NICE National Institute for Health and Care Excellence

NICE advice

Artificial intelligence in mammography

Medtech innovation briefing
Published: 5 January 2021

www.nice.org.uk/guidance/mib242

Summary

- The **technologies** described in this briefing are 5 artificial intelligence (AI) technologies for mammography. They can help to select mammography images that need further diagnostic tests to see whether detected features in the image are malignant.
- The **innovative aspects** are that the technologies can provide information to support radiologists, or other qualified people, to interpret mammograms.
- The **intended use** would be to support qualified people to interpret mammograms.
- The **main points from the evidence** summarised in this briefing are from 6 retrospective clinical validity studies, 3 conference proceedings, 3 conference abstracts, and 1 diagnostic accuracy study. They included a total of 71,470 mammography exams. They show that AI technologies may improve performance and save time in interpreting mammograms. There are a number of ongoing trials for these technologies.

- **Key uncertainties** around the evidence or technology are that few clinical validity studies have used UK datasets representative of the target population for screening. Also, there are no published prospective clinical studies.
- The **cost** models are different for each of the technologies. One technology costs £25,000 to £45,000 for a one-off purchase plus additional licences and ongoing updates and support. Another technology has a yearly subscription costing £13,400 in year 1, and £11,370 thereafter. The last technology has a cost per exam subscription between £0.60 and £3.00 per exam (excluding VAT). Using AI technologies could reduce resource use by helping reduce the workload of staff reading mammograms.

The technologies

Artificial intelligence (AI) technologies are being used increasingly in tasks usually done by humans. These systems are trained using large datasets, with machine learning approaches. A system is provided with training images, each with the desired output or classification (termed 'ground truth' in AI). Instead of learning how to classify new cases based on predefined rules, the system learns from the examples provided and detects patterns and features that predict the output. Because the AI technology may identify features that humans do not, algorithms trained in this way can outperform humans in some classification tasks (The Royal Society, 2017).

AI technologies exist for both full field digital mammography (FFDM, 2D imaging) and digital breast tomosynthesis (DBT, 3D imaging). The AI software detects and displays suspicious features in the image, and predicts the likelihood of malignancy, to help clinical diagnosis.

This briefing focuses on 5 AI technologies for mammography: Transpara Mammography and Transpara DBT (ScreenPoint Medical), HealthMammo Software (Zebra Medical Vision), and ProFound AI for 2D Mammography and ProFound AI for DBT (iCAD). Other relevant technologies may be available but are not included in this briefing. Reasons for this include not being identified in horizon scanning because at the time, they were not commercially available to the NHS, or the company choosing not to take part.

Transpara (ScreenPoint Medical)

- Exam type: FFDM (Transpara Mammography) or DBT (Transpara DBT).

- Setting: screening and diagnostic.
- Function: detecting and characterising suspicious features.
- Algorithm: deep learning convolutional neural networks, feature classifiers and image analysis algorithms.
- Training and validation set: over 1 million images from US and EU sites, including the OPTIMAM (NHS Breast Screening Programme) database.
- Standalone test sets: 5,327 cases (FFDM) and 2,319 (DBT) cases representing a screening patient population.
- Current software version: 1.6.0 (November 2019).
- Compatible FFDM modality vendors: Fujifilm; General Electric (GE); Hologic; Philips; Siemens.
- Compatible DBT modality vendors: Hologic; Siemens.
- Description: Transpara is a 2-step approach to the assessment of mammograms. An overall Exam Score gives the likelihood of cancer being present in an exam. This can be used under the supervision of a radiologist or other qualified person interpreting the mammogram, to triage screening exams, or as an independent reader. Detection Aid and Region Analysis give region-based information about abnormalities found in the case, including the location, likelihood of malignancy, and type of abnormality, which can be used to support the decision of the reader.

HealthMammo Software (Zebra Medical Vision)

- Exam type: FFDM.
- Setting: screening.
- Function: detecting and characterising suspicious features.
- Algorithm: deep learning convolutional neural networks.
- Training and validation set: more than 500,000 cases from 150 facilities across 3 continents.

- Standalone test sets: 835 cases from the US, UK and Israel, representing a screening population.
- Current software version: 2.2 (April 2020).
- Compatible FFDM modality vendor: Hologic.
- Description: HealthMammo Software automatically analyses images for suspicious findings and notifies the workstation or Picture Archive and Communication System (PACS). These findings can be used for decision support by the radiologist or other qualified person interpreting the mammogram, for worklist prioritisation and changing workflow, and as an independent second reader.

ProFound AI (iCAD)

- Exam type: FFDM (ProFound AI for 2D Mammography) or DBT (ProFound AI for DBT).
- Setting: screening and diagnostic.
- Function: detecting and characterising suspicious features.
- Algorithm: deep learning convolutional neural networks, feature classifiers and image analysis algorithms.
- Training and validation set: approximately 2 million images.
- Standalone test sets: 2,449 FFDM cases and 6,890 DBT cases representing a screening patient population.
- Current software version: 2.1.
- Compatible FFDM modality vendors: FujiFilm; GE; Hologic; Siemens, Philips.
- Compatible DBT modality vendors: GE, Hologic, Siemens.
- Description: ProFound AI automatically detects malignant soft tissue densities and calcifications and can be used for clinical decision support by the radiologist or other qualified person interpreting the mammogram. It records a Lesion Score for each suspicious feature detected, which represents the likelihood that the detection is malignant. A Case Score is also recorded, which indicates the likelihood that a malignancy is present in the case. This can be used to triage screening exams.

