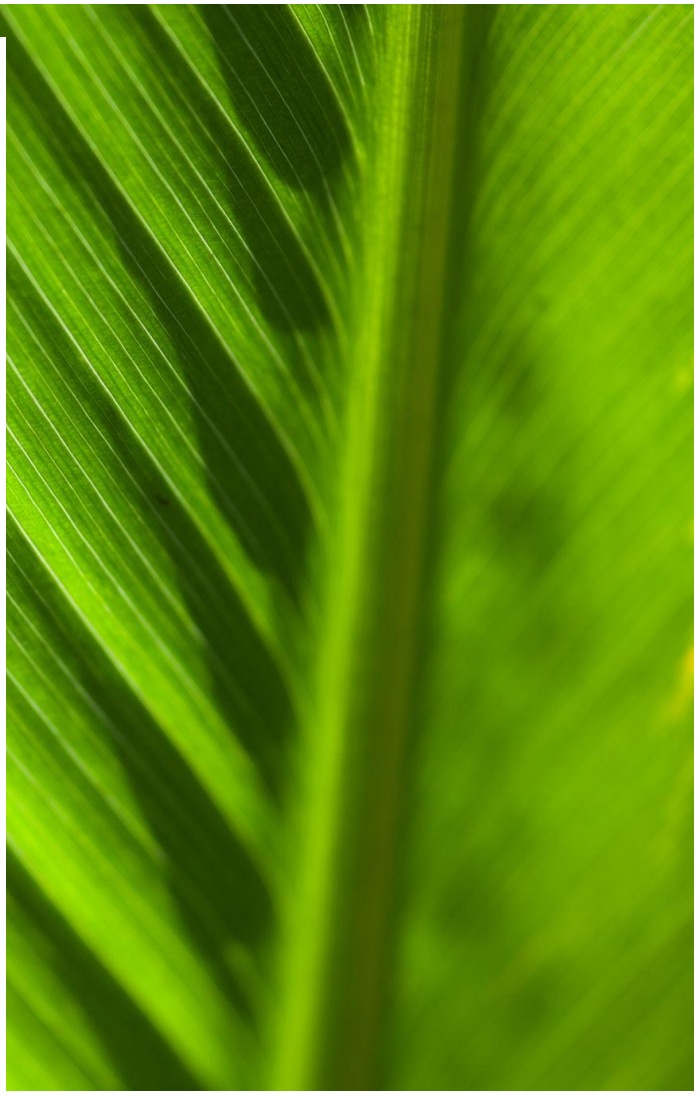
Evaluation of EMRAD AI in Breast Screening Project: Final Report

Summary



DECEMBER 29, 2020

TaoHealth Research & Implementation
Lead author: Dr Niamh Lennox-Chhugani

TaoHealth
Research
Implementation

This report is a brief summary of the evaluation. A more detailed technical report is available on request from info@taohealth.co.uk

Acknowledgements:

This evaluation would not have been possible without the cooperation of several people and organisations.

Staff from across the four study sites gave freely their time to support and engage in this study either as staff members within the Breast Screening Units or as ordinary women who are either attending breast screening now or will in the future. The Research & Innovation teams from the sites supported this work proactively encouraging people to take part in the study. Women from the general public also took part in the survey and in the subsequent focus groups.

The NHS EMRAD AI project team were instrumental in facilitating introductions to the service teams and other local stakeholders and played a proactive role in removing the barriers to the evaluation when they arose without seeking to influence the independence of the evaluation.

Contents

Introduction and background

Overview of the evaluation

This evaluation was conducted from October 2018 – September 2020 by the research team at TaoHealth Research & Implementation (TaoHealth). It was commissioned by NHS EMRAD[1] as part of the NHS England Wave 2 Test Beds Programme to: deliver project learning during and at the end of the project; inform project implementation and future investment decisions locally and nationally (as the NHS looks at how it can derive value from digital technologies), including artificial intelligence in the future.

NHS EMRAD (EMRAD), together with two commercial digital technology companies, Kheiron Medical Technologies (Kheiron) and Faculty (formerly ASI Data Science), GE Healthcare (EMRAD's provider of radiology IT systems) and East Midlands AHSN bid for funding under the NHS England Test Beds Programme in 2018 to train and implement artificial intelligence (AI) solutions within the national breast screening programme. Neither of the products being 'tested' were market-ready although the product from Kheiron was CE-marked.

The theory of change underlying the EMRAD screening imaging innovation programme is that the two AI tools, one an algorithm-based clinical decision support tool and the other a machine learning pathway optimisation tool, in the context of a scalable radiology IT system, will optimise the efficiency of the overall service, allow the same number of staff to process more scans-reducing reporting delays, freeing up staff to deliver high value activity and enable prompt and accurate diagnosis and treatment.

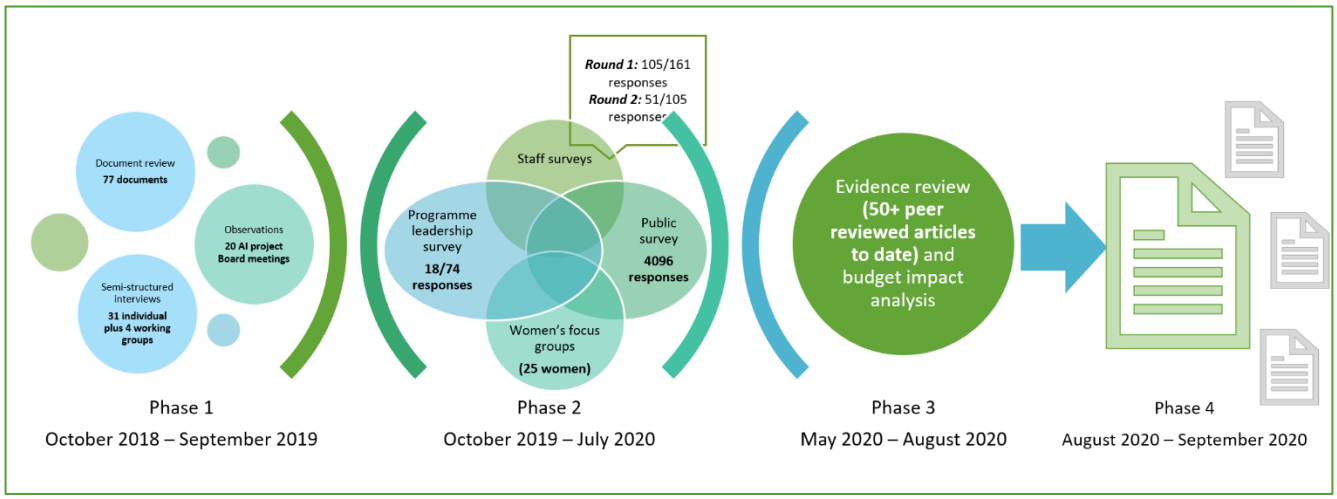
The main ***aims of the evaluation*** were to:

1. Understand the effect of combinatorial innovation in the NHS Breast Screening Programme (NHSBSP) on coverage and utilisation, user satisfaction, improvement in workforce productivity and improvement in health and care services with a specific focus on:
    a. Assessing the performance of a CE marked, AI tool to assist radiologists by reading breast images [Mia™] using a deidentified retrospective data set from two breast screening units, Nottingham Breast Institute (NBI) and United Lincolnshire Breast Screening Service (ULBSS) in this phase.
    b. Assessing the use of machine learning models with proven effectiveness in non-healthcare, live environments [developed by Faculty] to optimise the operational aspects (clinic scheduling and resource allocation) of the breast screening service, boosting system capacity, reducing delays and improving patient experience.

[1] NHS EMRAD stands for East Midlands Imaging Network and is a partnership of seven NHS trusts (Chesterfield Royal Hospital NHS Foundation Trust, Kettering General Hospital NHS Foundation Trust, Northampton General Hospital NHS Trust, Nottingham University Hospitals NHS Trust (host organisation), Sherwood Forest Hospitals NHS Foundation Trust, United Lincolnshire Hospitals NHS Trust, and University Hospitals of Derby and Burton NHS Foundation Trust). These trusts run 11 hospitals, covering more than five million patients.

2. Understand clinical and patient attitudes towards this technology, with a view to wider roll-out across the NHS.
3. Understand and share 'lessons learnt' as a nationally relevant template for the combined deployment of clinically and operationally focused AI tools in healthcare.
4. Make recommendations about future real-world testing and scale-up of AI technologies in the health system.

The tools were tested at Nottingham University Hospitals NHS Trust and United Lincolnshire Hospitals NHS Trust in the breast screening services.



The evaluation does not include a comprehensive assessment of the safety and effectiveness of Kheiron's Mia™ tool (the subject of a separate HRA approved study) or any assessment of Faculty's tool. It explores the process of testing and developing the tools in the real world, perceptions around the use of AI tools in the context of the NHSBSP, early evidence of the effect on NHSBSP performance and the process of innovating in the NHS.

