The Breast 49 (2020) 254–260

Contents lists available at ScienceDirect

The Breast

journal homepage: www.elsevier.com/brst

Stand-alone artificial intelligence - The future of breast cancer screening?

Ioannis Sechopoulos [a, b, *], Ritse M. Mann [a, c]

[a] Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Geert Grooteplein 10, 6525 GA, Nijmegen, the Netherlands
[b] Dutch Expert Centre for Screening (LRCB), Wijchenseweg 101, 6538 SW, Nijmegen, the Netherlands
[c] Department of Radiology, Netherlands Cancer Institute (NKI), Plesmanlaan 121, 1066 CX, Amsterdam, the Netherlands

ARTICLE INFO

Article history:
Received 28 November 2019
Received in revised form
17 December 2019
Accepted 21 December 2019
Available online 2 January 2020

Keywords:
Artificial intelligence
Deep learning
Mammography
Screening

ABSTRACT

Although computers have had a role in interpretation of mammograms for at least two decades, their impact on performance has not lived up to expectations. However, in the last five years, the field of medical image analysis has undergone a revolution due to the introduction of deep learning convolutional neural networks – a form of artificial intelligence (AI). Because of their considerably higher performance compared to conventional computer aided detection methods, these AI algorithms have resulted in renewed interest in their potential for interpreting breast images in stand-alone mode. For this, first the actual capability of the algorithms, compared to breast radiologists, needs to be well understood. Although early studies have pointed to the comparable performance between AI systems and breast radiologists in interpreting mammograms, these comparisons have been performed in laboratory conditions with limited, enriched datasets. AI algorithms with performance comparable to breast radiologists could be used in a number of different ways, the most impactful being pre-selection, or triaging, of normal screening mammograms that would not need human interpretation. Initial studies evaluating this proposed use have shown very promising results, with the resulting accuracy of the complete screening process not being affected, but with a significant reduction in workload. There is a need to perform additional studies, especially prospective ones, with large screening data sets, to both gauge the actual stand-alone performance of these new algorithms, and the impact of the different implementation possibilities on screening programs.

Introduction

The use of mammography for breast cancer screening in the general population has resulted in widespread use of this imaging technology. To provide just two examples, an estimated 33 million screening mammograms are performed in the US every year [1], while in the Netherlands, over one million such tests are performed yearly [2]. In the Netherlands, as in many countries throughout Europe, a national screening program has been established by the government. As part of this program, all women between the ages of 50 and 75 are invited for a free screening digital mammography exam every two years. As is common in Europe, these screening mammograms are double-read by two radiologists specifically trained and certified for reading screening mammograms, independently, blinded to each other's interpretation. If the two radiologists arrive at disparate decisions on the case, a third reader arbitrates the case, providing the final decision of whether to recall the woman or not for diagnostic work-up. This arbitration occurs in about 2% of the cases [3]. Thus, altogether, over two million mammographic interpretations are performed within the national screening program alone, every year, in the Netherlands. As a result of these interpretations, about 2.4% of screened women are recalled due to a suspicious finding [2], which, after work-up and biopsy, result in 6.8 cancers diagnosed per thousand women screened. In other words, over 97% of screening mammograms are interpreted as normal, and cancer is detected in less than 1% of examinations.

Although the introduction of screening mammography has contributed to an important reduction in breast cancer-related mortality [4], the very nature of population screening results in an extremely high number of negative tests. Given that this test is

* Corresponding author. P.O. Box 9101, Route 766, 6500 HB, Nijmegen, the Netherlands.
E-mail addresses: ioannis.sechopoulos@radboudumc.nl (I. Sechopoulos), ritse.mann@radboudumc.nl (R.M. Mann).

https://doi.org/10.1016/j.breast.2019.12.014

interpreted by highly specialized, human, experts, there would be various advantages in automating this process, at least partly, if at all possible.

Of course, various questions need to be answered convincingly before automation of interpretation of screening mammograms can be introduced. Is the image interpretation technology good enough? Are we aiming to automate screening interpretation fully or only partially? If the latter, then what is the best combination of computer vs. human interpretation? Why do this? What would the advantages be? Finally, and most importantly, would it be ethical to automate this process and therefore refrain from human reading in some or all of the mammograms?

The past: conventional computer aided detection algorithms

Stand-alone computer interpretation of mammograms, which would allow for automated screening, is not, strictly speaking, equivalent to Computer Aided Detection (CADe) and Computer Aided Diagnosis (CADx), but it is based on the same technology. Conventional CADe/CADx methods are based on the algorithms recognizing suspicious lesions due to their matching characteristics specified by the algorithm programmers. In other words, the programmer teaches the computer what a malignant mass or calcification cluster distribution looks like, by describing to the computer the different features that distinguish a malignant mass, for example, from normal tissue or a benign mass. CADe algorithms then search the image for areas that contain (some of) the programmed-in features, and, if the score for a possible finding is high enough, that location is marked as suspicious on the image. In CADx, the algorithm is input the region of the image that contains a suspicious finding, and the software evaluates only that area for determining how much it contains features that match the pre-programmed signs of malignancy. Therefore, the output of CADx software is usually a score that relates to the probability of finding analyzed being malignant.