Innovations

The NHS Breast Screening Programme (BSP) uses a system of 2 readers, and arbitration to interpret mammograms. However, it is currently facing a shortage of qualified people, especially radiologists. AI technologies in this setting could reduce workloads by replacing 1 of the 2 readers, or by performing triage according to likelihood of an image being malignant. This triage could be used to prioritise images according to the likelihood of malignancy, so that images with a higher chance of malignancy are reviewed sooner. It could also be used to automatically classify images showing a low likelihood of malignancy as normal, and remove these from the images to be reviewed.

Abnormalities detected within mammograms can include masses, microcalcifications, architectural distortions, and asymmetric density. Some of these changes may be small and difficult to interpret by eye, even for an experienced reader. Therefore, there is a risk of missing images that should be recalled for assessment (false negatives), and of recalling images for assessment that are normal (false positives). AI technologies could support decision making by those interpreting mammograms and reduce unnecessary or missing recalls.

Current care pathway

Women in England are invited for breast screening every 3 years from age 50 to 70, and after this they can self-refer. The NHS BSP is under the remit of the UK National Screening Committee, Public Health England. Mammograms taken in the NHS BSP are interpreted by 2 qualified independent readers. If these 2 outcomes are different, a third reader or group of readers will arbitrate. Mammography results are shared with the woman having screening and the GP within 2 weeks.

NICE's guideline on familial breast cancer includes using mammography for surveillance in those at greater risk of breast cancer, because of family history. Surveillance with mammography is also included in NICE's guideline on early and locally advanced breast cancer. In people who have had breast cancer, surveillance should be offered every year until they are eligible to be screened by the NHS BSP, or every year for 5 years if they are already of screening age.

The following publications have also been identified as relevant to this care pathway:

- Clinical guidance for breast cancer screening assessment by the NHS BSP gives guidance for the assessment of suspicious regions detected on screening mammography. This may be done in a clinic where people with symptoms are also referred, and the assessment pathway could therefore apply to both screening and symptomatic populations. Assessment options include follow up with ultrasound, further mammography or tomosynthesis, biopsy of lesions, and less frequently, MRI. AI technologies could be used in this setting to detect and diagnose breast cancer from the mammography or tomosynthesis exams.
- Quality assurance guidelines for breast cancer screening radiology, by the NHS BSP states that double reading of mammograms should be considered mandatory in centres offering digital mammography. Staff time could be saved if an AI technology could replace 1 of these human readers, or if it could help to target the focus of the human readers to suspicious regions.
- The UK National Screening Committee's interim guidance for those wishing to incorporate AI into the National BSP details the evidence that would be needed to support a modification of the NHS BSP to incorporate AI. Evaluation focus will be on the impact of changes to the benefits and harms of those screened, which are linked to the spectrum of disease detected. This is therefore important to consider alongside test sensitivity. High specificity is also important to reduce false positive recalls for assessment, and their impact on downstream testing and treatment pathways. How the technology interacts with the radiologist should also be considered, and the test accuracy analysis of AI technologies, split by population subgroups such as age and ethnicity.

Population, setting and intended user

These technologies would be used in breast screening and assessment clinics. They may also have a role in clinical symptomatic clinics where mammograms may only be read once. AI software would be used by those qualified to interpret mammograms, including radiologists, and radiographers who have had appropriate training (radiography advanced practitioners).

Additional training in image interpretation would be needed, specific for the AI system used.

Costs

Technology costs

Transpara is available as an on-premises subscription model, or with a multi-year licence. Pricing is based on the volume and type of study carried out (2D FFDM or 3D DBT). This also includes installation, training, ongoing support and future upgrades. Typical prices range from £0.60 to £3.00 per exam. The local hardware that the virtual server is installed on must be provided, secured, and maintained by the customer.

HealthMammo Software runs on a server or virtual machine purchased and set up by the customer, either on-premises or in the cloud. Pricing includes installation, setup and configuration, training, a software licence and ongoing support and maintenance. Pricing is for unlimited scan volume, including cloud costs, and for each installation and PACS connection it costs £13,400 for the first year and £11,370 in subsequent years.

ProFound AI is available as a one-off purchase, with a multi-year licence, or as an on-premises subscription model. The one-off purchase includes a perpetual licence and hardware server, installation and user training, and costs in the range of £25,000 to £45,000. Additional licences can be purchased separately and cost in the range of £15,000 to £25,000, depending on the licence type. Licence types are FFDM only, DBT only, or a combination of FFDM and DBT. Post-warranty support and software updates typically cost 12% to 15% of the purchase price. Subscription pricing is based on the volume and type of study done (2D FFDM or 3D DBT). This includes installation, training, ongoing support and future upgrades. The hardware server is sold separately. Typical prices range from £1 to £3 per exam.

Costs of standard care

Investment in AI technologies would be alongside standard care. The cost of taking a mammogram is about £25. The total cost, including interpretation of results and an outpatient appointment, is about £170 (weighted average cost of consultant and non-consultant led appointments [WF01B to D and WF02B to D], NHS England National Cost Collection, 2018/19).

Resource consequences

Transpara is currently used in 4 NHS trusts, and is expected to be installed in further trusts in the next 3 to 6 months.

HealthMammo Software is not currently used in any NHS trusts.

