The effect on interval cancer rates of the use of an AI-based algorithm in mammography

Screening mammography in women of appropriate age is widely accepted as an important strategy in the continuing reduction of the overall mortality of breast cancer, which is nevertheless still one of the most common cancers in women. A constant effort in screening mammography is focused, not only on increasing sensitivity, but also on reducing the numbers of interval cancers, i.e. those breast cancers which are detected symptomatically in the interval after a negative finding during the screening mammography exam. There are more and more reports of the benefits of the application of artificial intelligence-derived algorithms for the detection of breast lesions. A recent publication from a prominent German regional breast cancer screening center described a retrospective study of the potential impact of one such AI-algorithm on the rate of interval cancers [1]. We wanted to find out more about the issue of interval cancers in general and the use of AI algorithms in particular, so we spoke to Dr. A Gräwingholt, head of breast radiology at the Paderborn regional breast cancer screening center in Paderborn, Germany.

Dr Axel Gräwingholt
Email:
axel.graewingholt@t-online.de

Q *Before we get into your study of the impact of AI-algorithms on interval cancer rates, please remind us of the current status and practice of organized screening mammography in Germany.*

Under the current guidelines in Germany, women aged 50 - 69 years are invited for screening every two years. To facilitate this there is a large network of a total of 89 regional screening centers spread throughout the country. Our regional screening center in Paderborn in the west of Germany serves a total population of about 85000 women but to minimise the distance that the women need to travel to receive their examination we actually have four screening sub-units in our catchment area. Women are invited to the facility closest to their home via a central invitation system which proposes a date and time. No-shows are re-invited after four weeks. Approximately 60 % of women accept their invitation, a level that is in the upper tertile of the nationwide acceptance rates. In our unit, typically we see 12 000 to 15 000 women each year, although the attendance figure for 2020 dipped temporarily: in April last year our center was closed for a month because of COVID 19 lock-down measures. However we have now caught up with the back-log and are back on track with women being re-invited for screening 22 - 26 months after their prior examination.

We are quite proud of the fact that, once women have had their first appointment, very few decide subsequently to drop out of the screening program. Of the few women who do drop out, the principal reasons are a wish to get the results at once and an unwillingness to wait for the double reading to be carried out. Before reaching the age criteria for the screening program, some women were used to having their gynecologist-requested mammogram followed immediately by supplemental screening such as ultrasound and are disappointed that the more general screening workflow does not permit this.

Q *How is your center equipped for screening and what is the typical workflow?*

We have two Senographe Essential mammography/tomosynthesis devices from GE Healthcare. In regard to work-flow, breast screening in Germany is carried out according to European guidelines under which an average risk of breast cancer appropriate for the age group is assumed. At present, there are no formalised recommendations in the European guidelines for supplemental screening because the evidence is considered as still not clear enough, although other factors, such as manpower and financial resource requirements or the still-unresolved issue of over-diagnosis may also play a role in this. Nevertheless, women who carry gene mutations associated with an increased predisposition for breast cancer are in practice offered more intensified screening at special centers. Since we're talking about supplemental screening, it's worth mentioning the fairly recent DENSE trial. The results of this trial were interesting in that they showed some benefit in increased cancer detection and decreased interval cancers for women in a screening program with high breast density who were then offered supplemental MRI. However for me a more interesting approach is to find a way to stratify women into different risk groups based not only on breast density but also on a range of objective measurements and then implement whatever modalities are appropriate and necessary to detect any cancers. There are promising AI products already developed and available for such risk stratification – I think this will be the future basis of more personalized screening strategies.

Q *What about the performance statistics of your screening center?*

We have a cancer detection rate of around 6/1000 and a recall rate of approximately 3%, both of which are within the ranges specified by the European Guidelines. Caution should always be used when making detailed comparison of screening performances from center to center since the policy in each area can vary quite substantially. For example, in our region we actually have quite a large number of gynecologists who refer women for regular mammograms before the age of 50. Thus, it is likely that many cancers are detected before the age at which the screening program starts. It is reasonable to assume that if these mammograms of women under 50 years of age had not been carried out, more cancers would have been detected in the first screening round.

Q *What are the recommended guidelines regarding acceptable sensitivity/maximum recall rates? In practice how are you informed of interval cancers ?*

As for recall rates, the recommendations stipulate that these should be between 3 and 5% depending on whether they refer to subsequent or first round screening respectively. Sensitivity should be greater than 80%. Recall rates in Europe and Germany tend to be much lower than in the U.S.A. due to the fact that in our screening programs there are quite strict rules about permissible recall rates. Another reason for lower recall rates in Europe compared to the U.S.A is the mandatory double reading in Europe as well as the consensus or arbitration systems applied in almost all European screening programs. We have a population-based approach in Europe, whereas the U.S.A. has an individualized approach.