Artificial intelligence (AI) *can both be viewed as 'general' or 'narrow' in scope. Artificial general intelligence refers to a machine with broad cognitive abilities, which is able to think, or at least simulate convincingly, all of the intellectual capacities of a human being, and potentially surpass them—it would essentially be intellectually indistinguishable from a human being.*

Narrow AI systems *perform specific tasks which would require intelligence in a human being, and may even surpass human abilities in these areas. However, such systems are limited in the range of tasks they can perform.*

The terms 'machine learning' and 'artificial intelligence' are also sometimes conflated or confused, but machine learning is in fact a particular type of artificial intelligence which is especially dominant within the field today. ***Machine learning (ML)*** *gives computers the ability to learn from and improve with experience, without being explicitly programmed. When provided with sufficient data, a machine learning algorithm can learn to make predictions or solve problems, such as identifying objects in pictures or winning at particular games, for example.*

Source: Select Committee on Artificial Intelligence (2018)

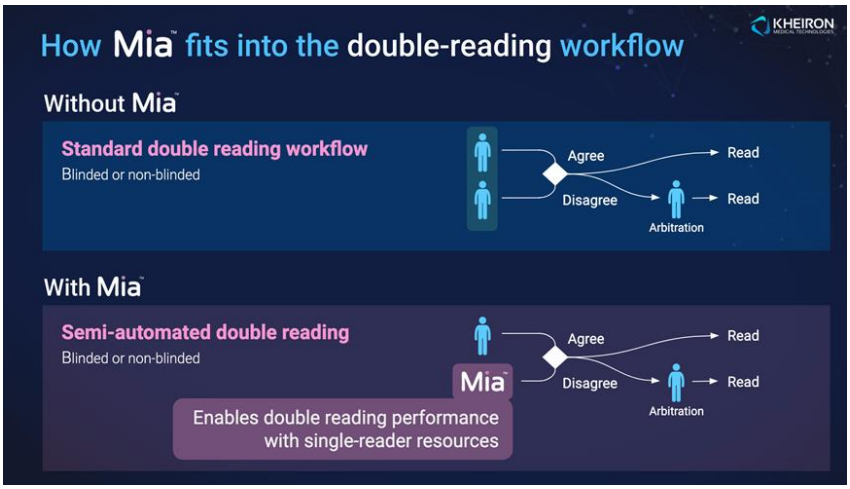
The tools being tested

Kheiron Medical Technologies - Mia™

Kheiron's Mia™ AI tool for breast cancer screening and reporting is CE-marked. It has been used to retrospectively read a large number of mammograms from different manufacturers, to train the algorithm, validate it and to determine which of the software's various operating points would be best used for future prospective pilot in the NHS breast screening programme (NHSBSP).

The way that Mia™ fits in to the NHSBSP workflow is summarised in Figure 1 which is reproduced from a document submitted by the partners to NHSX in September 2019.

Figure 1: Mia™ AI mammogram reader



Faculty AI – Service Optimisation Tool

Faculty AI's 'Platform' (formerly SherlockML) software is a secure machine learning environment for accessing and manipulating enormous amounts of data, designing, and testing AI models, and deploying those models in live environments. The platform has already been used on more than 200 commercial projects.

The main outputs of Faculty's contribution to the project are:

1. A synthetic NBSS data set (SDS) co-owned by Faculty and NUH (on behalf of EMRAD);
2. A deployment environment in NUH that is a controlled and governed environment in which new AI products can be deployed safely;
3. A round length[2] machine learning (ML) tool that includes (a) an interactive information dashboard that provides situational awareness to service managers about multiple dimensions of service activity and demand that can be used to support day to day decision making and (b)

[2] Round length is the term used to describe the time between breast screening appointments for each woman. In England this is usually 3 years.

a scenario planning tool that allows managers to model scenarios and predict demand and capacity changes under modelled conditions.

The round length machine learning tool aims to enable managers to make the best possible use of scarce resources like radiologists' time and expensive machinery, and to reduce stress on the clinical and administrative workforce delivering the programme.

Findings

Overview

The key findings of the evaluation are:

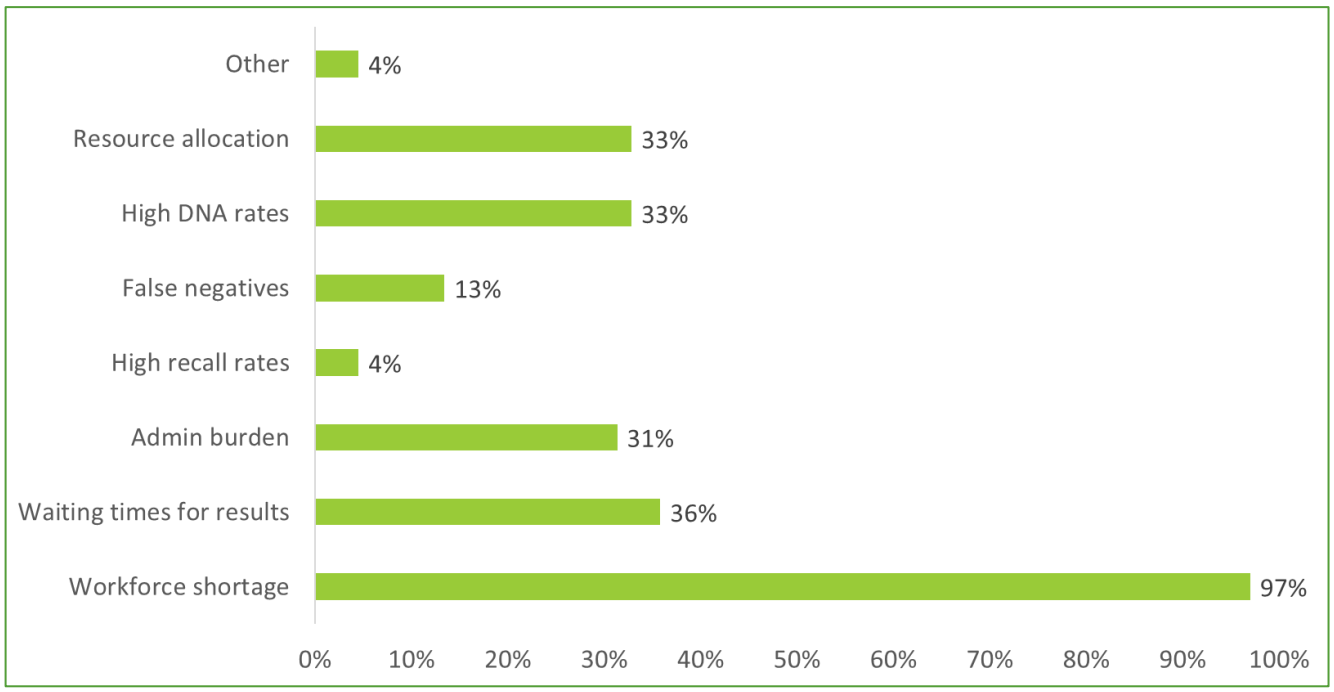
1. Clinicians working in the breast screening service are positive about the potential for AI as a second reader in breast screening but want to see more evidence from clinical trials and real-world validation.
2. Service administrative staff and managers (non-clinical) are less convinced about the potential for AI in service optimisation but have also seen less obvious development in this area over the period of the project.
3. Adult women of and under screening age are generally positive about the introduction of AI into the breast screening service but they also want to see evidence of effectiveness and safety especially where the technology is used as a second reader.
4. The same factors influenced the early stages of implementation of these novel technologies as any other digital health technology.
5. Some factors that are additional and unique to AI were evidenced during the implementation of this project.
6. The results of the retrospective study using the AI mammogram reader (Mia™), when used to model the impact of the technology on resourcing showed that using AI as a second reader could reduce the time required from human readers (radiologists and reporting radiographers) by 42%.

1. Clinicians attitudes to the use of AI in breast screening (AI second reader)

Key finding: Clinicians working in the breast screening service are positive about the potential for AI as a second reader in breast screening but want to see more evidence from clinical trials and real-world validation.