ProFound AI is not currently used in any NHS trusts but is being used in a private sector hospital in the UK.

One expert commentator knew of about 3 symptomatic breast clinics using AI decision support systems in the UK.

Trusts adopting these technologies will need additional computing resources, including a dedicated server or virtual machine. Additional staff time and computational expertise will be needed for installation, and may also be needed for maintenance and ongoing support. Evidence suggests that adopting these technologies could reduce the workload of staff reading mammograms. It could also further reduce the time taken to read individual exams.

Training to use Transpara is given by the company and includes an introductory presentation, demonstration, hands-on training with a sample set of training exams, and training in the local clinical environment. This is estimated to take less than 2 hours in total, and takes place with the installation.

Training to use HealthMammo is provided by the company after installation, for radiologists and IT staff, and training materials are supplied for new users. Training usually takes about 1 hour.

Training to use ProFound AI is provided by the company at installation, and includes a presentation of the technology, and hands-on training with real clinical cases. On average, this takes 1 hour.

NHS England's report of the independent review of adult screening programmes in England (October 2019) states that it is widely agreed that IT systems for breast screening urgently need renewing. Multiple inefficiencies, opportunities for error and corresponding benefits that will accrue from a new system have been identified, including the prompt and economical introduction of AI software in the future.

Regulatory information

Transpara 1.6.0 is a CE-marked class IIa medical device under the EU Medical Devices Directive (MDD) for both full field digital mammography and digital breast tomosynthesis.

HealthMammo Software 2.2 is a CE-marked class IIa medical device under the EU MDD.

ProFound AI for Digital Breast Tomosynthesis and ProFound AI for 2D Mammography are CE-marked class IIa medical devices under the EU MDD. The PowerLook Server is CE marked as class I under the EU MDD.

Equality considerations

NICE is committed to promoting equality of opportunity, eliminating unlawful discrimination and fostering good relations between people with particular protected characteristics and others.

Mammography is offered for routine breast cancer screening in women over 50. It may also be offered to men and women of any age, who are either suspected of having breast cancer, or who are at increased risk of breast cancer. It is therefore associated with the protected characteristics of sex and age.

Clinical and technical evidence

A literature search was carried out for this briefing in accordance with NICE's interim process and methods statement. This briefing includes the most relevant or best available published evidence relating to the clinical effectiveness of the technology. Further information about how the evidence for this briefing was selected is available on request by contacting mibs@nice.org.uk.

Published evidence

Six retrospective clinical validity studies, 3 conference proceedings, 3 conference abstracts, and 1 diagnostic accuracy study are summarised in this briefing. These include 71,470 mammography exams. Of these, 2,654 are duplicated across 2 studies, and 260 are duplicated across a study and a conference abstract. One clinical validity study, 2 conference proceedings and 1 conference abstract considered a screening population. The other study datasets were enriched with exams of people with cancer to increase the prevalence above that expected in a screening population.

One further conference abstract was provided by ScreenPoint Medical for Transpara, and a further 3 conference abstracts were provided by iCAD for ProFound AI. These abstracts did not name their respective technologies and so have not been included in the evidence summary.

Key findings include comparable or improved accuracy when artificial intelligence (AI) support is used, compared with an independent human reader, and faster reading times.

Overall assessment of the evidence

The overall evidence base is at an expected level for AI technologies emerging in the NHS. However, some of the AI technologies included in the briefing have a greater evidence base than others, in terms of both number and size of studies. The main limitation across the evidence base is that the studied datasets are mostly enriched with exams of people with cancer. However, evidence is beginning to emerge from datasets more representative of a screening population. Because the primary or intended use of these technologies is in a screening setting with lower prevalence of cancer, using enriched datasets is a significant limitation. This is recognised by the UK National Screening Committee (NSC) in their interim guidance, which requires clinical validity studies of AI technologies to be done using unenriched datasets, representative of the target population for screening. The guidance also recommends using paired, or within-person comparisons of accuracy between different AI systems, against double reading plus arbitration as the comparator. This would allow assessment of the concordance between systems, as well as their sensitivity and specificity with respect to the reference standard. Although the studies included in this briefing mostly consist of paired readings, none have provided a head-to-head comparison of AI systems. Also, most comparators were ground truth, determined by biopsy or conclusive follow up. The UK NSC interim guidance itself recommends prospective randomised test accuracy studies, preferably with multiple intervention arms to incorporate different AI technologies, and provide real world evidence. Evidence in this area is also emerging, and a study comparing 3 unnamed technologies (Salim et al. 2020) has been published. However, a further challenge in this area is that AI technologies are continuously evolving, and the evidence available is usually outdated by the time it is published. Also, any head-to-head comparisons would need to be done for comparable technologies, at the same stage in their development. A key priority of the UK NSC is to also assess interval cancers happening between routine screening visits, improved specificity, and detection of different cancer subtypes. This would be considered sufficient evidence for adoption of an AI technology that detects the same disease spectrum as standard care. For AI technologies detecting a significantly different disease spectrum, evidence of longer-term harms and benefits must be provided. Also, the optimal operating point should be established, balancing sensitivity and specificity. None of the included technologies appear to currently meet these criteria, and so the trials needed to improve the evidence base would be independent evaluation with retrospective datasets, followed by prospective randomised accuracy studies. Research is also needed to assess replacing a human reader with AI in a double reading scenario, when arbitration is needed, and the effect this has on accuracy and recall rate. The outcomes reported from the included studies are also limited. The area under the receiver operating characteristic curve (AUROC) is not clinically meaningful, and depends upon the thresholds used in practice, which may differ geographically. A further limitation of the evidence base and the technologies included in this briefing, is their lack of generalisability to all full field digital mammography (FFDM) and digital breast tomosynthesis (DBT) systems. Additionally, there is a lack of published evidence evaluating the generalisability of performance of AI technologies between different FFDM and DBT systems from the same vendor. Comparisons have not been drawn between technology performance across different screening centres. The UK NSC interim guidance states a clear preference for AI technologies that can show cross-vendor compatibility. It also indicates that those that cannot show this may only be recommended for use with FFDM and DBT systems for which evidence is available.

