AI-derived software in screening for breast cancer

The Breast Imaging Group in the Department of Medical Imaging at the Radboud University Medical Center in the Netherlands is widely renowned for its work in the improvement and evaluation of radiological techniques for the detection and monitoring of breast cancer. Recently the group has been evaluating an AI-generated software package from the Dutch company Screenpoint for the detection of cancerous lesions in the breast. We wanted to find out more about the department in general and their experience of Screenpoint's Transpara algorithm in particular, so we spoke to Dr. Ritse Mann, breast radiologist and head of the Breast Imaging Group.

Dr. Ritse Mann is head of the Breast Imaging Group in the Radboud University Medical center in Nijmegen, The Netherlands.
Email:
Ritse.Mann@radboudumc.nl

Q Before we get onto discussing your experience with the AI-derived software please tell us about your center in general.

In fact our breast imaging group is spread between two quite distinct hospitals: the Radboud University Medical Center (Radboudumc) and the Netherlands Cancer Institute.

Radboudumc has a local/regional role in breast cancer diagnostics and treatment similar to what's provided in many other community hospitals, with perhaps the difference that we have a very large screening program for women at increased risk. Thus, in Radboudumc, each year we treat about 200 women with breast cancer.

On the other hand, the Netherlands Cancer Institute has a nation-wide tertiary referral function and therefore sees a very large number — about 700 annually — of patients with (often large, developed) breast cancers.

From a scientific point of view, within the Radboudumc we carry out extensive pre-clinical research and development in the field of imaging and in the evaluation of AI tools for screening. In the NCI, research is predominantly focussed on the clinical assessment of patients with lesions and on image-guided de-escalation of therapy. As for equipment, both centers have Digital Breast Tomosynthesis (DBT) systems. In Radboudumc we have Siemens whereas in NCI it's Hologic. For ultrasound we have Siemens and Philips systems respectively and for MRI Siemens and Philips. Automated Breast Ultrasound (ABUS) is only available in Radboudumc. We don't use CESM; for screening purposes we routinely run abbreviated breast MRI protocols with ultrafast acquisitions.

So, all-in-all, between the two institutions, we have a broad experience with many of the currently commercially available breast imaging systems.

Q Regarding screening, what is the usual programme for screening women in The Netherlands?

As in most European countries, women in the Netherlands are invited for screening every two years in the age range of 50 to 75. The take-up rate can vary — there is generally a higher participation in rural areas than in the cities — but overall, about 75% of invited women accept screening, a rate which is higher than in most European countries. Of course there are some women who drop out of the screening programs. There is no precise up-to-date information on the reasons for this but previous studies have shown that drop-outs are partly the result of women having had a negative experience with mammography and/or being relatively more anxious. Other so-called drop-outs are simply the result of women who choose to undergo their mammograms at a different centre.

Currently women at high risk of breast cancer, e.g. those with BRCA mutations, undergo annual screening from age 25 with abbreviated protocol MRI. Women at familial risk undergo breast MRI at a lower frequency.

Regarding another risk factor, namely breast density, we determine this by automated analysis of the images, although our radiologists always have the possibility to overrule the software-generated values. In the clinic, women with dense breasts currently undergo supplemental automated ultrasound. On the basis of the results of the recent DENSE trial, the Dutch parliamentary authorities have recommended that, in screening, women with extremely dense breasts be offered MRI once every 4 years, but we are awaiting practical implementation of this policy decision.

Q What about the performance statistics regarding screening?

It's important to be able to monitor the overall performance of screening services, but in practice this is quite difficult for us since the Dutch national breast screening program is fully extramural, i.e. it's organized totally outside of the hospitals, with hospital radiologists being hired by the screening organizations to carry out the reading of the screening images. This means that breast screening actually carried out within hospitals is restricted to women at increased risk. All this makes it difficult to rapidly generate statistics regarding the performance of the screening programs as a whole or to assess personal performance.