The goal of CADe/CADx is to aid the radiologist during their interpretation of the exam, in one of two ways. In the traditional use of CADe, after the radiologist interprets the case and arrives at a decision, he/she turns on the CADe marks, if any, and reviews them to ensure that none of the locations highlighted as suspicious are deemed actionable findings. The use of CADe in this manner aims to reduce false negatives due to lesions being overlooked by the radiologist during his/her search for findings. During the lab-setting testing of CADe, using retrospective studies with selected data sets, the performance of the CADe algorithms was found to be very promising [5–8]. This led to the rapid introduction of these algorithms in the clinic, especially in the US. However, later studies that evaluated the actual impact on clinical performance of CADe introduction showed that its promise was not fulfilled [9,10]. The high rate of false positive marks, typically far above one per image, leads to a reduction in specificity. This is because, compared to the very low actual prevalence of cancer, if the radiologist is swayed to accept only a small percentage of marks as actionable, the rate of negative recalls will increase, while the chances that a overlooked cancer is found thanks to the algorithm are small [11]. In addition, the CADe programs seem to be used in a different fashion than intended, as in practice sensitivity also decreases. This is most likely caused by the fact that areas not marked by the CAD system are more easily dismissed.

Another scenario for use of computer algorithms in interpretation of screening mammograms uses a hybrid CADe/CADx approach. In this setting, CADe marks are not automatically shown, but during interpretation of the case, the radiologist may query the computer on its opinion of an area already identified by him/her as potentially being suspicious. Only at this point are hidden CADe marks displayed, if available, usually together with a likelihood of malignancy for the queried area. In trying to arrive at a decision to recall or not based on the potential finding, the radiologist thus seeks a second opinion, and instead of asking his/her colleague, asks the computer for help. Such use case would have a positive impact on both sensitivity and specificity, but is limited in magnitude, since it can only affect the outcome in difficult border-line cases, where the radiologist would actually use the algorithm. In normal cases with no detected findings, or cases with findings in which the radiologist is certain of his/her decision (even if that decision is actually wrong), then the algorithm would not be used, and therefore it could not have an impact [12].

The main limitation of conventional CADe/CADx algorithms is the need for the characteristics of malignancy to be specified by the programmer. This is a challenging, cumbersome, subjective process, which is inherently limiting in terms of amount of information that can be provided to the algorithm by the programmer. The fact that hardly any of the conventional algorithms ever approached the performance of breast radiologists implies that it is very difficult to capture all the signs of breast cancer recognized by humans in handcrafted mathematical formulations. The introduction of the new generation of artificial intelligence (AI) algorithms for computer-interpretation of mammographic images solves this limitation.

Introduction of AI

This new generation of algorithms is, for the most part, based on deep learning convolutional neural networks (CNNs). These networks consist of a series of simple mathematical operations, grouped in *layers*, that sequentially first break down the image being analyzed into ever-smaller components. Depending on the network, this might be followed by putting back together these components into an image of the same or similar size as the original input, learning the spatially related characteristics that determine a pre-defined ground truth at every level. Depending on the nature and intent of the algorithm, the output could be a single binary positive/negative decision, or a probability of malignancy score for each lesion detected. The latter might include a heat map highlighting the area(s) that contained the characteristics used to come to the final classification. In mammograms containing breast cancer this usually means the cancer itself, but as is the case for human interpretation, the final classification might also be influenced by retraction patterns or skin characteristics.

The design of the network, in terms of the number and type of layers and of the connections among them define what the network outputs, which is still defined by programmers. However, the multiple layers of mathematical operators also include thousands of numerical parameters, denoted *weights*, that have to be set to specific values. The values of these thousands of weights are what define the characteristics searched for in the images, and how these different characteristics influence the final output. These values are determined by the software itself, during the training process. For this, the network learns by repeatedly being shown example images and the correct output for each image, and the weight values are adjusted each time so that the network arrives to an output ever-closer to the correct one.

This is the major difference between conventional CADe/CADx algorithms and the new generation of AI-based algorithms; the computer learns to distinguish malignant findings from normal tissue or benign findings by itself, not because it is taught what a malignant finding looks like by the programmer. This results in a performance that is substantially improved compared to that of conventional algorithms [13,14], since the process is objective and data-driven. In other words, it is the network that determines what really differentiates a malignant lesion in an image from a benign one, in a way that descriptive code written by a programmer would never be able to achieve.

Interestingly, the exact same network can be trained, by use of corresponding examples, to detect a tumor in a mammogram, a cat in a photograph, or a fault in a photograph of fabricated piece of machinery [15–18].

Achieving stand-alone computer interpretation: when is it feasible?