We are informed of interval cancers by a separate cancer registry dealing with our region. In the cancer registry, their QA systems use record linkage processes to identify a list of potential cases of interval cancers.These cases are then sent to us for verification. In a second stage, we are then asked by the reference center for our region to investigate these interval cancers and particularly all prior examinations which had been dismissed as normal in the screening programs. We then assign these "normal" priors into four categories –True, Occult, Minimal Signs and Missed.

"True" means that the cancer did indeed develop in the interval period. "Occult" indicates that the lesion could still not be detected by mammography even in the interval. The "Minimal signs" category means that there were only minimal signs that could have been detected in the priors. The "Missed" category is self-explanatory.

In practice such in-depth investigation that we carry out after verification can be quite a challenge mainly because of the difficulty we occasionally have in getting hold of the medical records, the images, etc. Sometimes it is just not possible. Other difficulties can be the application of data protection rules and sometimes also the refusal of women to authorize the transfer of their diagnostic records and images to us.

Of course the discovery of interval cancers is first of all very stressful for the women concerned since they are naturally pre-occupied with questions such as the probable improved prognosis they would have had if their cancer had been detected in the screening round, an issue that is all the more stressful if it turns out that the prior exam in question was not normal but that the cancer had been missed. Another factor is that when women first hear about interval cancers they tend to investigate and read more about the subject. When they find out that in general the prognosis for interval cancers is poorer than for screen-detected cancers, they become frequently ,and understandably, more frightened.

Apart from the women involved, the rates of interval cancers are very important as a quality performance parameter for the screening center itself, since, put simply, they are a measure of the number of cancers that have been missed or misinterpreted. Consequently, interval cancers serve as a powerful incentive for the continuing education of the screening readers. In fact, it is not just interval cancers that the screening center should be concerned about — it is also important to quantify how many large cancers are found in subsequent screening rounds. It is probable that quite a few of such large cancers had also been missed in a prior screening round.

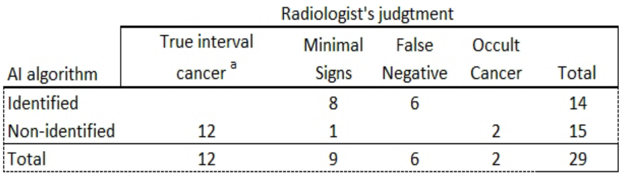
A complicating factor in the interpretation of the numbers of interval cancers is that the rates are also dependent on the extent of medical attention the women receive in the periods between scheduled screening rounds. For example if gynecologists carry out frequent ultrasound exams within the interval, the woman is more likely to have an interval cancer detected than women who do not undergo interval examinations. It should also be remembered that interval cancers are not necessarily always symptomatic.

However, despite these complicating factors, interval cancers are generally very important for the screening center since they are an unmistakable indicator of the quality level and trends of the screening unit and the screening program as a whole.

Q *And now let's get on to your recent retrospective study of the impact of AI-developed algorithm on interval cancers.*

The basic idea behind the study was to investigate the possibility of decreasing the rate of interval cancers by adding a procedure to the reading process, which will not radically change the screening process as a whole.The system we investigated was an an artifical intelligence algorithm called ProFound AI for 2D Mammography developed by iCAD, Inc. We wanted to find out whether the use of ProFound AI could detect certain lesions in the mammogram and draw the attention of the reader to them, so potentially reducing the number of interval cancers. We also wanted to know what impact this could have on the recall rate. A particular focus was on cases where the cancer had been missed or where the cancers showed only minimal signs. It is an inescapable fact that every radiologist — and sometimes even two radiologists — can miss cancers in the screening process for whatever reason, be it tiredness or disturbance during the reading process.

As part of the quality assurance process the interval cancers were each assigned to one of the four categories mentioned earlier.



	Radiologist's judgment				
AI algorithm	True interval cancer [a]	Minimal Signs	False Negative	Occult Cancer	Total
Identified		8	6		14
Non-identified	12	1		2	15
Total	12	9	6	2	29

Table 1. Number of interval cancers according to classification by radiologist and by the ProFound AI for 2D Mammography algorithm from iCAD.

"True" means that the cancer did indeed develop in the interval period. "Occult" means that the lesion could not be detected by mammography even in the interval. The "Minimal Signs" category means that there were minimal signs that could have been detected in the priors. The "Missed" category is self-explanatory.

a. In three cases the AI system identified a lesion in a breast area other than where the interval cancer was detected later.

Q *So, what were the results of the study? And their significance in terms of the use of the ProFound AI for 2D Mammography algorithm?*

We found out that the use of the algorithm enabled identification of lesions in the images of the prior examinations of many of the interval cancers. A summary of the results is shown in Table 1.