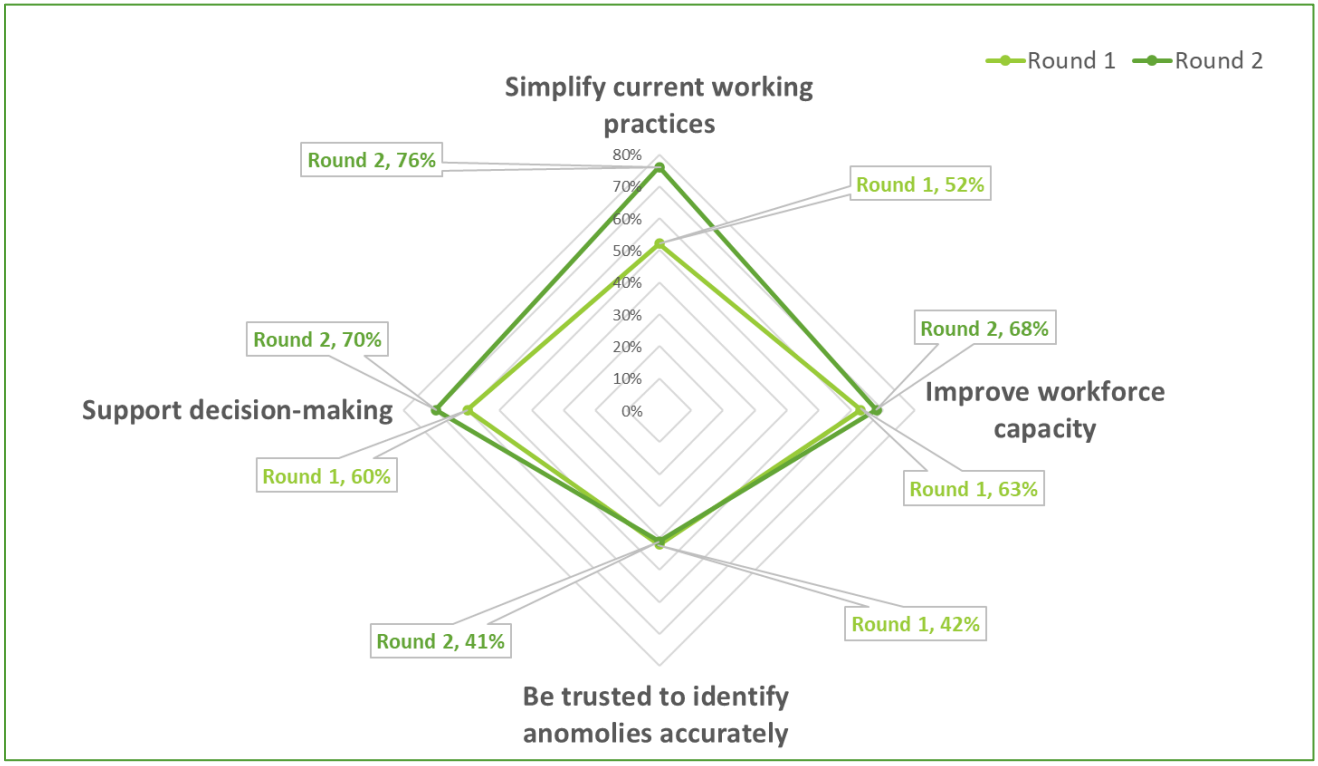
To understand the extent to which the perceived benefits of Mia™ aligned to actual challenges faced by the service, we asked clinical staff what they thought were the greatest challenges facing the service at present. We asked them to select their top three from a long list of challenges based on a literature review and consultation with a small group of clinical staff. In both rounds of the survey, workforce shortage was by far the biggest concern (95%) and high 'do not attend' (DNA) rates the next biggest (33%) [Figure 2].

Figure 2: Challenges facing breast screening services - clinicians views



Whilst the clinicians were positive about the potential benefits of AI in breast cancer screening in simplifying current working practices and in supporting decision-making, they are less convinced of the potential to improve workforce capacity and there is no change in their view that AI can be trusted to identify anomalies accurately [Figure 3].

Figure 3: Perceived benefits of implementing AI as a second breast image reader - Clinical staff



Asked about their level of comfort using AI as a second reader in the process of reading population breast screening mammograms, clinical staff moved from 51% agreeing they would be comfortable, to 54% over the period although the proportion saying they would not be comfortable also went up from 4.5% to 8% indicating that clinicians had more decisive views by round 2 [Table 1].

Table 1: Comfort using AI in breast screening

I would be comfortable with AI technology being used as a second reader in breast cancer screening.	Round 1	Round 2	Change
Strongly agree	8.96%	13.51%	4.55%
Agree	41.79%	40.54%	-1.25%
Neither agree nor disagree	44.78%	37.84%	-6.94%
Disagree	4.48%	8.11%	3.63%
Strongly disagree	0%	0%	0.00%

When asked to expand on this qualitatively and indicate what would give them greater confidence that AI second readers were safe, effective, would improve their working life and the experience of women coming into the service, the largest proportion (43%) want to see evidence from trials and 24% want the results of clinical audits in situ.

Quotes from the survey

"Initially it would have to be introduced as an additional tool rather than a replacement. The use of audits would then be able to determine the effectiveness and benefits of using AI."

"[I need to see] publicised, peer reviewed results against real life."

Clinicians were positive about the potential effect of introducing the AI reader on the experiences of women attending the service with 51% in round 2 agreeing that it would have a positive effect on women's experience, increasing from 42% in round 1.

Clinical staff expressed concern about the impact of introducing AI as a second reader in terms of possible job losses and reduced opportunities for reporting radiographers to develop their skills. On the other hand, some saw potential in the introduction of AI second readers in reducing severe pressure currently placed on the screening workforce, releasing staff to activities that are "patient-facing" as well as helping with the continuing workforce shortfall

Clinical staff were also concerned about the safety, accuracy and reliability of the AI reader and saw the publication of clinical trials as an essential prerequisite to adoption of the technology in the breast screening service workflow.

Quotes from the survey

"I believe the use of AI could be an exciting development in improving the service, however I would want to see the evidence that it is a safe tool to use."

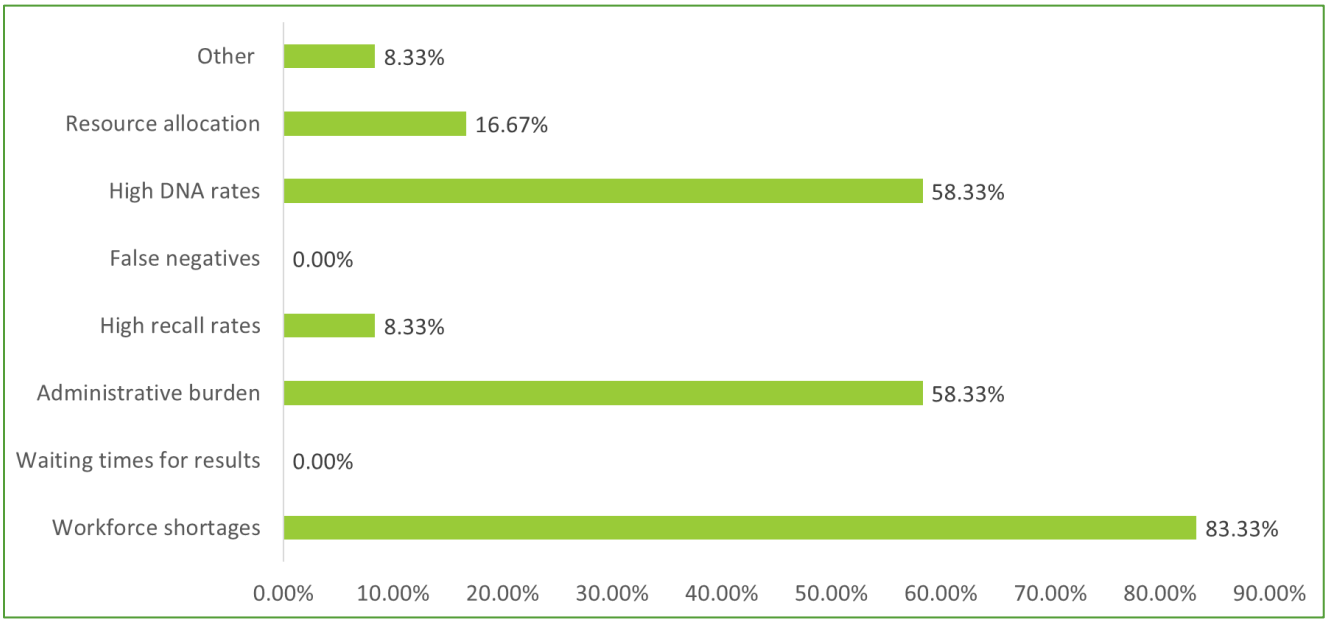
"I think AI tools in the breast screening programme may be useful in booking patients, sending them invitation to screening etc, but probably not advisable/safe to use in image reporting."

2. Service administrators and managers attitudes to the use of AI in breast screening (Service optimisation tool)

Key finding: Service administrative staff and managers (non-clinical) are less convinced about the potential for AI in service optimisation but have also seen less obvious development in this area over the period of the project.