As discussed by Doi years before the introduction of AI for medical image analysis, and therefore before it was feasible, the requirements on stand-alone computer interpretation of mammograms are very different from those for the use of CADe/CADx algorithms [19]. For the latter, as already discussed, and its name implies, the interpreting radiologist is *aided* by the computer. In one implementation the radiologist can discard the CADe marks as false positives, while with the interactive software the radiologist might not even turn it on for cases in which he/she is sure of his/her decision. As a result, the final performance is determined by the radiologist.

Of course, a better-performing CAD would most probably result in a better outcome, in most circumstances. But, at least in theory, this is not necessarily true, if the radiologist decides to discard or ignore the CAD opinion the times it is correct, then the contribution of the software to the final performance could be limited. In addition, if, for example, the CADe performs at the same level as the radiologist, but with a perfect overlap in what it detects and what it misses, then it actually would not contribute to the final outcome at all. In the other extreme, if the performance of the CADe is rather poor but it is even sometimes correct, and the radiologist is patient enough (a tall order to ask), then the final outcome might, even just a few times, be improved, because some cancers missed by the radiologist are pointed out by the CADe algorithm. Therefore, in an idealized world in which the radiologist does not get frustrated or loses confidence in the computer, but evaluates each and every one of its marks, the final outcome could be improved by the incorporation of even a sub-optimal CAD algorithm. Still, even with a good CAD algorithm, the gain that can be achieved by the use of an algorithm as an aid is relatively limited. The use of AI as a stand-alone reader of mammograms might have a much larger impact on the associated workload of screening, and, if the system is really good, on the quality of the screening program.

On the other hand, conventional wisdom states that for stand-alone computer interpretation of medical images to be acceptable and incorporated into clinical use, the computer has to approach, or exceed, human performance. The obvious expectation is that if any radiologist is to be, at least partly replaced by computers, then performance cannot be lower. In general, we should not accept an increase in missed cancers or recalling additional healthy women due to having incorporated a computer to read any number of screening mammograms.

However, there are exceptions to this requirement. In the first place, sub-par performance of computer interpretation might be the only option to read screening mammograms. That is, in a developing country, perhaps the human or economic resources are simply not available, and therefore reading a substantial number of screening exams could be only possible with computers. In fact, this could be the case in the near future in some developed countries, where the shortage of radiologists is increasing [20]. In these cases, reduction of the total workload by automated interpretation, making widespread screening feasible, with a reasonable performance, even if below that of a well-trained screening radiologist, would be better than no screening at all.

Another scenario in which stand-alone performance below that of radiologists could still be acceptable would be if that incorporation would make introduction of other improvements feasible. Specifically, let's consider the scenario of replacing digital mammography as the screening imaging modality for a higher-performing technique, e.g., digital breast tomosynthesis (DBT). Over a number of prospective screening trials, DBT has been found to result in higher detection performance, but with a doubling of the reading time [21–29]. Therefore, introduction of DBT-based screening, which has already taken place in the US, where screening is institution-based [30], is also of intense interest in countries where screening is regional- or national-program based. However, if the resources, economic and/or human, are not available to introduce DBT screening due to its increase in reading time, again, perhaps DBT + a sub-human stand-alone computer, in at least a portion of screening, results in an overall better performance than mammography + only human interpretation.

It is in the spirit of these realities and possibilities that, since the introduction of the new generation of algorithms for medical image interpretation, the potential for stand-alone interpretation of breast screening images is again being studied and discussed. Although, currently, we have evidence that the performance of these algorithms is approaching that of breast radiologists, there is yet no definitive evidence that this level has been reached.

Stand-alone AI performance: how does a computer compare to a human radiologist?

So, where do we stand? How good are the current cutting-edge AI algorithms in interpreting breast images? A few studies have evaluated the stand-alone performance of AI algorithms, some of them commercial, in reading digital mammography or DBT images and compared it to that of radiologists [31–34]. All involve retrospective reading of enriched case sets, using receiver operating characteristic (ROC) methods.

In the largest study so far, Rodriguez Ruiz et al. aimed to compare the performance of an AI algorithm for interpretation of digital mammograms, using as varied a data set as possible [31]. To be able to perform this comparison including a multitude of conditions, the investigators gathered nine different data sets that included the mammographic images and the interpretations of these images from multiple breast radiologists. These sets had been assembled and used in other, previously-published ROC observer studies, in which the original investigation was to compare digital mammography to a different modality (mostly, DBT). Ignoring the data from the competing modality, Rodriguez Ruiz et al. gathered the digital mammography information (images and reader probability-of-malignancy ratings), and compared the reader detection performance to that of the AI algorithm. In total, the nine sources of data included over 2600 exams, acquired with systems from four different vendors, installed in both the US and countries across Europe, which included over 650 cancer cases. The total number of interpreting radiologists, that involved breast radiologists from both the US and across Europe, was 101, which, considering the number of exams each radiologist read, resulted in an analysis of over 28,000 interpretations. After obtaining the probability-of-malignancy present for each case estimated by a commercial AI system, the authors compared the resulting ROC curves, both between the AI and the average of the radiologists for all nine data sets pooled together, and against each individual radiologist. The area under the ROC curve (AUC) of the AI system was found to be non-inferior to that of the average radiologist (AI AUC: 0.840 [95% CI: 0.820–0.860] vs. radiologist AUC 0.814 [0.787–0.841]) (Fig. 1). Compared to individual readers, again in terms of AUC, the AI outperformed 61% of the radiologists.

