were technical reasons for CAD's inferior performance. The processing power of the first versions of CAD was limited, which limited the analysis of different views of mammograms and review of old imaging [10]. Although the radiologist could compare mammograms to old studies, limitations in processing power prevented CAD from doing so.

The key issue was that CADx platform was developed using supervised learning. In supervised learning, the computer is trained on samples with known pathology (truth) and then tested for its ability to predict the likelihood of malignancies in a test sample (truth and lies). Despite the allure of supervision, the pedagogy is not neutral. Because the computer sees more cancers during its training than its test, there is verification bias, and the specificity drifts [14,15].

Complicating the computer's education is the fact that imaging interpretation varies considerably between radiologists [16,17]. Radiologists are not the most consistent of teachers. The truth sometimes lied.

Because CAD was marked as a "second pair of eyes," it was marketed as a "second set of revenue generators" to imaging centers. After CMS approved reimbursement for CAD, vendors advertised a >$7 reimbursement for every dollar invested in CAD [14].

Meanwhile, costs of breast cancer screening continued to increase, and according to recent estimates are $1 billion annually in the Medicare population alone [18]. Though CAD may be cheaper than double reads by radiologists [19,20], CAD adds a tab of $400 million per year compared with interpretation by a single radiologist [12]. Improving a radiologist's area under the receiver operating characteristics curve costs, but CAD costs without improving the receiver operating characteristics.

In summary, we believe CAD failed because of insufficient processing power and supervised learning. Its widespread implementation unmasked the lack of its effectiveness.

FUTURE APPLICATIONS: WHY CAD 2.0 MIGHT NOT FAIL

Only 84% of breast cancers are detected by interpreting radiologists. The 16% of cancers that are missed by radiologists likely reflect limitations in image perception by the human eye [21]. AI can help if we focus on the unique features of the missed cancers.

CAD failed, in no small part, because it was recruited to do what radiologists already do well—pick up the 84% of cancers. In screening for breast cancer, AI should function not as a second pair of eyes, but as a Hubble telescope, to see what we cannot see.

What should be done differently?

CAD 2.0 should retrieve and manage data for the radiologist. Its role should be quantitative analysis. Rather than replicate large-scale maps, CAD 2.0 should zoom into the areas of blindness. Radiologists know their blind spots. One example is the dense breast, which is known to reduce the sensitivity of the mammogram to dangerously low levels [22]. Another example is patients with reconstructed breasts. A radiologist's largest blind spot is what imaging cannot see, and this includes information from the "-omics" such as proteomics and genomics.

The major shift, owing to more powerful graphical processing units, is the change from supervised learning to unsupervised learning. AI will no longer be bound to radiologists for determination of the truth, but may, through its own pattern-learning abilities (which will get better with time and sample) teach radiologists what they are liable to miss.

As AI figures stuff out on mammograms, the lesions can be compared with and clustered with similar lesions that have known pathology. Then, using content-based image retrieval, these lesions can then be presented to the radiologist with a set of images that share similar features and have known pathology [23]. That is, AI can create a Tower of Babel of mammograms, ultimately giving radiologists interpreting mammograms a reliable dichotomous answer—benign or malignant.

As computer power doubles in short periods of time [24], CAD 2.0 could be applied to not just one view of the same image, but also into different views, prior images, and even non-imaging data, such as pathology images.

Newer technologies can use "transfer learning" where algorithms developed on a certain data set—for example, one image-view mammograms—are applied to different data sets [25]. Perhaps the biggest application for this technology will be in breast tomosynthesis [26]. Even traditional CAD is promising in tomosynthesis [27]. Deep learning has been used in tomosynthesis with success [25].

There are lessons to learn from the failure of CAD 1.0, but CAD should not be written off. With greater computing power, unsupervised

learning, and amalgamation of information from disparate sources, CAD 2.0 may succeed, particularly where radiologists fail.