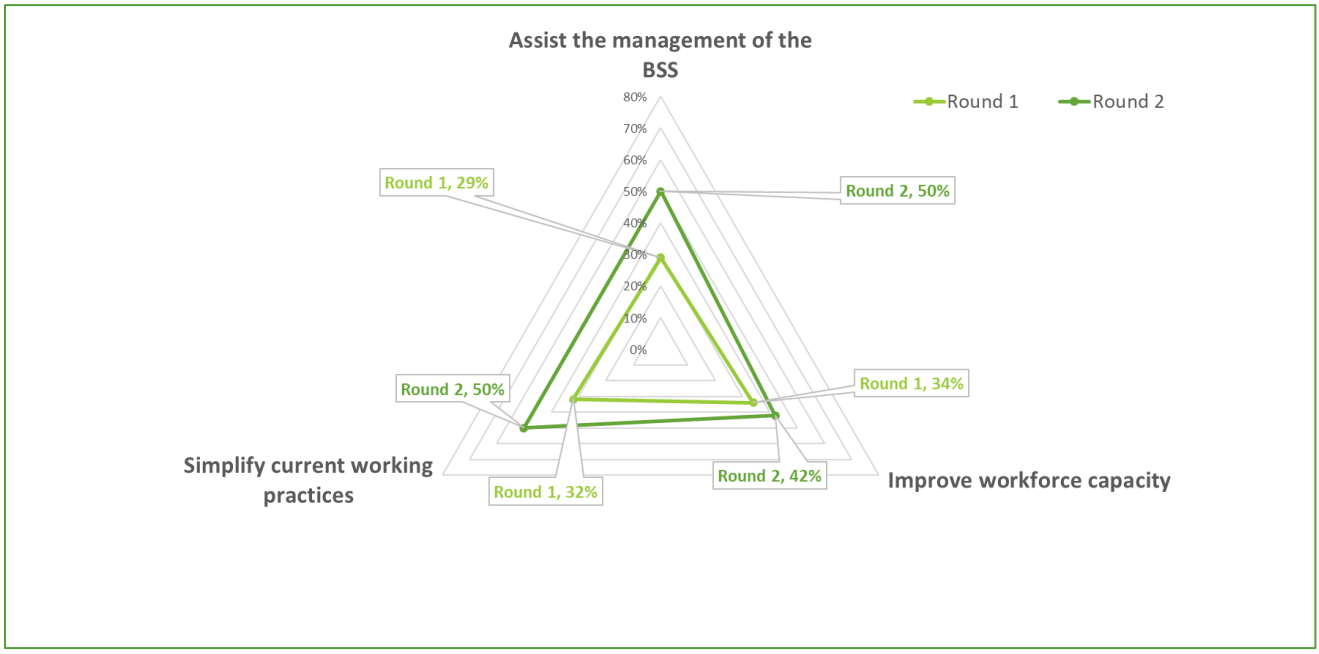
The top three challenges for the service according to non-clinical staff remained the same across the two surveys:workforce shortages (83%); high 'do not attend' (DNA) rates (58%); and administrative burden (58%) [Figure 4].

Figure 4: Challenges faced by breast screening services - non-clinicians views



By the time of the second survey round, non-clinicians were more likely to agree that AI would have potential benefits in supporting the management of breast screening services (BSS) against all three dimensions [Figure 5].

Figure 5: Perceived benefits of implementing AI to support operations - Non-clinical staff



Quote from the survey

"I think if this will increase the experience of women it will be a good thing. We do get a lot of anxious women wanting to know results sooner than our protocol." "It will be easier to assign sensitivity thresholds with an AI to reduce false positives."

The non-clinical workforce was less positive about the predicted benefits of introducing AI service optimisation tools into the workflow than clinicians were about the AI second reader (mean score 47.33% compared to 63.75%). It is likely that the absence of any real-world testing of the service optimisation tool accounts for this difference. This was reflected in the common assertion that respondents to the survey did not have enough information to express a view on the introduction of AI into the service management workflow. Some concerns were expressed about the potential impact of job losses in relation to the service optimisation tool.

Quotes from the survey

"Because I do not know much about it and nothing artificial is usually good. However, if I don't understand how it works then my answer will be biased."

"I feel uneasy about it until there is more information and research to prove the reliability and benefit and also wonder what this means for jobs."

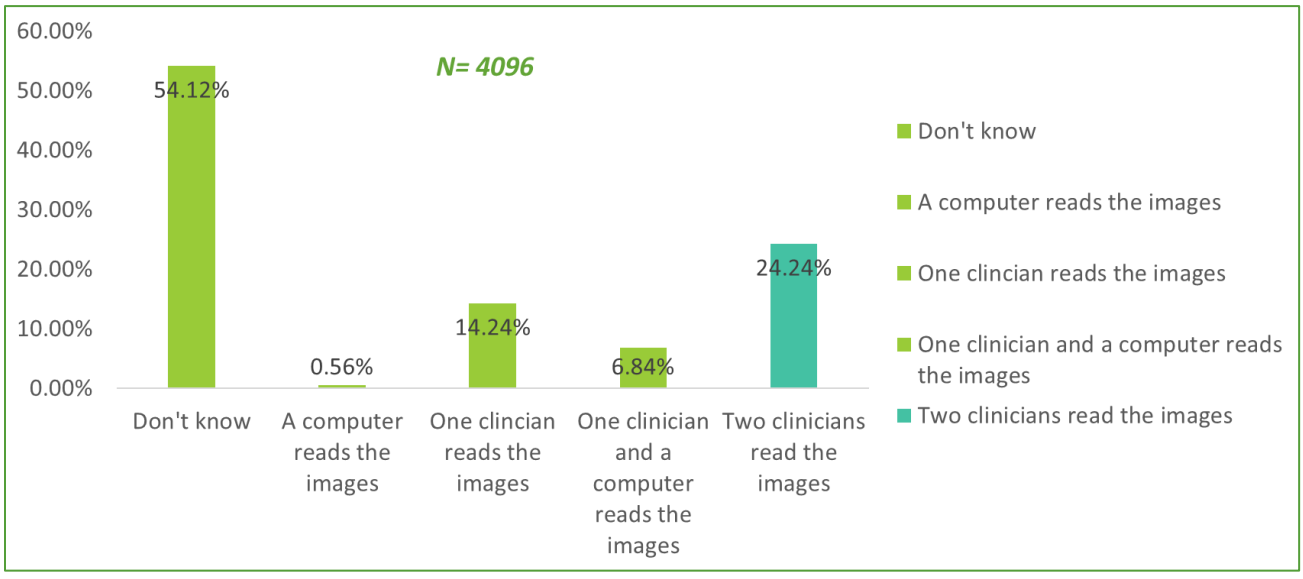
"I would like to see it in place first before I make a comment."

3. Women's attitudes towards the use of AI in the breast screening programme

Key finding: Adult women of and under screening age are generally positive about the introduction of AI into the breast screening service but they also want to see evidence of effectiveness and safety especially where the technology is used as a second reader.

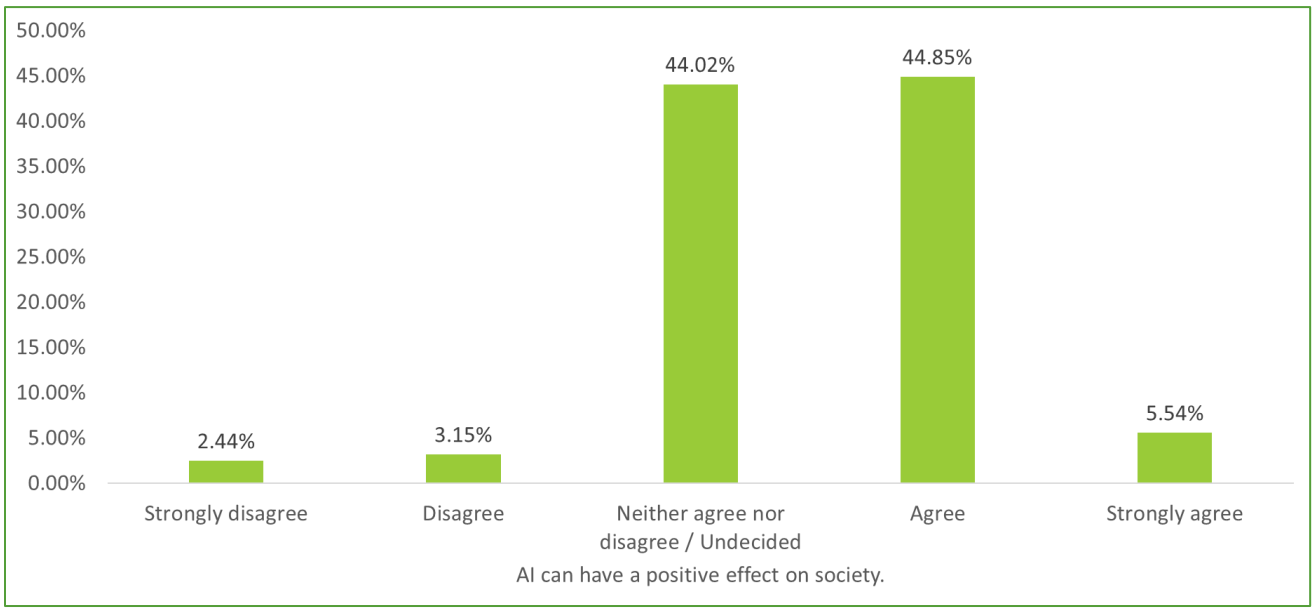
The majority of women surveyed (54%) did not know what the current mammogram reading process and only 24% had an accurate understanding of the process [Figure 6].