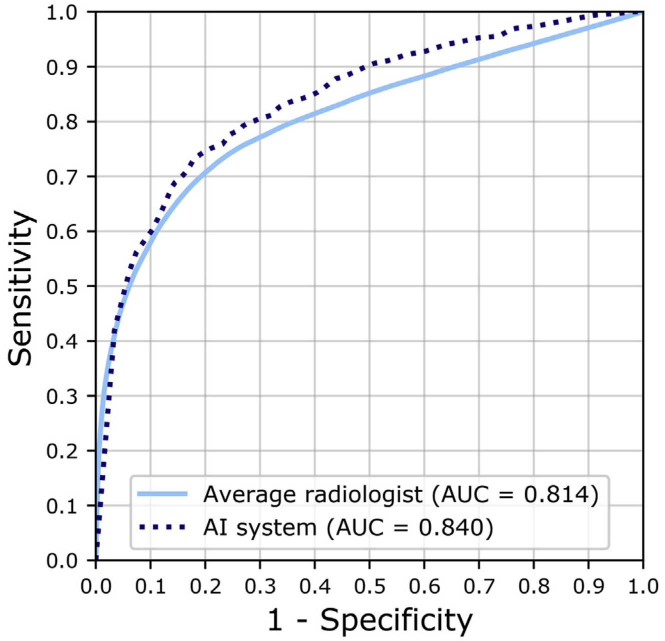

Fig. 1. ROC curves of the reader-averaged radiologist (solid line) and of a commercial AI system (dashed line) for detection of cancer in over 2600 digital mammograms. The AI system performance was found to be non-inferior to that of the radiologists. Adapted from Rodriguez Ruiz et al. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. J Natl Cancer Inst. 2019; 111 (9):916–922.

However, interestingly, the AI algorithm performed worse than the highest-performing radiologist of all of the nine data sets.

In a slightly more recent study, Conant et al. reported the stand-alone performance of a different commercial AI system when analyzing a DBT data set consisting of 260 cases acquired at seven different sites, with systems from a single vendor, all across the US [32]. The performance of 24 radiologists interpreting this same data set without access to the AI system was also reported. The AI system was set to operate at a high sensitivity (actual sensitivity achieved with the evaluated data set was 91% [95% CI: 81%–96%]), which resulted in a specificity of 41% [95% CI: 34%–48%]. The mean sensitivity and specificity of the radiologists was found to be 77.0% (min/max: 38.5%–93.8%) and 62.7% (22.1%–84.6%), respectively. Although the AUC of the AI system was not specified in this publication, the ROC curves of the AI system and of the average of the radiologists can be seen to be very similar (Fig. 2).

These two examples of comparisons of the stand-alone performance of commercially available AI systems to that of radiologists interpreting digital mammography and DBT images show the potential for these systems to achieve human-like performance for this clinical task. The particular strength of the Rodriguez Ruiz et al. study is the comprehensive nature of the data, covering multiple image acquisition systems, and the images interpreted by radiologists from various countries from both sides of the Atlantic. At the same time, the Conant study showed that an AI system for DBT seems to be at similar performance levels compared to radiologists as systems for digital mammography.

However, both of these studies do not allow for the conclusion that stand-alone AI interpretation of x-ray breast images is at a par with screening breast radiologists in the real screening setting. In the first place, the data sets were enriched, resulting in a much higher prevalence of cancer than that found in a screening data set. In addition, they were all retrospectively read, in a research setting. Furthermore, both AI algorithms are not able to compare the current images to any prior examination images. Although this comparison is an important source of information for the interpreting radiologist, AI systems are for now unable to take this information into account. In the data sets included in the Rodriguez Ruiz et al. study, some of the human reader results included providing the radiologists with the prior images, and some did not. Interestingly, a difference in how the AI results compared to these data sets vs. how it compared to the data sets that excluded prior images, cannot be easily seen. The Conant et al. study excluded the use of priors. Therefore, although the exclusion of priors allows for a comparison of performance when given the same amount of information to both the AI and the radiologists, this may underestimate the accuracy of radiologists during real screening.