Mia (Kheiron Medical Technologies) is an AI technology not included in this briefing. It is currently undergoing a large-scale retrospective study of algorithm generalisability with the East Midlands Radiology Consortium (EMRAD), to assess the feasibility of its adoption as an independent reader of breast cancer screening mammograms. This is part of wave 2 of the NHS test bed programme, funded by NHS England. This briefing did not identify any published evidence for the Mia system.

Lång et al. (2020)

Study size, design and location

Retrospective clinical validity study of AI to identify normal FFDM exams in a screening population. A cohort from the Swedish Malmö Breast Tomosynthesis Screening Trial of 9,581 double-read mammography screening exams, including 68 screen-detected cancers, and 187 false positives. The AI system assigned a cancer risk score to each exam. The effect of excluding reading by radiologists, at different thresholds, was investigated.

Intervention and comparator

Intervention: Transpara (version 1.4.0, ScreenPoint Medical) to identify normal (low-risk score) mammograms in a screening population.

Comparator: ground truth, defined by histology of surgical specimen or biopsy, cross referenced to regional cancer register.

Key outcomes

If mammograms with AI risk scores of 1 or 2 were considered normal, 1,829/9,581 (19.1%; 95% confidence interval [CI] 18.3 to 19.9) of exams could be removed from screen-reading. This included 10/187 (5.3%; 95% CI 2.1 to 8.6) false positives, and no cancers. This could reduce the workload of radiologists, and costs associated with screen-reading. Of the 5,082/9,581 (53%; 95% CI 52.0 to 54.0) with low-risk scores (5 or lower), 52/187 (27.8%; 95% CI 21.4% to 34.2%) were false positives, and 7/68 (10.3%; 95% CI 3.1% to 17.5%) were exams showing cancer. The radiologists considered all but 1 of these cancers to be clearly visible.

Strengths and limitations

The AI technology had been trained and validated using about 180,000 normal, and 9,000 abnormal mammograms from 4 different vendors, independent of those used in this study. The dataset was small when the proportion of screen-detected cancers is considered. Also, the study did not try to explain why the AI system assigned low-risk scores to cancers visible to the radiologists. However, the mammograms were taken from a consecutive cohort of women presenting for population screening at a single centre in Sweden, so selection bias is unlikely. Results may not be generalisable to a UK population, although the recall rate after screening was 2.7%, which is comparable and is more representative of a screening population than studies using enriched datasets. The authors acknowledge limitations around the study population, and using only 1 mammography and AI vendor combination. The company provided technical support for the study.

Lauritzen et al. (2020, conference abstract, presented at ECR 2020)

Study size, design and location

Retrospective study of AI as a rule-out tool for screening mammograms. Included 18,020 double-read exams from the Danish Capital Region breast cancer screening program. There were 143 screen-detected cancers, and 447 non-cancer recalls (false positives). Exams were sorted into categories from 1 to 10 (low to high chance of malignancy) to assess the number of exams, and non-cancer recalls that could be avoided by detecting normal exams before radiologist reading.

Intervention and comparator

Intervention: Transpara (version 1.5, ScreenPoint Medical) as a rule-out tool.

Comparator: outcome of double reading of screening mammograms.

Key outcomes

At a threshold of 5, 58.52% of studies were classified as normal and included 3.5% screen-detected cancers and 23.71% non-cancer recalls. Categories 1 and 2 contained 26.29% of studies, including 5.82% non-cancer recalls and 1.36% screen-detected cancers. Category 1 contained 14.58% of studies, with 2.68% non-cancer recalls, and no screen-detected cancers.

Strengths and limitations

This is a conference abstract that has not had peer review and lacks detail. The authors acknowledge that the AI technology identified some exams of people with cancer as normal exams and that improvements could be made by having radiologists examine these. The total number of exams showing cancer was also limited, and the authors highlight that a larger study should be done.

One of the authors was reported in the abstract as chief executive officer of ScreenPoint Medical, and another as an employee.

Balta et al. (2020, conference proceedings, presented at SPIE 2020)

Study size, design and location

Retrospective clinical validity study of AI as a rule-out tool in FFDM. It included 18,015 consecutive screening exams acquired at a single institution on devices from 2 vendors. There were 77 exams excluded, and the final dataset included 114 biopsy-proven screen-detected cancers. Each exam was independently read by 2 radiologists, and each exam with at least 1 decision to recall for assessment was reviewed at a consensus meeting for a final recall decision. AI was used to assign an AI score to each exam, representing the chance of cancer being present, and this was used to assess the impact of setting different thresholds for exams to be read by 1 radiologist instead of 2.