Figure 6: What is the process for reading mammograms?



Fifty percent of survey respondents were positive about the potential effect of AI on society but were almost as likely to be undecided on this (44%). Women under screening age were more likely to be positive about AI at 53% as opposed to 47% for women of screening age [Figure 7].

Figure 7: Artificial intelligence can have a positive effect on society



Respondents were asked to provide a free text response on their views on using AI in the breast screening programme, both as a second reader and to support programme management.

Quotes from the survey

"A computer if programmed correctly will not produce any errors when a human can." (50-59 years)

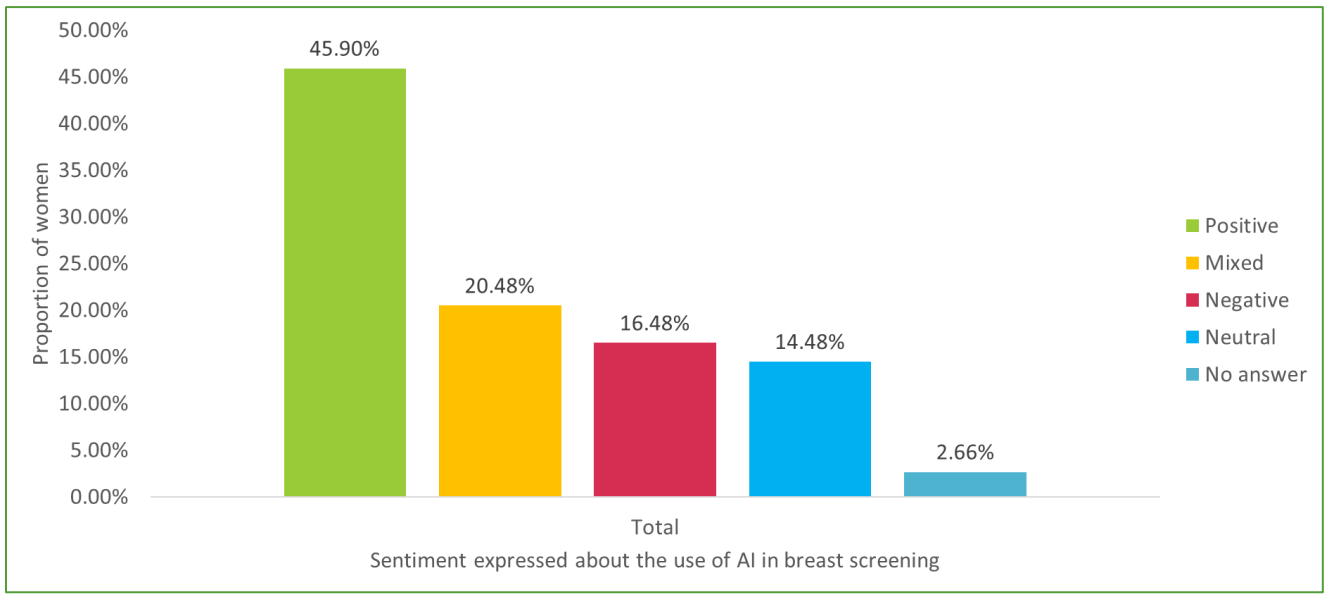
"The removal of emotion and personal circumstances can lead to more consistent and fair decisions." (40-49 years)

"AI can be used to aid in education and understanding of many issues." (20-29 years)

"It can be used in a range of areas to improve speed, accuracy, reduce costs of certain tasks. It is already used in lots of ways. It can standardise tasks/tests etc. as not prone to same biases of humans (other biases may exist and need to be taken into account). Reduce problems caused by human error and differences in my opinion." (30-39 years)

Sentiment analysis of these free text responses showed that the largest proportion of women were positive about this (46%) with the next largest (20%) expressing mixed views and 16% expressing a negative view. Women of screening age were more strongly positive of the two groups and women under screening age more likely to hold mixed views of the two groups [Figure 8].

Figure 8: Views on the use of AI as part of the breast screening programme



Perceived benefits expressed most frequently in the free text were improved reliability (n=263), increased efficiency (n=162), and greater safety (n=139). Most women were stated they were not sure what the benefits would be (n=543). Many women expressed the view that AI in breast screening would and *should* happen (n=847, 78% women who provided a response to this question) in the future.

Potential benefits were explored in more detail as part of the **focus groups** that followed the survey. Many of the women who took part expressed the view that the use of AI in healthcare and specifically in the breast screening programme was inevitable (n=11), with some seeing a positive contribution being made by AI (n=4). The main benefits that women saw AI in breast screening offering were in increased efficiency (n=23), improved reliability (n=12), improved outcomes (n=8) and improved safety / fewer errors (n=8). They also hypothesised that introducing AI into the breast screening programme might release staff to higher value activities and save money for the service (n=6) and help address the workforce shortage within the breast screening programme (n=17).

Quotes from the focus group discussions

"My GP has introduced AskMyGP – I was blown over from the response - personalised to me. I would find this easier to do and would prefer to spending 2-3 hours going to the surgery." ***(Woman under screening age)***

"I'd like to think that this AI will shorten the time taken from the mammogram being taken to getting the results." ***(Breast cancer survivor)***

"AI in the background - you could really get a lot of people through the system. I'd have a level of comfort from a mammogram point of view. I guess if there was a problem you would have a review by a radiologist. I guess on the back end you would still be getting some kind of personal touch." ***(Woman of screening age)***

"I wanted to choose a mix of AI and human. So much of my life was waiting for results and being on hold, I think it was about speed and accuracy for me. I don't have enough experience of normal mammograms to know how to answer." ***(Breast cancer survivor)***

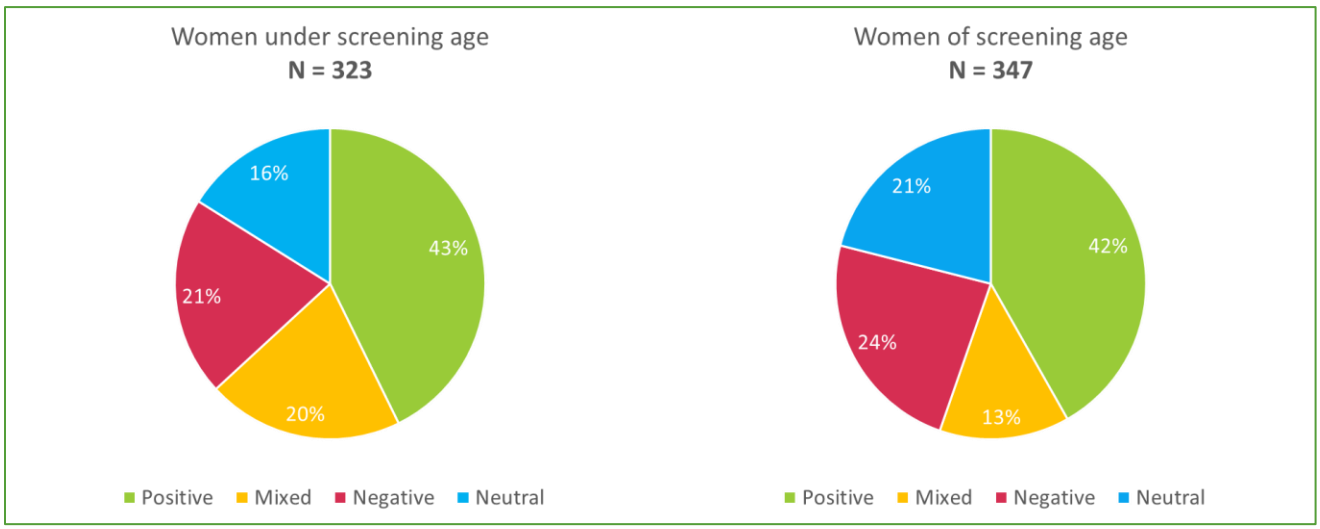
The word "trust" was mentioned 137 times in the public survey in the free text response to two questions which sought to understand more detail on respondents' attitudes to the use of AI.

- Tell us more about why you selected the level of agreement with ***Artificial intelligence can have a positive effect on society*** that you did.
- How would you feel about artificial intelligence being used to read mammograms?