Having said that, our ***cancer detection rates*** are around 6/1000 with screening mammography; these detection rates vary by breast density category (around 3/1000 in category a; 6.8/1000 in category d) In a screening context, it's difficult to say if the imaging modality, e.g. DBT or mammography, has an effect on the detection rates. However if we consider women presenting with symptoms, the yield of tomosynthesis for cancers elsewhere in the breast is about 8/1000, so approximately 25% higher than for mammography.

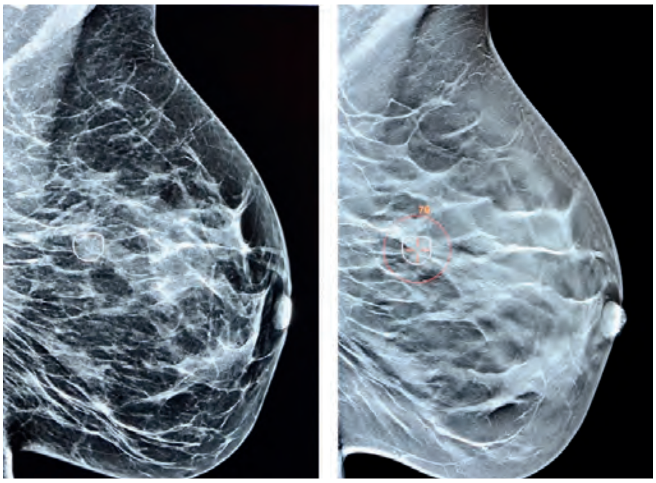
In the Dutch screening programs it is recommended that the ***recall rates*** should be below 2.3%. In practice we are slightly above this target but obviously this depends on the risk the women has of developing breast cancer. On average, about 1 in 4 recalled women actually has breast cancer, regardless of the imaging modality used for screening. These rates are low and are similar to those observed in the Nordic countries.

As for the rates of ***interval cancers***, in population screening these are about 2-3/1000 and are higher in women with dense breasts (up to 5/1000 in category d). The DENSE trial that I mentioned earlier clearly showed that we could effectively overcome the issue of interval cancers almost entirely by shifting the screening of women with dense breasts to MRI.

Regarding ***biopsies***, Positive Predictive Values (PPV) of over 40% have been reported for screen-recalled masses visible under ultrasound. Stereotactic biopsy tends to have a much lower PPV of around 20%. MRI guided biopsy is similarly positive in one in four to one in five patients.

Q *And now let's turn to the Transpara software from Screenpoint*

Our experience with the software actually goes back a long time, since for as long as I have been working at Radboudumc we have carried out research projects on the software and have been beta testing it. The software is now fully up and running and is used in routine, in tomosynthesis mode (we don't carry out normal mammograms within the Radboudumc any more, except for a few very specific situations such as women who have also been screened with MRI).



A spiculated mass identified at DBT by the Transpara software (and missed by the radiologist).

Left panel. As a detection aid, the software can act as a second pair of eyes to support radiologists. AI markers immediately highlight suspicious calcifications and soft tissue lesions on the sytnthetic mammogram.

Right Panel. When clicking on the mark, the software takes you directly to the most suspicious slice in the tomostack, also providing a lesion score. This helps by providing objective information equivalent to that of a radiologist on the areas of suspicion. Transpara findings are graded between 1-100

Personally I always use the software for any images in the so-called "grey area", i.e. images where I may have some doubt and would appreciate a second look. What is really interesting is that often, even before I get to see the images, our technicians will spontaneously check the mammograms with Transpara.

The software itself is fairly straightforward, so any issue is not about actually using the system but getting to trust it. The importance of trust in the software is shown by the fact that, any time there has been a stupid mistake made by the software — and this can happen, although rarely — there is a subsequent, small dip in its use until trust is regained.

In practice, the system works both as a standard CAD tool pointing at specific abnormalities, and also as a decision aid. In other words the radiologist can ask the system for an opinion on a particular area that might be suspected of being abnormal. I really like this aspect, since it basically provides me with an extra pair of eyes in a setting where, most of the time I tend to work alone.