Intervention and comparator

Intervention: Transpara (version 1.6.0, ScreenPoint Medical) as a rule-out tool to reduce radiologist screening workload.

Comparator: all images read by 2 radiologists, with consensus meeting for all recalled cases.

Key outcomes

All AI score thresholds were evaluated and for scores of 1 to 7, the cancer detection rate was the same as if the exam had been read by 2 readers. In some cases, cancer was missed by the AI system, but the second reader would recall it. The overall recall rate would decrease from 5.35% (n=958) with double reading of all exams, to 4.79% (n=857) if exams with scores of 1 to 7 were read by a single radiologist only, and the workload for radiologists would decrease by 32.6%.

Strengths and limitations

This is from conference proceedings and may not have had peer review. The cases were collected consecutively in a screening population, so selection bias is unlikely, and the sample is representative of the target population. The authors acknowledge that no follow up was available for normal exams, and so incidence of interval cancers is unknown.

The affiliation of 1 author was ScreenPoint Medical.

Dahlblom et al. (2020, conference proceedings, presented at SPIE 2020)

Study size, design and location

Retrospective clinical validity study of AI as a tool to rule-in using DBT after FFDM. An AI score of the chance of cancer being present was given by the system to 14,768 FFDM exams, which also had DBT exams available. The effect of different AI score thresholds for adding a DBT exam to FFDM, on number of cancers detected, additional DBT exams needed, detection rate, and false positives, was assessed.

Intervention and comparator

Intervention: Transpara (version 1.4.0, ScreenPoint Medical) as a rule-in for DBT exam after FFDM.

Comparator: FFDM images read by Transpara (version 1.4.0, ScreenPoint Medical) only.

Key outcomes

At an AI score threshold of 9.0, 26% (n=25) more cancers would be detected than on FFDM alone. This is 61% of the 41 cancers originally detected only on DBT, and 16 (12%) would still be missed. At this threshold, 1,797 (12%) of exams would have both FFDM and DBT.

Strengths and limitations

This is from conference proceedings and may not have had peer review. The cases were collected in a screening population that is likely to be representative of the target population.

Halling-Brown et al. (2019, conference abstract, presented at RSNA 2019)

Study size, design and location

Retrospective clinical validity study of AI as a rule-in tool. There were 2,683 screening mammograms from the OPTIMAM Medical Image Database (OMI-DB), which collects NHS Breast Screening Programme images from centres in the UK. All had biopsy-proven cancers of different grades and 1,212 had a previous mammogram available. The AI system was calibrated to make recall decisions at 50%, 10% and 4% recall rates, using a separate independent screening dataset.

Intervention and comparator

Intervention: Transpara (ScreenPoint Medical) as a rule-in tool.

Comparator: biopsy-proven ground truth.

Key outcomes

When applied to the 2,683 mammograms with biopsy-proven cancer, the calibrated AI had sensitivities of 99.3%, 87.7% and 76.1% for recall rates of 50%, 10% and 4%, respectively. At a recall rate of 4%, 16.8% of the 1,212 previous screening mammograms would have been recalled. Also, at a 4% recall rate, the AI recalled a greater proportion of higher than lower grade cancers (80.7% grade 3 compared with 68.2% grade 1) at $p<0.001$.

Strengths and limitations

This is a conference abstract that has not had peer review and lacks detail. The mammograms used in the study had not been used previously to train, validate or test the AI system. The results may be generalised to the UK because they used UK cases, and a recall rate of 4% to assess previous screening mammograms, which is typical in a UK screening setting. However, these mammograms were selected because of a later positive mammogram, and do not represent a typical screening population.

One of authors was reported in the abstract as an employee of ScreenPoint Medical, and another as a shareholder.

Rodriguez-Ruiz et al. (2019a)

Study size, design and location

Retrospective clinical validity study of AI as a stand-alone reader in FFDM assessing 2,654 digital mammography exams in study dataset. Ground truth was verified by histopathological analysis or follow up as: cancer (n=653), benign (n=768) and normal (n=1,233). About half of exams were from a screening population and half were from a clinical symptomatic population, collected from studies across 7 countries using mammography systems from 4 different vendors.

Intervention and comparator

Intervention: Transpara (version 1.4.0 ScreenPoint Medical).

Comparator: average radiologist performance against ground truth.

Key outcomes

AI performance was statistically non-inferior to that of the average of 101 radiologists. The AUROC for the AI system was 0.840 (95% CI 0.820 to 0.860) compared with 0.814 (95% CI 0.787 to 0.841). AI had AUROC higher than 61.4% of radiologists.

Strengths and limitations

The AI technology had been trained and validated using over 189,000 mammograms (5% with cancer), independent of those used in this study. However, the source of these images was not reported. The mammograms used in this study were from studies across different countries, so the results are likely generalisable to a large population. It is reported that there is overlap between 2 of the studied datasets, and that centres provided a Breast Imaging Reporting and Data System BI-RADS score. This indicates how concerned the interpreting reader is about the findings, or probability of malignancy, which could cause inconsistencies in interpretation. The study datasets were enriched with malignant cases, which is not normal for the intended screening population. Knowing this may have influenced the judgement of the radiologists scoring the images. There is a discrepancy in the number of mammograms stated in the abstract and main body of the text.

Two of the 15 authors are named in the paper as employees of ScreenPoint Medical, and technical support for the study was provided by the company.