The software identified all of the lesions in the "Missed" category and almost all of the lesions categorized as haviing "Minimal Signs". Together the cancers detected through the use of ProFound AI accounted for 48% of the interval cancers.

> *"...by the addition of the ProFound AI algorithm as a support tool in the reading process, the benefits of finding more cancers outweigh the disadvantage of a (probably small) increase in recall rate..."*

The overall conclusion from our retrospective study is that by adding ProFound AI as a supportive tool in the reading process, the benefits of finding more cancers outweigh the disadvantage of a most likely small increase in recall rate.

Q *Broadening out from your study on interval cancers, what about the potential of AI-developed software support in breast imaging in general?*

Whether for reading purposes or for risk assessment I feel that adding AI derived products to screening programs will improve the quality of the program and therefore help in the detection of more cancers at the important stage, namely as early as possible. As far as reading is concerned, the reading time in 2D mammography is much less than that in tomosynthesis, so the impact of using AI-derived algorithms is of course not as great in 2D mammography as in tomosynthesis. If, one day, tomosynthesis were to be used in Europe as the baseline examination modality, it is difficult to see how this could be implemented in practice without the help of powerful AI-derived algorithms to reduce the reading time.

Q *Do you think that one day the European model of double reading could mutate to something like the US model of a single reader plus software as a second reader?*

Actually, I do think so, but with the reservation that the possibility should be maintained of having the results discussed in a consensus conference or at least through an arbitration reading in order to keep the recall rates low. From my experience in using these AI-derived algorithms, I have great confidence in the ability of the algorithms to detect cancerous lesions and they miss very few. In fact, I believe that we detect more cancers when we use an algorithm as a supportive tool. However if there are no consensus or arbitration processes, recall rates might rise a little too much.

So as a compromise and as my personal opinion I think that a model of "*Single reading + AI + consensus/arbitration on the findings*", could be the optimal approach to keep costs down but still have the same — or better — results than we have now when we use a double reading strategy. Of course, this model needs to be evaluated and validated in well-designed, large randomized control trials.

Q *Let's now broaden our conversation even more to the question of the reaction of radiologists to AI in general? Is AI perceived as a threat or a help?*

Unfortunately, I know of many radiology colleagues who fear that they could be replaced by AI procedures sometime in the future. This fear is partly justified, in that AI could easily replace the routine — and frequently boring — work many radiologists carry out in the routine reading and interpretation of image after image, day after day.

However, the time gained by the use of AI can profitably be used to increase the time the radiologist spends (and needs to spend) on complicated cases or with increased direct contact with patients. Even the very best radiologists can always improve their performance in the reading and detection of cancers or other diseases in general. The time needed for complicated cases could be made available by the use of AI.

My feeling is that in fact a new evolution of radiology as a medical subspecialty has only just begun. In the not-too-distant future, I foresee radiologists using all the various algorithms available for detection, prediction, monitoring and whatever other applications may be developed. Thus, the future role of the radiologist will be markedly different from the current one of purely interpreting images. I see the radiologist being more and more involved in guiding patients as to the optimal way of successfully navigating what is all too often a "diagnostic jungle", with the ultimate benefit of an optimal patient outcome.

Q *To finish off, let's go back to breast screening. What do you think about the future potential of personalized risk-based screening programs versus population-based programs, especially in light of the controversy on over-diagnosis?*

In my opinion personalized risk-based screening will be the future of screening. Unfortunately, generally accepted, reliable and reproducible methods for the assessment of individual risk haven't been widely available so far. However, new AI products that can assess the risk of a woman being diagnosed with breast cancer within the next 2 years show promising results. Of course, this promise will have to be confirmed and validated. Randomized control trials will have to be carried out to ensure that risk-based screening really does help to detect more cancers — and especially more aggressive cancers — and to verify that recall rates can be maintained at a sustainable and affordable level. It shouldn't be forgotten that the speed of development in the AI industry is so rapid that the resulting algorithms will be ever-more powerful.

For me the beauty of these products is that their results are totally independent and unaffected by any personal favoritism the radiologist may have, for example his personal views on the merits of supplemental imaging modalities. I see the potential of risk-based screening approaches as an option for harmonizing breast cancer care inside individual countries and throughout Europe. If AI systems can guide us on the optimal way to screen individual women there could be a dramatic benefit not just for each patient but in the cost-efficiency of healthcare systems as a whole.

REFERENCE

1. Graewingholt A & Rossi PG. Retrospective analysis of the effect on interval cancer rate of adding an artificial intelligence algorithm to the reading process for two-dimensional full-field digital mammography. J Med Screen. 2021 Jan 12;969141320988049. *doi: 10.1177/0969141320988049.*