Theoretically there is a danger that the radiologist could be distracted by software-generated marks and not give full attention to other parts of the images. (However, even if you only read what the system points out, you still wouldn't be doing too badly at all). I have to say that overall, the system is as good as I am — however, when we're together, we're better, especially in a clinical context. Thus the radiologists should always continue to keep their eyes open.

As I mentioned earlier we carried out clinical studies in-house, so right from the start I was quite confident that it would work well in routine. The challenge was more in training my colleagues than in anything else. For example, the scores given by Transpara do not directly correspond to a specific likelihood of cancer, so this can be a bit confusing if you are not used to it. Basically what you need to learn is simply that a score of 40 means that the lesion is probably benign and can be safely ignored. Otherwise specificity is ruined, together with trust in the system.

Q *Published data from the clinical trials have shown that the performance of the software is non-inferior to that of radiologists. How do your radiologists react to this — do they perceive the software as a threat or are they grateful for help?*

In fact it's a bit of both. Screening *per se* can be tedious, and it's easy to make preventable errors. This is where eventually I'd guess we will leave it to the computers. But, if this is where the radiologist earns most, it definitely is a sort of a threat. However, in a clinical setting, where the mammogram is reported directly after it is obtained and the results also communicated right away to the patient, the situation is different — it's much easier to embrace the help of a system.

In practice, the case-based scores generated by the system help the radiologist to assess whether or not an extra few seconds should be spent on a mammogram, and in this context I think the scores are really valuable. Thus, if you don't see anything at first glance and the Level of Suspicion (LOS) score is very low, it is easy, safe and reassuring to just move on.

The performance of the software can vary slightly depending on which vendor's mammography system is being used. Since currently the largest datasets used for training the software are from Hologic, the best results are with Hologic mammograms. However I use it mainly in the Radboudumc where we have Siemens mammography machines, i.e. I trust it for the other vendors too.

Similarly the software's efficiency in detecting lesions varies a little depending on the type of lesion. For example there may be slightly more false-positives with calcifications, but that is easy to deal with. In fact the software is deliberately set up to do just this — radiologists will not easily miss a mass, but they might just overlook a small group of microcalcifications.

Theoretically neural-network-based software could "learn by itself" as it handles more and more images, but in practice the Transpara software doesn't work like this. Currently the algorithms are fixed when they are brought to market, so the performance won't change until a new upgrade is installed. It would be really nice in the future to have a local self-learning system that could continuously try to improve itself. This would mean that on a daily basis a benchmark would have to be carried out on a very large independent and validated dataset to check that the performance doesn't actually become poorer. Of course, it should be realized that such a system would also need access to the ground truths, i.e. pathology and follow-up data. In addition, we would need clear and formal regulation to make sure that any such self-learning system works — for example that the validation set is truly independent.

For the moment the Transpara software is thus not optimized specifically for a local situation, but rather is the same everywhere. This has obvious drawbacks, but does have the advantage that if we see an individual patient who happens to come from a population that is not usually seen in our hospital, the software will still work satisfactorily. In that sense it guarantees a sort of equality that I, as radiologist, cannot offer.

Q *So overall, what is your impression of the software? What are the most significant pros and cons?*

The software is very easy to use. Currently, we use a form of the software that is integrated in our workstations. This is easier than as a stand-aside tool on a tablet computer, which is how we started out. We mainly use its decision support feature, i.e. I hardly look at it in clearly normal or abnormal cases, whereas in more difficult cases it simply provides me with some additional confidence.

There have been several instances where I, or one of my colleagues have detected a cancer only because the system pointed it out so it's possible that the software increases the sensitivity but by exactly how much is difficult to say. I would guess by a few percent at best.

What we appreciate is that in practice when we're dealing with somewhat more irregular mammograms and DBT examinations, with the software it is much easier to move on to the next case without any lingering doubts. This certainly speeds up the overall evaluation rate, because it is these cases that typically hold the radiologist up.