Rodriguez-Ruiz et al. (2019b)

Study size, design and location

Retrospective clinical validity study of AI as a rule-out tool in FFDM for 2,654 digital mammography exams in same study dataset as reported in Rodriguez-Ruiz et al. (2019a).

Intervention and comparator

Intervention: Transpara (version 1.4.0, ScreenPoint Medical) as a rule-out tool to reduce radiologist screening workload.

Comparator: all images read by radiologist.

Key outcomes

Setting the likelihood threshold at 5 (high likelihood above 5) resulted in a trade-off between approximately halving (-47%) the radiologist workload, and excluding 7% of true-positive exams. A threshold of 2 resulted in workload reduction of 17%, and excluding only 1% of true positives. The area under the curve was not affected by pre-selection except at the extreme AI-generated likelihood score of 9.

Strengths and limitations

In addition to limitations for related study, Rodriguez-Ruiz et al. (2019a), the authors stated that they were unable to analyse results for detection mode (screening or clinical), for histopathological type of cancers, or for breast density, because this information was not available from the original studies.

Rodriguez-Ruiz et al. (2019c)

Study size, design and location

Retrospective clinical validity study of AI as a decision support aid in FFDM for 240 (100 cancer, 100 normal, 40 false positive cases recalled for assessment) digital mammography exams from 2 different mammography vendors in the US and Europe. Fully crossed study design with 14 radiologists reading half the images with AI support and half without AI support, in 2 sessions at least 4 weeks apart.

Intervention and comparator

Intervention: images read by radiologists with Transpara (version 1.3.0 ScreenPoint Medical) AI decision support.

Comparator: images read by same radiologists without AI support.

Key outcomes

AUROC was higher with AI support (0.89 compared with 0.87, p=0.002), sensitivity increased with AI support (86% compared with 83%, p=0.046), specificity trended toward improvement but not significantly so (79% compared with 77%, p=0.06). Reading times per case were similar (149 seconds with AI compared with 146 seconds without AI). AUROC of the AI system alone was similar to average AUROC of the radiologists (0.89 compared with 0.87).

Strengths and limitations

The sample size was small, and the method for randomly selecting the final population from the available exams of each type was not reported. The study covered the US and Europe, giving some generalisability to populations in those areas. Only exams from 2 device vendors were considered, so the AI performance on exams taken on other devices cannot be assumed. Despite the main inclusion criteria being women attending for screening with no symptoms or concerns, the selection of cases produced a dataset enriched with exams of people with cancer. This is therefore not representative of the intended population. Although the mammograms were collected in 2 centres, all radiologists were based in the US, and interpretation may differ geographically or locally.

ScreenPoint Medical was responsible for data generation, and payment of external study costs. A research agreement exists between the academic institution responsible for the work and the company. However, no financial compensation was provided.

Rodriguez-Ruiz et al. (2020, conference proceedings, presented at

SPIE 2020)

Study size, design and location

Retrospective study of AI to replace 1 radiologist in a double reading setting. In total, 31,650 radiologist assessments of 2,892 mammograms were available in a database. Bootstrapping was used to select different combinations of mammogram and radiologist assessment. This was to simulate scenarios of double human reading and double hybrid reading with the second reader replaced by AI. For this study, an AI score of 10, indicating chance of cancer, represented a recall by the system.

Intervention and comparator

Intervention: images read by radiologists with Transpara (version 1.4.0, ScreenPoint Medical) as second reader. Consensus by an independent radiologist if needed.

Comparator: images read by 2 independent radiologists, with consensus by an independent radiologist if needed.

Key outcomes

When using AI as a second reader, the workload (number of human assessments including arbitration) was reduced by 44%, compared with double human reading. Sensitivity was similar, at 81.4% compared with 81.5% (p=0.88). Specificity was improved by 5.3% with AI from 69.9% to 75.2% (p<0.001).

Strengths and limitations

This is from conference proceedings and may not have had peer review. The authors acknowledge that using a single reader to arbitrate may not be representative of centres who use panels for this. The dataset was enriched with exams of people with cancer. It also used mammograms from a range of previous studies, which have their own limitations. It is not clear if the same images and radiologist readings were used for both scenarios in each case. These mammograms are reported to have been used previously to benchmark the standalone performance of the AI system, compared with that of radiologists.

The affiliation of the lead author was ScreenPoint Medical.

Sasaki et al. (2020)

Study size, design and location

Retrospective clinical validity study of AI as a standalone reader of FFDM exams for 310 people of Japanese family origin in an outpatient setting, including 69 with malignant lesions. The AI system gave each case a cancer likelihood score, and recall thresholds of 4 and 7 were used to determine sensitivity and specificity.

Intervention and comparator

Intervention: Transpara (version 1.3.0, ScreenPoint Medical).

Comparator: performance of 3 radiologists reaching a consensus, against ground truth.

Key outcomes

The AUROC was higher for human readers than for Transpara, at 0.816 compared with 0.706 ($p<0.001$). Sensitivities were 89% for human readers. For AI they were 93% for a cancer likelihood cut-off score of 4, and 85% for a cut-off of 7. Specificities were 86% for the human readers, and 45% and 67% for AI, with cut-offs of 4 and 7 respectively.