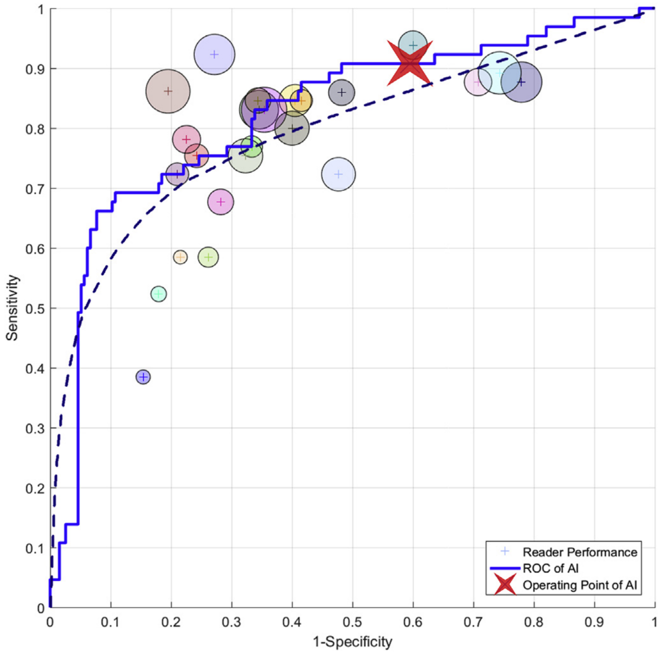

Fig. 2. ROC curves of the reader-averaged radiologist (dashed line) and of a commercial AI system (solid line) for detection of cancer in over 260 digital breast tomosynthesis images. The sensitivity/specificity pairs for each radiologist are shown with circles, whose size represents the average reading time per case. The red star indicates the operating point of the AI system. Adapted from Conant et al. Improving Accuracy and Efficiency with Concurrent Use of Artificial Intelligence for Digital Breast Tomosynthesis. Radiology: Artificial Intelligence. 2019; 1 (4):e180096. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

As mentioned, the AI algorithm evaluated by Rodriguez Ruiz et al. did not achieve the same performance as any of the best radiologist in each set included in the study. Similarly, two of the 24 radiologists in the Conant et al. study had a higher sensitivity/specificity pair than that obtained by the AI system. In the absence of information not available to the AI system, such as prior images, further improvements of the AI algorithms, in order to achieve, or even surpass, the performance of the best human readers, are theoretically possible. Due to the self-learning nature of this technology, continued training of the network with additional cases can continuously increase its performance. Of course, performance improvement with additional cases undergoes diminishing returns, but, given the variability across images, in terms of system vendors, acquisition techniques, image post-processing algorithms and versions, lesion variability, etc., the training with a large number of cases across all these variations may potentially further improve performance. This helps highlight the main difference, and most powerful aspect, of this new AI technology compared to conventional CAD algorithms; improvement in performance can be achieved by simply continuing the training of the network with additional data, whereas any improvement in conventional CAD software would require human revision of the software with new or improved algorithms.

The one algorithmic change that could be expected to improve the performance of AI systems would be the ability to consider temporal information within a case, by being able to compare the current images to priors, as done by human readers. However, somewhat surprisingly, the introduction of such an improvement in one AI algorithm did not result in any significant performance improvement [35]. Of course, further work on this aspect is needed.

Further performance evaluation studies

Beyond any further improvements in the AI systems that might take place in the future, further evaluation studies to determine the actual stand-alone performance of AI systems when reading breast cancer screening images are needed. Specifically, large scale studies, comparing the performance of these systems when reading actual screening data sets, containing a large number of cases, need to be performed. In addition, the AI performance with these data sets needs to be compared to the actual, prospective, radiologist readings at screening. Only with such studies will the potential for these algorithms to have a role as a stand-alone option in screening for breast cancer be clear.

Assuming that these studies will also show at least equal overall performance of AI systems and radiologists, it is safe to assume that neither will detect all cancers present in the screening set, and that they will likely not detect exactly the same cancers either. Therefore, it would still be important to determine if there are any differences between the biological and molecular profiles of the cancers detected by AI systems and those detected by interpreting radiologists. If AI systems detected, overall, an equal or higher number of cancers (assuming same specificity), but there tended to be an increase in the number of indolent cancers detected, while the fraction of more aggressive cancers declines, then all that would be gained would be an increase in overdiagnosis, and a decrease in the benefit of the screening program. To date, the retrospective enriched data sets used for AI performance evaluation studies have been too small to investigate this matter. Future large-scale studies with actual screening data sets, therefore, need to include an analysis of the characteristics of the human- and AI-detected tumors to be able to estimate the actual impact on patient outcomes of using these systems.

Options for use as stand-alone

But what could be the role of AI algorithms, when used in stand-alone mode, in breast cancer screening? The answer that screening will be performed automatically by AI algorithms, with no role by human radiologists seems currently still too simplistic, albeit this may be achieved when AI algorithms clearly outperform even the best human readers. For the moment, published studies have proposed more sophisticated strategies for incorporating automated reading of images into screening.