> *"... The system can be configured so that performance parameters such as sensitivity and/or recall rates can be set. These thus become medico-political choices...."*

However to be honest, I think that the impact of the system will remain only minor so long as the system is used concurrently. We might become a little better and a little faster in our everyday practice, but it won't be a game-changer.

Only when we are willing to step aside and use the system for a form of independent reading (either as first, second or third reader), will it make a huge difference in workload and costs of screening. Even then, it won't necessarily make screening a lot better, since we are bound by the technology limits of mammography and/or DBT to show early cancers. However the software could bring about a homogeneity in the performance of screening across the world. This in turn could enable an adequate selection of women at higher risk for supplemental screening tests.

The system can be configured so that performance parameters such as sensitivity and/or recall rates can be set. These thus become medico-political choices. It shouldn't be forgotten that even with human reading we don't aim for maximum sensitivity, but rather for an acceptable balance between sensitivity and recall.

There is a different version of the software for mammography and DBT, but in essence it works similarly in each case. For DBT the software first analyses the individual tomoslices, and when any significant findings are identified, these are projected on to the synthetic 2D mammogram. When these are clicked the software goes to the relevant slice in the tomostack.

Currently, I still always scroll once through each of the tomostacks, but I must admit that if the software hasn't signalled anything I don't scrutinize each slice in as much detail. Looking at each slice has the advantage of preventing reportable findings from being overlooked (e.g. large cysts are often automatically ignored by the system, but sometimes it is good to devote a line or two to cysts in the report).

Although the system can be set up in various ways, in our experience it is easiest to just choose one way of working and get used to that.

DBT systems from different vendors have different characteristics, e.g. angle of sweep. Personally, I have only experience with the Transpara software in Siemens DBT, which has a relatively wide sweep angle. However, recently results were published on its use on Hologic DBT (with a small angle) with equally good results. We have contacts with many other centers, who use different DBT machines, as well as different thresholds for recall. Overall, the results have been consistent, which implies that the software is robust and trustworthy, — and, unlike many other AI applications, is satisfactory no matter the particular setting in which it operates.

Q *Do you think that sooner or later, the current European practice of double reading will evolve to a system of a single human reader plus AI software as a second reader?*

Personally, I don't think we can ignore for much longer the use of computers in this field. As humans we simply make too many silly errors, typically because we are distracted for a moment, or because we accidently hit the next button. Such errors would easily be solved by implementing an AI system as a third reader.

However if you ask me, the second reader approach is more difficult. Studies we have conducted so far in consecutive screening series show that humans and AI have similar detection performances but there is a huge difference in the cases actually recalled. Hence, the findings of the AI and humans need to be integrated and arbitrated to keep the recall rate under control. Theoretically it is possible to do this by group arbitration, but it remains to be seen what the relevant contribution would be of the AI and human detected cancers in such a situation. For example it could be possible that the humans overrule all additional cancers detected by AI, which would render its value zero.

If the legal requirements can be satisfied and appropriate QC programs set up, I would be more in favor of using AI as a first reader. This would involve simply selecting the subset of cases for which human reading might be useful (i.e. preselection) — all other cases would then be excluded from human intervention. This would dramatically reduce the workload and might actually boost both sensitivity and specificity of the screening programs.

> *"... If the legal requirements can be satisfied and appropriate QC programs set up, I would be more in favor of using AI as a first reader ..."*

Q *What about DBT being eventually favoured over mammography in Europe as the preferred screening modality?*

I think that this will happen. Whether we will actually read all the DBT images is another matter. Good AI software may be embedded in the image reconstruction and simply highlight all potential findings in the synthetic mammogram. There is, after all, little reason why a synthetic mammogram should look exactly like a normal mammogram (if so we would actually be deliberately masking tumors again, which is ridiculous when put that way).