Strengths and limitations

The AI technology had been trained and validated using a multi-vendor multicentre mammogram database. This included 9,000 biopsy-proven cancers, independent of those used in this study. The small number of exams studied were selected by a radiologist from 11,891 exams so that 22% had malignant lesions. The cases were therefore prone to selection bias, and not representative of a screening population. The people in the study were of Japanese family origin, so results may not be generalisable to a UK population. The authors acknowledge limitations around training the AI system using mammograms mainly from a Western population (Transpara's training and validation set is based on images from US and EU sites).

Conant et al. (2019a)

Study size, design and location

Retrospective clinical validity study of AI as a decision support tool in DBT using 260 DBT exams from 7 sites in the US. These included 65 exams found to have cancer. This was a fully crossed study design with 24 radiologists (13 of whom were breast subspecialists) reading 260 images with AI support and 260 without AI support, in 2 sessions at least 4 weeks apart.

Intervention and comparator

Intervention: exams read by radiologists with PowerLook Tomo Detection 2.0 decision support (since rebranded as ProFound AI, iCAD).

Comparator: exams read by same radiologists without AI support.

Key outcomes

With AI, average radiologist AUROC increased by 5.7% (95% CI 2.8 to 8.7), and reading time was improved by 52.7% (95% CI 41.8 to 61.5). Case-level sensitivity and lesion-level sensitivity increased by 8.0% (95% CI 2.6 to 13.4) and 8.4% (95% CI 2.9 to 13.9) respectively. Specificity increased by 6.9% (95% CI 3.0 to 10.8), and recall rate in non-cancers decreased by 7.2% (95% CI 3.1 to 11.2). All results were statistically significant ($p<0.01$).

Strengths and limitations

Exams selected were enriched with exams of people with cancer, but readers were blinded to proportions. No study cases were used for development or training of algorithm. Those reading exams had not practiced at the acquisition sites. There was a randomised reading order to limit recall bias. Readers knew they were being timed but were blinded to the measurement. Sample size calculation were used, and study end points were explicitly stated. The method of random selection of exams from those eligible for analysis was not disclosed.

Study funded by iCAD, and 6 authors are either employees of or consultants for the company. Data and publication were controlled by an author not affiliated with the company.

Conant et al. (2019b, conference abstract, presented at ECR 2019)

Study size, design and location

Retrospective clinical validity study in the same study dataset as Conant et al. (2019a).

Intervention and comparator(s)

Intervention: exams read by radiologists with ProFound AI (version 2.0, iCAD) decision support.

Comparator: exams read by same radiologists without decision support.

Key outcomes

Outcomes reported included those reported in Conant et al. (2019a), plus outcomes relating to subgroups of people with cancer. Average radiologist sensitivity improved 6.8% with AI for soft tissue lesions, 6.2% for invasive carcinomas, 12.0% for calcifications only, and 14.6% for ductal carcinoma in situ (DCIS). Specificity improved 7.9% with AI for soft tissue, and 8.4% for lesions with BI-RADS classifications of 1 or 2. For calcifications only, average radiologist specificity reduced 2.7%, and for all exams not showing cancer it improved 6.9%. Reading times in exams showing cancer were reduced 39.5% for soft tissue, 41.4% for invasive carcinomas, 46.8% for calcifications, and 40.3% for DCIS. For exams without cancer, reading times were reduced 54.4% with AI for soft tissue lesions, 59.1% for BI-RADS classifications of 1 or 2, 43.6% for calcifications only, and 55.9% overall for all of these exams.

Strengths and limitations

This is a conference abstract that has not had peer review and lacks detail. Other strengths and limitations relating to the dataset and study design are as reported for Conant et al. (2019a).

Zebra Medical Vision Ltd (2020, unpublished data from Food and Drugs Administration)

Study size, design and location

Diagnostic accuracy study of AI in FFDM, including a retrospective cohort of 835 FFDM screening mammograms from the UK, US and Israel. Ground truth classifications were 435 biopsy-confirmed cancers, and 400 exams with no cancer, confirmed by 2-year follow up. Mammograms were selected to include a representative range of lesion types, breast densities, ages, and histology types.

Data is unpublished from Food and Drugs Administration section 510(k) premarket notification: performance data section.

Intervention and comparator

Intervention: HealthMammo Software (Zebra Medical Vision).

Comparator: ground truth mammograms showing cancer confirmed by biopsy and mammograms not showing cancer confirmed by a negative follow up of 2 years.

Key outcomes

HealthMammo showed an AUROC of 0.9661 (95% CI 0.9552 to 0.9769). In standard mode, the sensitivity was 89.89% (95% CI 86.69 to 92.38) and specificity was 90.75% (95% CI 87.51 to 93.21). In high sensitivity mode, sensitivity was 94.02% (95% CI 91.39 to 95.89) and specificity was 83.5% (95% CI 79.55 to 86.82). In high specificity mode, the sensitivity was 84.41% (95% CI 80.41 to 87.27) and specificity was 94.00% (95% CI 91.23 to 95.94).

Strengths and limitations

The study is unpublished and so not peer reviewed. It uses a relatively small sample size. Images are from 3 countries including the UK, and so could be representative of target population. There is a risk of selection bias because it was not reported how included cases were selected. Although intended for screening, the dataset is enriched with exams of people with cancer, and no comparison has been made with human readers.

Sustainability

No sustainability claims have been made by the companies.

Recent and ongoing studies

ScreenPoint Medical advised that several international clinical studies of Transpara are ongoing using unenriched datasets. The focus of these is early detection, reducing interval cancers and workload reduction. Also, 2 prospective clinical trials using Transpara are being set up. Several papers are currently under review in journals, with several more in preparation. There are 6 presentations planned for the Radiological Society of North America (RSNA) 2020 annual meeting.