When we included synonyms such as "sure", "confident" and "believe", this incidence increased to 696**.** Almost all the respondents who mentioned trust, either did not trust AI or felt that it could not be trusted if used in isolation without human oversight.

Of the women who chose to explain why they had given the response that they did to the statement ***"Artificial intelligence will have a positive effect on society"*** (n=670), when these responses were analysed for sentiment, 39% made mixed or negative statements about the impact of AI. Women under screening age had slightly more mixed or negative views at 41% than women of screening age at 37% [Figure 9].

Figure 9: Sentiment analysis of free text in response to AI having a positive impact on society



Women who had a negative or mixed view of the effect of AI in society were unsure of why they felt this way in many cases (n=96) although they felt it was an inevitable part of their lives in the future (n=20), those that did express a view cited concern about the reliability and safety of technology (n=123), a lack of trust in the technology itself of the systems that sit around it (n=65), a fear about a combination of over-reliance on AI and job losses that might ensue (n=32) and the absence of the human touch in interactions (n=46).

Quotes from the survey

"If used well, AI has an important part to play in diagnosis of disease. However, there are also dangers of it being used to further profit-driven goals." (50-59 years)

"It reduces human interaction but I agree there are a lot of amazing applications of AI that can, for example, keep the disabled independent." (40-49 years)

"It has the potential for profound good or profound harm. It must be controllable." (20-29 years)

"AI has the ability to remove skills and development of people and learning is wasted. I'm concerned for my children in the further, I believe that humans will be removed from day to day things and life will miss person centred contact." (30-39 years)

When we compared the women's views of the effect of AI on society and their views on the use of AI in the breast screening programme, Women with positive views on the effect of AI on society as a whole were slightly more likely to hold positive views on the use of AI in breast screening but interestingly, also more likely than women who had a negative or undecided view on the effect of AI on society, to hold a negative view of AI in breast screening. In other words, women who have positive views about AI's effect on society have more decisive views in AI in breast screening (positive or

negative). Women with the highest proportion of negative views on AI in breast screening were those with neutral views on the effect of AI on society [Table 2].

Table 2: Cross tabulation of the perception of the effect of AI on society with perception of use in breast screening

	View of the use of AI in breast screening					
View on the effect of AI on society (rank)	**No answer**	**Negative**	**Mixed**	**Undecided**	**Positive**	**Grand Total**
Strongly disagree	3	26	12	7	52	**100**
Disagree	5	77	15	16	16	**129**
Neither agree nor disagree / Undecided	68	414	329	423	569	**1803**
Agree	29	148	452	139	1069	**1837**
Strongly agree	4	10	31	8	174	**227**
Grand Total	**109**	**675**	**839**	**593**	**1880**	**4096**

When the perception of the use of AI in the breast screening programme was compared with women's self-reported understanding of AI overall, we found that the higher women rated their understanding of AI, the more likely there were to have a positive view of its use in the breast screening programme [Table 3].

Table 3: Cross tabulation of self-reported understanding of AI with perceptions of AI in breast screening

	View of the use of AI in breast screening					
Understanding of AI in general	**No answer**	**Negative**	**Mixed**	**Undecided**	**Positive**	**Grand Total**
None	33	66	25	90	90	304
Aware that it exists but little understanding	33	287	269	261	583	1433
Some understanding	40	310	525	236	1148	2259
Extensive understanding	3	12	20	6	59	100
Grand Total	**109**	**675**	**839**	**593**	**1880**	**4096**

The perceived concerns and risks of using AI in the breast screening programme raised by women in their free text responses were a lack of clarity around how the technology would be governed and regulated once in place (n=163) and the lack of 'human touch' that may result (n=143).

A large number of women (n=643) expressed the view that they expected that the AI tool being used as a second reader specifically would be rigorously tested and there would be robust evidence made available on its safety and accuracy. A small but not insignificant group (n=242 or 22%) of women who expressed a view about the future of AI in breast screening said that AI should not be used within the breast screening programme.

The main concerns that were expressed by women during the focus group discussions were the absence of the 'human touch' (n=37), lack of clarity around how the AI tools will be governed and potential discriminatory bias avoided (n=33) and how data privacy will be protected (n=25).

Quotes from the focus group discussions

"There are all those sci-fi movies where it goes rogue and I am not saying it is not completely far-fetched, most of it is, but I think it about having some real strong list of ethical principles about how you use it but in free market capitalism you are not going to have that are you? that would be bad for society, it will make money for the money people, it will leave behind the poor people and there will be some good people along the way who will do good things with it." ***(Women of screening age)***

"I think behind the scenes it is great, but I think you need a lot of face to face compassion and understanding." ***(Woman under screening age)***

"If there was some sort of consent, confidentiality, some sort of understanding of the rules. It would be nice to know that some sort of trustworthy organisation was monitoring it." ***(Breast cancer survivor)***

"I have been reading negative stuff about AI like facial recognition and how it's a bit biased - would it be biased against certain racial populations?" ***(Woman under screening age)***

The kind of action that women thought would address some of their concerns was designing the workflow so that humans would always be involved. For some women this was human oversight of the AI technology which undertakes most of the activity including decision-making (n=10), for others they see the human role as pre-eminent with AI used to augment clinical activity and decision-making (n=15). The women assumed that this technology would never be used without clear evidence of its effectiveness (n=24) and expected its impact on the equity of access to breast screening to be closely monitored (n=18) through governance processes.

Women were divided on whether they would want to be informed if AI tools were being used as part of their experience of breast screening but agreed overall that women should be given the information as part of the process of informed consent when taking part in the breast screening programme (n=15).

4. Factors common to digital health technology influencing the implementation and adoption of the AI solutions

Key finding: The same factors influenced the early stages of implementation of these novel technologies as any other digital health technology.

Information management and data governance:

Whilst there were some minor technical challenges during the project, the data challenges were more significant.

It was clear that ethical approval was a necessary first step for the retrospective study of the AI second image reader, but not clear that it would be required for the development of the synthetic data set. This took some time to clarify due to it novelty in the NHS, causing some delay to development.

The processes of data deidentification, transfer, cleaning and preparation for use in AI training and validation, and associated decision-making and sign-off, also took longer than planned as this had not been done on this scale before by any of the test-bed partners.

The development of the synthetic data set required access to the National Breast Screening System hosted by Public Health England and commissioned from Hitachi. This added a further layer of complexity and delays to decision-making on access to this data set to develop the synthetic data.

Organisational readiness:

This project was the first of its kind testing AI technologies in an NHS setting at this scale. Only one technology was market ready with a CE mark and it still needed to be trained and validated on a UK population. In that sense, the project was less a real-world test of AI and more a real-world collaborative development of AI that included an understanding of clinical and operational workflows.

The project was able to use the foundation of the EMRAD network. This network of seven NHS trusts had already successfully delivered a number of technology projects and had a standing experienced team of project management professionals with strong links into all of the member trusts.

Committed clinical leadership in the two test sites chosen for the project (Nottingham University Hospitals NHS Trust and United Lincolnshire Hospitals NHS Trust) ensured that the commercial partners had access to critical assessment of their ongoing development. The absence of any testing in the clinical or operational workflow meant that wider teams were less exposed to the project and its outputs than originally planned.

The absence of testing in clinical and operational workflows, the fact that neither commercial partner had established a price for their technology nor a procurement route to market meant that the exploitation of the technologies was incomplete by the end of the project. There are plans for this work to continue.

Project decision-makers, that is the group overseeing the project in the context of other network priorities, were split between those who thought the project had made the expected progress and those who thought it had made less than expected. The reasons they gave for slow progress were the

novelty of the technologies and the complexity of the project and they highlighted the importance of trust between the partners in the project as overcoming these challenges. They were impressed by the project's ability to adapt the project scope and learn iteratively from ongoing evidence.

Context

During the period of this project the General Data Protection Regulation (GDPR) came into force across the EU in May 2019. This placed new requirements for information governance particularly in the responsibilities of data controllers and processers.

Each new introduction of regulations or guidelines in respect of AI and data protection required the project team spending time reading and understanding the impact of the guidance for the project and communicating this to stakeholders, recommending changes to project delivery and implementing these.

Socio-cultural differences between commercial technology firms and public sector providers were evident from the outset. This 'business case' type approach to innovation in the NHS contrasted with the agile approach that was used by technology developers.

Project decision-makers survey

"One of the biggest challenges for me personally has been bridging the NHS-industry communications gap. The complexity of the technology, the scale, the novel nature of the task at hand, and the continuously steep learning curve that comes for companies working on the development and testing of these technologies, often makes it difficult to communicate externally at the right level of simplicity/ transparency/ cadence, whilst also not introducing too much uncertainty or confusion such that trust is damaged." ***(AI project programme leadership team member #17)***

The **Covid-19 pandemic** had several effects on the project. The national breast screening programme was suspended or paused in March 2020 until July 2020. Many staff in the two breast screening units that took part in this project were redeployed to support other services in their Trusts responding to increased demand although most of these have now returned to the breast screening units. Early in the national response, the project team produced a mitigation document which was shared with all stakeholders including Innovate UK and NHS England which set out the risks and how they would be managed and monitored through project governance.