REFERENCES

1. Silver D, Huang A, Maddison CJ, et al. Mastering the game of Go with deep neural networks and tree search. Nature 2016;529:484-9.
2. Jiang Y, Nishikawa RM, Wolverton DE, et al. Malignant and benign clustered microcalcifications: automated feature analysis and classification. Radiology 1996;198:671-8.
3. From the Federal Register Online via GPO Access (P970058). U. S. Food and Drug Administration; 1998. p. 970058. Available at: http://www.fda.gov/ohrms/dockets/98fr/123098b.txt.
4. Rao VM, Levin DC, Parker L, et al. How widely is computer-aided detection used in screening and diagnostic mammography? J Am Coll Radiol 2010;7:802-5.
5. Ko JM, Nicholas MJ, Mendel JB, Slanetz PJ. Prospective assessment of computer-aided detection in interpretation of screening mammography. AJR Am J Roentgenol 2006;187:1483-91.
6. Taplin SH, Rutter CM, Lehman CD. Testing the effect of computer-assisted detection on interpretive performance in screening mammography. AJR Am J Roentgenol 2006;187:1475-82.
7. Gur D, Sumkin JH. CAD in screening mammography. AJR Am J Roentgenol 2006;187:1474-1474.
8. 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;198:708-716.
9. 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.
10. Giger ML, Chan HP, Boone J. Anniversary paper: history and status of CAD and quantitative image analysis: the role of medical physics and AAPM. Med Phys 2008;35:5799-820.
11. Cole EB, Zhang Z, Marques HS, et al. Impact of computer-aided detection systems on radiologist accuracy with digital mammography. AJR Am J Roentgenol 2014;203:909-16.
12. Lehman CD, Wellman RD, Buist DS, et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Intern Med 2015;175:1828.
13. Taylor P, Potts HW. Computer aids and human second reading as interventions in screening mammography: two systematic reviews to compare effects on cancer detection and recall rate. Eur J Cancer 2008;44:798-807.
14. Elmore JG, Carney PA. Computer-aided detection of breast cancer: has promise outstripped performance? J Natl Cancer Inst 2004;96:162-3.
15. Gur D, Sumkin JH, Rockette HE, et al. Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system. J Natal Cancer Inst 2004;96:185-90.
16. Beam C, Sullivan D. Variability in mammogram interpretation. Adm Radiol J 1996;15:49-50. 52.
17. Breast Cancer Surveillance Consortium. Smoothed plots of sensitivity (among radiologists finding 30 or more cancers), 2004-2008—based on BCSC data through 2009. Available at: http://www.bcsc-research.org/statistics/benchmarks/screening/2009/figure10.html. Accessed September 23, 2017.
18. Gross CP, Long JB, Ross JS, et al. The cost of breast cancer screening in the Medicare population. JAMA Intern Med 2013;173:220-6.
19. Azavedo E, Zackrisson S, Mejàre I, et al. Is single reading with computer-aided detection (CAD) as good as double reading in mammography screening? A systematic review. BMC Med Imaging 2012;12:22.
20. Sato M, Kawai M, Nishino Y, et al. Cost-effectiveness analysis for breast cancer screening: double reading versus single CAD reading. Breast Cancer 2012;21:532-41.
21. Trister AD, Buist DSM, Lee CI, et al. Will machine learning tip the balance in breast cancer screening? JAMA Oncol 2017;3:1463-4.
22. Boyd NF, Guo H, Martin LJ, et al. Mammographic density and the risk and detection of breast cancer. N Engl J Med 2007;356:227-36.
23. Wei L, Yang Y, Wernick MN, Nishikawa R. Learning of perceptual similarity from expert readers for mammogram retrieval. IEEE J Sel Top Signal Process 2009;1:53-61.
24. Moore GE. Cramming more components onto integrated circuits. Proceedings of the IEEE 1998;86:82-5.
25. Samala RK, Chan HP, Hadjiiski L, Helvie MA, et al. Mass detection in digital breast tomosynthesis: deep convolutional neural network with transfer learning from mammography. Med Phys 2016;43:6654-66.
26. Skaane P, Bandos AI, Gullien R, et al. Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population-based screening program. Radiology 2013;267:47-56.
27. Reiser I, Nishikawa R. Computerized mass detection for digital breast tomosynthesis. In: Suri JS, Rangayyan RM (eds). Recent advances in breast imaging, mammography, and computer-aided diagnosis of breast cancer, SPIE press, Bellingham, WA, pp. 409-28.

Ajay Kohli, MD, is from the Department of Radiology, Drexel University College of Medicine, Philadelphia, Pennsylvania. Saurabh Jha, MD, is from the Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia, Pennsylvania.

The authors have no conflicts of interest related to the material discussed in this article.

Ajay Kohli, MD: 2930 Chestnut St # 604, Philadelphia, PA 19104; e-mail: akohli23@gmail.com.