Zebra Medical Vision advised that several international clinical studies are ongoing. These use HealthMammo to validate and collect evidence of the benefits of using the system in a screening setting.

iCAD advised that several international clinical studies of ProFound AI for 2D Mammography and ProFound AI for DBT are ongoing. These include a focus on reducing the number of interval cancers and improving cancer detection rates in screening.

Expert comments

Comments on this technology were invited from clinical experts working in the field and relevant patient organisations. The comments received are individual opinions and do not represent NICE's view.

All 3 experts were familiar with or had used AI technologies before, but not necessarily those in the scope of this briefing.

Level of innovation

All experts considered the technologies to be novel or innovative. One indicated that AI is not currently used in standard care, and another highlighted that it could change the way breast screening is offered. All experts mentioned that other similar, competing technologies are in development. However, these may not yet be commercially available. One expert expressed a desire for more technologies to be included in the evidence section. They provided publications for these technologies to be included. NICE excluded some technologies based on their published process and methods as described in the technologies section. One expert mentioned related technologies, including abbreviated MRI and automated ultrasound. They also mentioned ongoing research to refine the screening intervals or tests done, according to risk of breast cancer.

Potential patient impact

All experts agreed that there were potential benefits for patients, including reducing the number of unnecessary recalls, extra visits, and the anxiety these may cause. One expert felt that variability in the recall threshold between different screening services could be reduced, and care could be better standardised. One mentioned a potential decrease in breast cancer mortality because of earlier detection of screen-detected and interval cancers. Another felt that cancer detection would increase, but that this would not improve false negative rates relating to interval cancers. One expert saw an opportunity to develop algorithms that better detect the highest grade cancers, and another suggested that the need for needle biopsy of low-risk lesions could be reduced. One expert raised the important concern that without AI technologies, the NHS may struggle to continue to give breast screening to the current people who are eligible, because of the diminishing workforce. This could lead to patient harm. One expert also felt that those presenting with symptoms could particularly benefit from AI technologies.

Potential system impact

The overall financial impact of adopting the technologies was unclear. One expert suggested that it could be cheaper, and another expected it to cost roughly the same, or perhaps more. Two experts felt that cost savings could be made if the technologies reduced unnecessary recalls, but this would depend on the technology cost. Two experts suggested changes to the pathway by reducing the number of human readers for each case, or by eliminating a human reader entirely. One of these suggested that this would lead to more efficient use of time, and the other expert commented on reducing time taken to read mammograms. Two experts expected that adoption of AI technologies could reduce the number of people needed in the service to read screening mammograms. Two experts also highlighted shortages in the workforce. One expert expressed concerns that without support, such as that offered by these technologies, the Breast Screening Programme may struggle to continue to offer the service to those eligible. Two experts felt that there would be significant work needed to ensure the computing infrastructure of the

National Breast Screening Service, and its Picture Archive and Communication System (PACS) was able to support adoption. One expressed that this will be a major issue. Two experts also felt that if AI technologies were used to help those reading mammograms, and not as an independent reader, specialist training would be needed.

General comments

One expert felt the role of the technology is not yet clear. One expert felt it was additional to current standard care. The other expert agreed with this if it was used as a decision support tool, but thought it could replace current standard care if used to read mammograms independently. Two experts referred to usability issues with computer aided detection, the predecessor to AI. All experts felt that more information was needed around how a human reader would interact with the software. No specific safety or regulatory concerns were identified, but 2 experts anticipated potential medicolegal issues if human readers were replaced and cancers were missed. Two experts suggested a number of areas lacking evidence, including the localisation of cancers, the spectrum of disease detected, and independent evaluation of AI technologies in a screening population. One expert highlighted that studies of US radiologists have very low applicability to the UK, because of their higher recall rate and lower reading volume per year. One expert stressed the importance of making sure that the performance of these technologies does not decline if mammography vendors update the software on their machines.

Expert commentators

The following experts contributed to this briefing:

- Dr Sian Taylor-Phillips, professor of population health, University of Warwick. Dr Taylor-Phillips authored the UK National Screening Committee (NSC) position statement, and is funded by an National Institute for Health Research Career Development Fellowship to develop guidance on evaluation of new tests for UK screening programmes. She is also a member of the UK NSC Artificial Intelligence (AI) Task Group (unpaid) and has been commissioned by Public Health England to evaluate mammography AI in breast screening on behalf of the UK NSC.
- Dr Matthew Wallis, consultant radiologist, Cambridge University Hospitals (CUH) NHS Foundation Trust. Dr Wallis co-authored evidence presented in the published evidence section of this briefing. CUH received grant funding related to the OPTIMAM database.

- Dr Lucy Warren, clinical scientist, Royal Surrey County Hospital (RSCH) NHS Foundation Trust. Dr Warren works on the OPTIMAM image database (OMI-DB). Images from this have been shared with over 40 AI companies and may have been used to train technologies included in this briefing. She is also involved in a collaboration between RSCH and ScreenPoint Medical to independently evaluate their AI algorithm.

Development of this briefing

This briefing was developed for NICE by Newcastle External Assessment Centre. NICE's interim process and methods statement sets out the process NICE uses to select topics, and how the briefings are developed, quality-assured and approved for publication.

ISBN: 978-1-4731-3716-5