5. Factors unique to AI influencing the implementation and adoption of the AI solutions

Key finding: Some factors that are additional and unique to AI were evidenced during the implementation of this project.

Information management and data governance:

This project was not testing real-world deployment of market-ready digital tools, but developing innovative tools using artificial intelligence in a real-world context. In the case of Mia™, this meant training, validating and testing their AI breast image reader on a large retrospective data set extracted from the GE Healthcare system that EMRAD Trusts use to process and store all medical imaging. Faculty first tried to develop their service optimisation tools using data extracted from EMRAD trusts but it became clear that this would not be possible and so they pivoted to developing a synthetic data set (SDS) based on an extraction of real-world data at one EMRAD trust site, then developing the service optimisation tools using this SDS.

Regulation and governance:

NHSX's AI Lab was set up in 2019 during the delivery of this project. One of the areas of focus for the Lab is the regulation of AI within the health system. During the project, NHSX introduced the *Evidence Standards for Digital Health Technologies* in March 2019 and this project took part in testing the first assessment questionnaire tool. This enabled to the project to reflect on some of the requirements of likely future regulation and provide feedback to NHSX on the tool.

The project was delivered against a backdrop of increasing awareness amongst the public about the use of AI in daily life, the role of algorithms in decision-making and the benefits and disbenefits of technology in general. It also surfaced some of the differences in the working practices and mind sets of people working in the technology sector and people, including clinicians, working in the NHS. These social factors were noted throughout the duration of the project. A small but not insignificant group (n=242 or 22%) of women we surveyed expressed the view that AI should not be used within the breast screening programme and it will be interesting to see if this group increases or decreases in time.

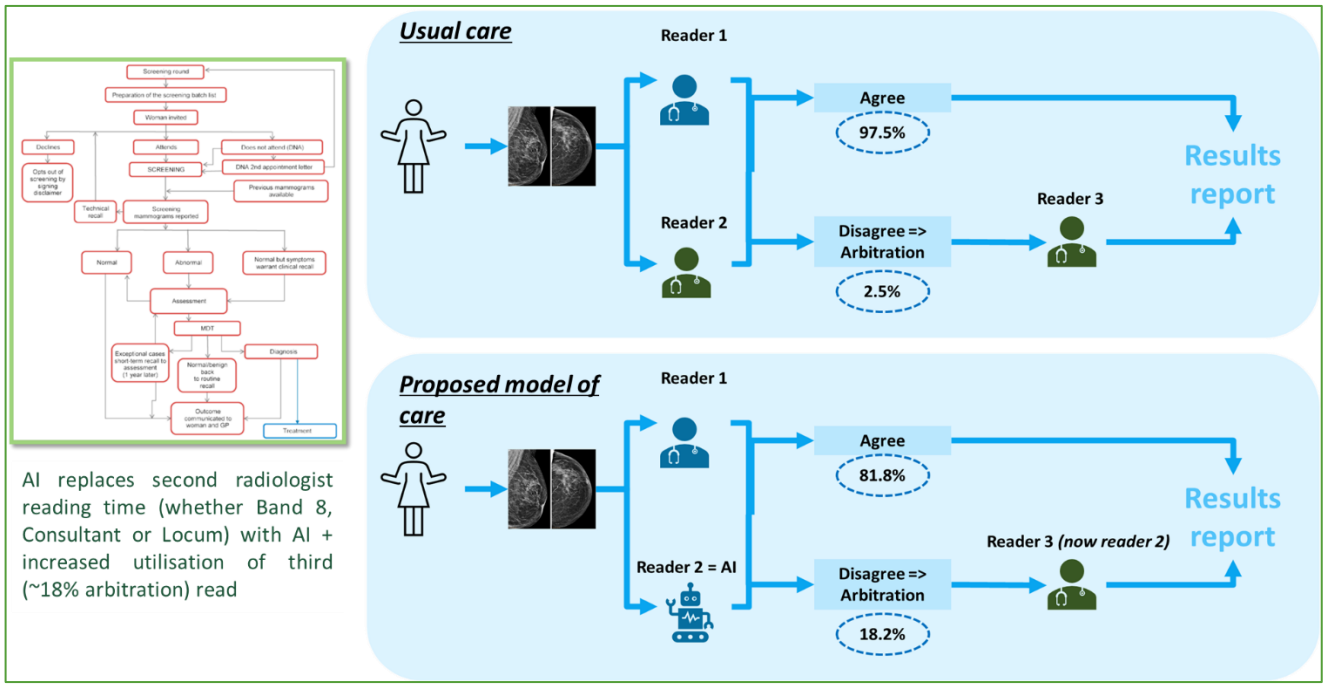
6. The likely impact of the technologies if implemented in the workflow

Key finding: The results of the retrospective study using the AI mammogram reader (Mia™), when used to model the impact of the technology on resourcing showed that using AI as a second reader could reduce the time required from human readers (radiologists and reporting radiographers) by 42%.

It was not possible to model the likely impact of the service optimisation tool during the course of the project as this was not developed and tested until the last 2 months of the project and then only on a small scale.

The likely impact of the AI second reader tool was modelled by comparing the usual care model (two human readers) with the proposed model (one human reader and one AI reader) [Figure 10].

Figure 10: Usual care pathway versus new AI second reader pathway



Analysing the information received on activity and service costs from Nottingham University Hospitals NHS Trust and United Lincolnshire Hospitals NHS Trust, assumptions and predicted effects were identified that could be supported by evidence, and what the material impact would be of deploying the AI second reader tool within the breast screening programme. As no peer-reviewed clinical evidence validating downstream benefits gathered as part of prospective trials was available at this stage, the analysis focused on what immediate and direct resource implications were revealed during the initial retrospective study phase covering three sites[3]. Since there are currently no agreed prices

[3] Kheiron Medical Technologies added a third site to the retrospective study in September 2019 (not an EMRAD Consortium member and in a different English region) which uses different mammography equipment to add to the EMRAD test site data collection. This site was not included in the NHS Test

for Mia™, nor any published market pricing, it was not possible to compute the cost of deployment in a real-world setting.

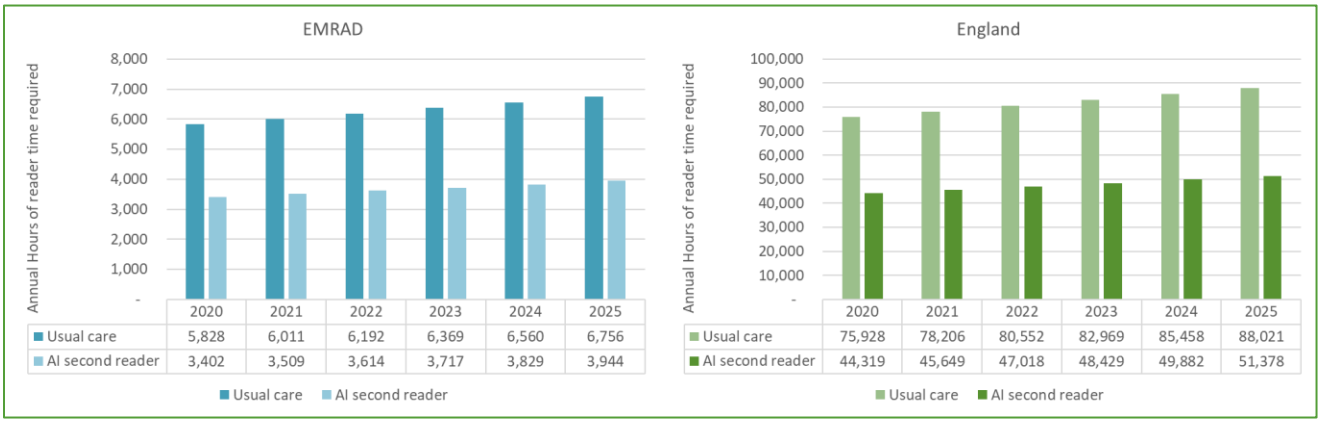
The most material and tangible benefit identified, around which there was sufficient supporting evidence was around reducing second reader time.

Based on historic trends in activity and productivity, the effect of introducing an AI second reader was modelled using the arbitration levels which were identified in the retrospective study (18.2%). This gave an estimate of human screening hours required for each care model, usual versus AI. While the AI eliminated second reader hours, it also increased arbitration (i.e. 'third' reader hours) rates, thereby resulting in.