Such approaches would inevitably mean changes in the role of the radiologist. I like to compare such a future role with that of a hematologist; no-one expects the hematologist to do a manual cell-count on every blood sample. Instead hematologists should be capable of explaining what it means when the computer reports something abnormal, and be able to implement a logical follow-up. That doesn't free the hematologists from the responsibility for the diagnosis, though. To get back to radiology, it is very likely that in the near future a mammography machine will no longer just yield only an image, but will accompany the image with a full report on the characteristics of the breast including density and any abnormalities observed.

In such a scenario, it is clear that rules will need to be re-defined. Currently radiologists are officially obligated to look at every image that is stored in the PACS system. It isn't generally realized that *de facto* such a rule has already been obsolete for some time — in CT and MRI this usually doesn't happen. Radiologists shouldn't be freed of responsibility, but rather they need to ensure that they check that what the computer produces is logical, and that the results are incorporated in the patient-care pathway in such a way that is beneficial to the patient.

Coming back to legal responsibility, in essence AI mammography is a piece of machinery, like a CT or MRI scanner. If the radiologist uses it on a patient, he/she is responsible, not the manufacturer. This means that there will have to be certain standards and benchmarks against which the program can be checked. It would be for the radiological community as a whole to organize this.

Q *How would radiologists react to the prospect of only dealing with positive or difficult cases and never seeing "normal cases"?*

Some re-adjustment will obviously be needed. Also radiologists would probably become less confident in reporting normal cases as normal. However, this will mainly just mean another change in our profession — there are very few radiologists I know that started to read mammograms because most of them are normal.

We shouldn't forget that any new ways of working will have to be explained to the women involved. This should be done simply and clearly, always stressing the safety brought about by the changes. In general people adapt really quickly to technological advancements, and I see no reason why AI-assisted radiology couldn't be acceptable (patients don't complain about tests from the hematologists and clinical chemists either).

Q *What about the replacement of the current general population screening strategy by one based on personalized risk assessment to reduce radiologists' workload ?*

I doubt that personalized screening will lead to a reduction in workload. Screening is the one thing that reduces mortality and enables de-escalation of therapy. Personalized screening mainly tackles the underdiagnosis that is currently abundant in screening, by offering more or better screening to women for whom standard mammography is insufficient. In these women, we might only observe a shift from one screening modality to another, but this will not reduce the overall amount of work. Only the subgroup for which we could safely do less screening would enable us to truly lower the workload for radiologists. Instead I see the implementation of computer-based image interpretation as more useful in this respect.

However this doesn't mean that screening strategies should be frozen in their current form — indeed, some adaptation of screening is imminent. For example, we should soon get supplemental or replacement screening in women with very dense breasts, because mammography in this population really underperforms.

Q *Is the debate on the over-diagnosis/ over-treatment aspects of screening finally becoming less heated? If so why?*

A little. Overdiagnosis is still a major issue, prticularly among those physicians who actually have to carry out the treatment and follow-up. However, virtually no one doubts any more that early detection does decrease mortality. For some lesions that were formerly treated as cancer, watchful waiting has now become acceptable, so this makes the effect of overdiagnosis less dramatic. Likewise, image-guided de-escalation of therapy in small breast cancers also reduces the negative impact of overdiagnosis. Despite this, overdiagnosis is still a problem, and we should continue to pay attention to it. But if we use imaging not just to find the cancers early, but also to minimize the impact of treatment for any cancers detected, we could largely overcome the problems underlying this debate.

Q *what about likely future trends in the incidence of breast cancer?*

There seems to be a real increase in the number of women that will get breast cancer at some point in their life. The statistics in the Netherlands have been increasing gradually over time. Our current estimates are that 1 in 6.6 women will have breast cancer, which is about the highest incidence in the world. Probably there are dietary, lifestyle and hormonal factors that play a major role here, along with the fact that life expectancy is still increasing. However there is no single behavioral practice which could be implemented to reduce the increase in the incidence of breast cancer, like stopping smoking decreases the risk of lung cancer.

So, all we can do right now is alleviate its harms. In that sense, the importance of screening can only increase.