Currently, the most common approach being investigated is the automated identification of normal cases that either do not need to be evaluated by radiologists at all, or that could be single read instead of double read (in the common double reading screening scenario in Europe) [36–39]. In this light, Rodriguez Ruiz et al., in another study with the same dataset mentioned earlier, evaluated the impact on the overall outcome when the cases with the lowest AI-generated probability-of-malignancy scores are pre-designated as normal and therefore not interpreted by radiologists. In a more aggressive setting, with a threshold for human-reading that results in pre-designating about 50% of cases as normal, the investigators found that 7% of cancer cases would be incorrectly flagged as normal. With a more conservative setting that results in a reduction in the workload of 20%, only 1% of cancer cases would be mislabeled. At the same time, these two thresholds would reduce the false positive recalls by 27% and 5%, respectively. Therefore, the slight loss in sensitivity is compensated by the gain in specificity, resulting in an unchanged AUC. This, of course, means that adjustment of the operating point after pre-selection could result in equal sensitivity and specificity as without pre-selection, but with a considerably reduced workload.

Lång et al. [38], using the same AI system but on a subset of cases from the prospective Malmö Breast Tomosynthesis Screening Trial [24,25], found very similar results. Interestingly, radiologist review of the cancer cases that the AI scores incorrectly labelled as normal determined that the missed lesions were clearly visible. This points to the possibility that there is still "low-hanging fruit" that would allow for improvement in the stand-alone performance of AI algorithms, reducing the proportion of mislabeled cancer cases.

In a third study evaluating the same strategy, Yala et al. developed an AI algorithm specifically to identify normal cases to be used for this pre-designation strategy, and evaluated what its impact on performance would be [37]. The investigators found that the use of this software would result in a reduction of 20% in the workload, with non-inferior sensitivity and a statistically significant increase in the specificity, from 93.5% to 94.2%. If these results were to be confirmed, though slight, this scenario would increase the overall performance in terms of accuracy, in addition to the reduction in workload.

Finally, Kyono et al. also developed an AI algorithm optimized for this purpose [39]. Using a portion of the TOMMY tomosynthesis trial dataset [28], the authors determined the proportion of non-cancer cases that would be correctly labelled as normal by their developed algorithm. Interestingly, they evaluated this using data sets enriched to three different cancer prevalence levels: 15%, 5%, and 1%. As is necessary for this task, maintaining a very high negative predictive value (NPV), the system was able to reduce the workload to be human-read by 34% in the set with highest prevalence, and by 91% at a screening-like 1% prevalence.

With these studies, the current scenario for incorporation of stand-alone AI in screening mammography would result in an equal performance to a completely human-read screening scenario, but with a reduction in workload. Such a reduction, be it by 20%, 50%, or even 90%, in number of cases to be reviewed, could still result in an improvement in the screening performance in the cases read by humans in ways that are not captured by these four studies.

In the first place, since these studies simulated outcomes using retrospective reading results, the impact on radiologist performance of reading a pre-selected case set is ignored. Would radiologists' performance vary when they know that an AI algorithm scored the cases they are reading as being above a certain suspiciousness threshold? Although the actual cancer prevalence within the human-read cases would, of course, be increased, in absolute terms the prevalence would still be low. Would the reduction in workload allow for an increase in the reading time per case, and therefore improved performance, or would the pre-designation of the cases as at somewhat more suspicious cause an increase in the false positive recalls? Answer to these questions could only be arrived at with prospective studies, in which reader performance after pre-selection could be gauged.

More indirectly, incorporation of an AI technology that allows for a reduction in workload could allow for introduction of other imaging technology for screening. Specifically, as discussed earlier, reduction in the cases to be human-read would ameliorate the increase in reading time required by DBT reading. This would allow for replacement of mammography-based screening with DBT, presumably resulting in improved performance, as shown by the prospective DBT screening trials. Hence, incorporation of an AI technology that results in equal performance but a reduction in the workload, could indirectly impact the outcomes of screening.

Ethical issues

Can we do this? Even if the answer from a technical point of view is yes, how about the answer from an ethical point of view? Is it acceptable, or will the screened population accept, that some of their images will not to be reviewed by any human? Anecdotally, discussion with women that regularly participate in screening programs, including breast cancer survivors and volunteer patient advocates, have expressed their support for mechanisms that result in improved outcomes, whatever their nature. An AI algorithm that is stable and always performs at the same high level may, in the end, achieve this. It can be safely assumed that an AI algorithm does not have attention deficits due to disturbing phone calls or the need for a cup of coffee, albeit, of course, quality assurance protocols will need to be devised.

Importantly, automated analysis of health information is already prevalent in current healthcare. Red blood cell count, for example, during a blood test, is not performed by a human using a microscope to count each and every single cell in a blood sample, of course. Rather, a machine performs the count. We are all used to this, and thinking about still doing such tasks by hand seems unthinkable. Although interpreting a screening mammography examination may be a more complex task than counting cells in blood, performing the latter task automatically was years ago also regarded as challenging. Hence, once the performance of algorithms interpreting mammography images is shown to be equivalent, or superior, to that of screening radiologists, perhaps with time applying this technology to improve screening outcomes will seem normal and expected.

Of course, a discussion of the various non-technical aspects involved in automated screening interpretation will first be necessary, before such incorporation is possible. These involve not only ethical issues, but also, legal, social, and even economic considerations. For a comprehensive discussion on these topics, the reader is referred to a recent review of these issues by Carter et al., included in this special issue [40].

Conclusions

We seem to be at the doorstep of a revolution in breast cancer screening. The developments in AI interpretation of medical images over the last few years seem to have opened the door for incorporating stand-alone computer interpretation of images into breast cancer screening programs. Current evidence shows that these algorithms are approaching, if not yet have reached, expert human performance, although definitive studies that compare their performance to actual screening results are not yet available. If and when such performance levels are achieved and demonstrated, it seems feasible that, at least, an important reduction in the workload for human interpretation could be achieved, with no decrease in performance. Even if future improvements are not achieved, and therefore the impact on performance discussed earlier remains unchanged, there might be subsequent changes down- and upstream that could result in an improvement in the quality of the screening program. A reduction in workload with an unchanged AUC could allow more time for interpreting radiologists to spend on the cases that do need human review, presumably improving accuracy. In addition, a reduction in human workload could ease the challenge of transitioning to a more accurate but slower to interpret, imaging technology, such as DBT, again resulting in an overall improvement in performance.

Further improvements in algorithms and training sets, combined with evidence from more definite, prospective, actual-screening-prevalence trials, could finally usher in the age of computers having a direct role in breast cancer screening. The next few years will be very dynamic in this field.

Funding source

No funding has been received for this work.

Ethical approval

No ethical approval was required for this work.

Declaration of competing interest

ScreenPoint Medical, a company that develops and markets AI systems for breast imaging, is a spinoff company of the Radboud University Medical Center. While there is no financial relationship between ScreenPoint Medical and the authors, they do work closely with its CEO, who is also a Professor at their department. There is a master research agreement (MRA) between Radboud University Medical Center (Department of Radiology and Nuclear Medicine) and ScreenPoint Medical that describes terms of cooperation. The authors have research and speaking agreements, through their institution, with Siemens Healthcare.

References

[1] Comparison of screening recommendations indicates annual mammography starting at 40 prevents most cancer deaths | Wiley News Room – Press Releases, News, Events & Media. https://newsroom.wiley.com/press-release/cancer/comparison-screening-recommendations-indicates-annual-mammography-starting-age-. Accessed November 21, 2019.

[2] National Evaluation Team for Breast cancer screening in the Netherlands (NETB). NETB Monitor 2014 - nation-wide breast cancer screening in The Netherlands, Results 2004 -2014. Rotterdam: Erasmus MC and Radboudumc; 2019 Jan.

[3] Klompenhouwer EG, Duijm LEM, Voogd AC, et al. Variations in screening outcome among pairs of screening radiologists at non-blinded double reading of screening mammograms: a population-based study. Eur Radiol 2014;24(5):1097–104.

[4] Plevritis SK, Munoz D, Kurian AW, et al. Association of screening and treatment with breast cancer mortality by molecular subtype in us women, 2000-2012. J Am Med Assoc 2018;319(2):154–64.

[5] Birdwell RL, Ikeda DM, O'Shaughnessy KF, Sickles EA. Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection. Radiology 2001;219(1):192–202.

[6] Warren Burhenne LJ, Wood SA, D'Orsi CJ, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. Radiology 2000;215(2):554–62.

[7] Freer TW, Ulissey MJ. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. Radiology 2001;220(3):781–6.

[8] Destounis SV, DiNitto P, Logan-Young W, Bonaccio E, Zuley ML, Willison KM. Can computer-aided detection with double reading of screening mammograms help decrease the false-negative rate? Initial experience. Radiology 2004;232(2):578–84.

[9] Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. N Engl J Med 2007;356(14):1399–409.

[10] Lehman CD, Wellman RD, Buist DSM, Kerlikowske K, Tosteson ANA, Miglioretti DL. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Int Med 2015;175(11):1828–37.

[11] Ikeda DM, Birdwell RL, O'Shaughnessy KF, Sickles EA, Brenner RJ. Computer-aided detection output on 172 subtle findings on normal mammograms previously obtained in women with breast cancer detected at follow-up screening mammography. Radiology 2004;230(3):811–9.

[12] Hupse R, Samulski M, Lobbes MB, et al. Computer-aided detection of masses at mammography: interactive decision support versus prompts. Radiology

2013;266(1):123–9.

[13] Fotin SV, Yin Y, Haldankar H, Hoffmeister JW, Periaswamy S. Detection of soft tissue densities from digital breast tomosynthesis: comparison of conventional and deep learning approaches. In: Proceedings of SPIE 9785. International Society for Optics and Photonics; 2016. 97850X.

[14] Kooi T, Litjens G, van Ginneken B, et al. Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal 2017;35:303–12.

[15] Samala RK, Chan H-P, Hadjiiski LM, Helvie MA, Cha KH, Richter CD. Multi-task transfer learning deep convolutional neural network: application to computer-aided diagnosis of breast cancer on mammograms. Phys Med Biol 2017;62(23):8894–908.

[16] Huynh BQ, Li H, Giger ML. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. JMI 2016;3(3): 034501.

[17] Aboutalib SS, Mohamed AA, Berg WA, Zuley ML, Sumkin JH, Wu S. Deep learning to distinguish recalled but benign mammography images in breast cancer screening. Clin Cancer Res 2018;24(23):5902–9.

[18] Becker AS, Marcon M, Ghafoor S, Wurnig MC, Frauenfelder T, Boss A. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. Investig Radiol 2017;52(7): 434–40.

[19] Doi K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Comput Med Imag Graph 2007;31(4): 198–211.

[20] The breast imaging and diagnostic workforce in the United Kingdom | The Royal College of Radiologists. https://www.rcr.ac.uk/publication/breast-imaging-and-diagnostic-workforce-united-kingdom. Accessed April 30, 2019.

[21] Skaane P, Bandos AI, Niklason LT, et al. Digital mammography versus digital mammography plus tomosynthesis in breast cancer screening: the oslo tomosynthesis screening trial. Radiology 2019;219(1):23–30.

[22] Hofvind S, Hovda T, Holen AS, et al. Digital breast tomosynthesis and synthetic 2D mammography versus digital mammography: evaluation in a population-based screening program. Radiology 2018;287(3):787–94.

[23] Østerås BH, Martinsen ACT, Gullien R, Skaane P. Digital mammography versus breast tomosynthesis: impact of breast density on diagnostic performance in population-based screening. Radiology 2019;293(1):60–8.

[24] Zackrisson S, Lång K, Rosso A, et al. One-view breast tomosynthesis versus two-view mammography in the Malmö Breast Tomosynthesis Screening Trial (MBTST): a prospective, population-based, diagnostic accuracy study. Lancet Oncol 2018;19(11):1493–503.

[25] Lång K, Andersson I, Rosso A, Tingberg A, Timberg P, Zackrisson S. Performance of one-view breast tomosynthesis as a stand-alone breast cancer screening modality: results from the Malmö Breast Tomosynthesis Screening Trial, a population-based study. Eur Radiol 2016;26(1):184–90.

[26] Bernardi D, Macaskill P, Pellegrini M, et al. Breast cancer screening with tomosynthesis (3D mammography) with acquired or synthetic 2D mammography compared with 2D mammography alone (STORM-2): a population-based prospective study. Lancet Oncol 2016;17(8):1105–13.

[27] Ciatto S, Houssami N, Bernardi D, et al. Integration of 3D digital mammography with tomosynthesis for population breast-cancer screening (STORM): a prospective comparison study. Lancet Oncol 2013;14(7):583–9.

[28] Gilbert F, Tucker L, Gillan M, et al. The TOMMY trial: a comparison of TOMosynthesis with digital MammographY in the UK NHS breast screening programme. Health Technol Assess 2015;19(4).

[29] Romero Martin S, Raya Povedano JL, Cara Garcia M, Santos Romero AL, Pedrosa Garriguet M, Alvarez Benito M. Prospective study aiming to compare 2D mammography and tomosynthesis + synthesized mammography in terms of cancer detection and recall. From double reading of 2D mammography to single reading of tomosynthesis. Eur Radiol 2018;28(6):2484–91.

[30] Friedewald SM, Rafferty EA, Rose SL, et al. Breast cancer screening using tomosynthesis in combination with digital mammography. J Am Med Assoc 2014;311(24):2499.

[31] Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. J Natl Cancer Inst 2019;111(9):916–22.

[32] Conant EF, Toledano AY, Periaswamy S, et al. Improving accuracy and efficiency with concurrent use of artificial intelligence for digital breast tomosynthesis. Radiology: Artif Intell 2019;1(4):e180096.

[33] Rodríguez-Ruiz A, Krupinski E, Mordang J-J, et al. Detection of breast cancer with mammography: effect of an artificial intelligence support system. Radiology 2018;290(2):305–14.

[34] Wu N, Phang J, Park J, et al. Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Trans Med Imaging 2019. https://doi.org/10.1109/TMI.2019.2945514.

[35] Kooi T, Karssemeijer N. Classifying symmetrical differences and temporal change for the detection of malignant masses in mammography using deep neural networks. JMI 2017;4(4):044501.

[36] Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. Eur Radiol 2019;29(9): 4825–32.

[37] Yala A, Schuster T, Miles R, Barzilay R, Lehman C. A deep learning model to triage screening mammograms: a simulation study. Radiology 2019;293(1): 38–46.

[38] Lång K, Dustler M, Dahlblom V, Andersson I, Zackrisson S. Can artificial intelligence identify normal mammograms in screening? European Congress of Radiology; 2019. Vienna, Austria.

[39] Kyono T, Gilbert FJ, van der Schaar M. Improving workflow efficiency for mammography using machine learning. J Am Coll Radiol 2020;17(1PA): 56–63.

[40] Carter SM, Rogers W, Win KT, Frazer H, Richards B, Houssami N. The ethical, legal and social implications of using artificial intelligence systems in breast cancer care. Breast 2020;49:25–32.