As trials progress and the technology and processes benefit from learning and refining, it is hoped that the arbitration rates will decrease, and effective capacity release would increase. Thus, the pressure on existing resource (i.e. radiographers) would in turn decrease, either allowing more screening with the same level of staff, or a lower staffing to meet existing needs. These variables will need to be considered as part of future health economic analysis during prospective trials and real-world implementation.

The potential effect of adopting an AI second reader at scale across all seven EMRAD trusts and all English trusts that offer breast screening services was extrapolated. Activity data for the last three years was directly obtained from EMRAD trusts and cross referenced this with KC62 returns to NHS Digital[4]. The KC62 returns were used to establish trends across England. The graphs below indicate the potential effect of using AI as a second reader [Figure 11].

Figure 11: Potential effect of adoption of AI second reader at scale across EMRAD and England.



This analysis is limited in its scope and has limited meaning without information on change to cost for provider trust and evidence from prospective trials on workflow and downstream effects. It does give some indication of the potential effect on workforce utilisation and the potential for an AI second reader to help resolve the acute workforce shortage in radiology.

Bed project. They used the data gathered from the retrospective studies in all three sites to train and validate the Mia™ algorithm and presented their findings at RSNA 2020.

4 https://datadictionary.nhs.uk/data_sets/central_return_data_sets/nhs_breast_screening_programme_central_return_data_set__kc62_.html

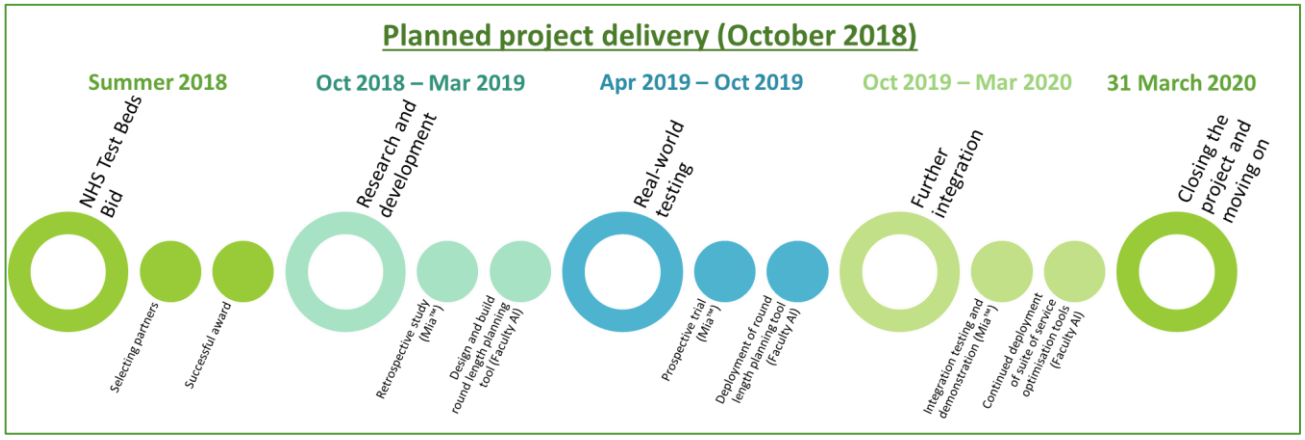
Project outcomes

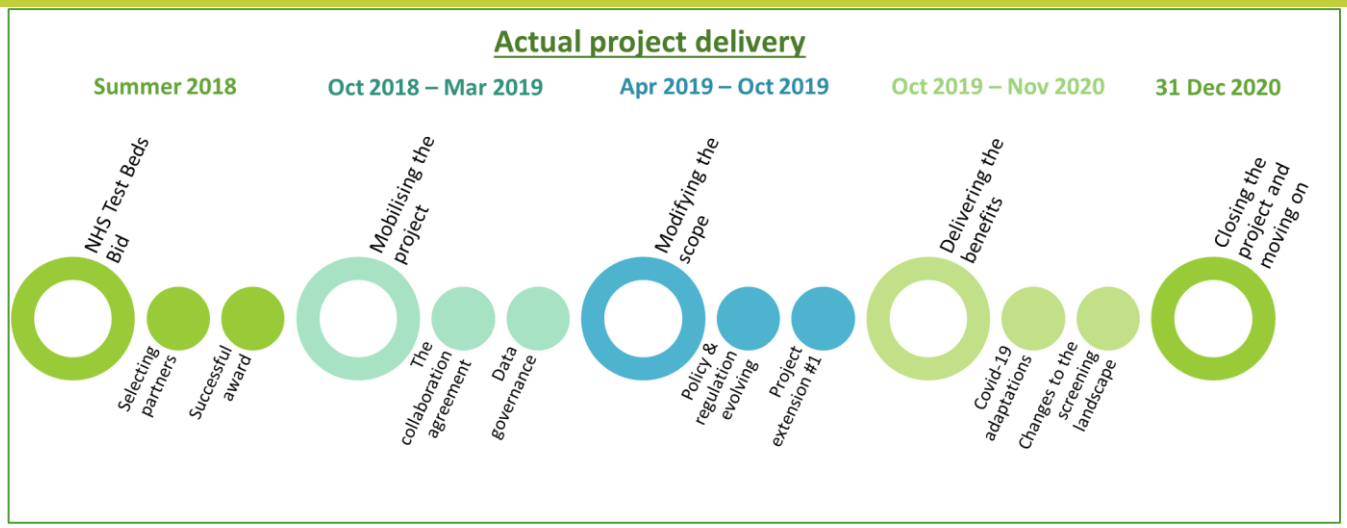
Over that time the project has delivered the following outputs:

Change Domain	Output
Technology:	• Mia™ retrospective study comprising completion of training, validation and testing; • A synthetic data set based on NBSS data to train and validate operational machine learning tools; • Round-length planning tool developed and tested.
Organisational readiness:	• Information governance blueprint for development of AI tools in an NHS context; • A project team with the skills and experience to test AI radiology products in the NHS environment; • Commercial / NHS partnerships for future development, deployment and uptake.
Value proposition:	• Outline business case for real-world deployment; • Financial and budget impact baseline model;
Adopters:	• Change in staff and public attitudes to the use of AI in breast screening with greater awareness across clinical staff groups in participating sites; • A baseline understanding of women in the wider population's attitudes to the use of AI in the breast cancer screening programme in England.
Wider context:	• Contribution to the emerging regulation of and policy context for AI in health through collaborative work with NHS England and Improvement, CQC and NHSX.

When this project is reviewed in terms of progress against the plans originally set out in December 2018, there is a significant difference between what was originally planned and what has been delivered summarised in the Figure 12.

Figure 12: Planned versus actual project delivery





Explaining the progress

Complexity and novelty were two reasons for the slower than originally planned rate of progress that the project decision-makers pointed to. The dimensions of complexity that have been evident in this project are structural, socio-political and emergent (Maylor, 2013).

1. Structural

Delays in securing project **funding** and in recruiting some of the resource to support project implementation as well as changes in project **scope** in response to unanticipated information governance hurdles and requirements for research approval led to delays in the project delivery overall and in some cases abandonment of planned workstreams (economic evaluation and clinical deployment of Mia™) or postponement (service optimisation tool development, training and testing). Having insufficient **information governance** expertise as part of the core team from the point of writing the proposal for Test bed funding is a key lesson learnt by the programme team. Information governance was so central to the project's progress that early involvement of IG specialists in planning may have enabled more proactive mitigation strategies to be put in place.

2. Socio-political

Cultural differences between technology start-up commercial partners and NHS trusts. Both technology start-ups had limited experience of working in the NHS either in a research or delivery capacity. This became evident early on during the discussions around information governance and data sharing. Commercialisation and scaling up highlighted these differences again as technology companies worked with the NHS to develop an evidenced business case that would gain support from Trust boards. As the business case development progressed, the different **expectations of outcomes** between Kheiron and EMRAD became obvious. EMRAD were keen to move quickly to adoption, Kheiron were more cautious, citing the need to deliver clinical trial data first. The **policy context** as well as political and social priorities were moving rapidly over the course of the project and heavily influenced the progress of the project. The project itself influenced policy and regulatory developments as evidenced by the number of times it is referenced as a case study in policy documents (Commons, 2018) (NHSX, 2019). The promotion of AI technology in healthcare in the UK is driven by **political**

commitment directly from the Secretary of State[5] and this high profile alongside the **multiple stakeholders** with an interest in the outcome of this project (NHSX, PHE, NHS England & Improvement, Office for Life Sciences, Innovate UK, CQC, ICO and NHS EMRAD Trusts), means that the project has had to meet a range of interests which have not always been aligned.

3. Emergent

The **Covid-19** pandemic had an effect on the project that could not have been anticipated and had to be adapted to as the risk emerged and was realised.

Attributing outcomes

It is challenging to attribute outcomes in complex projects such as this one (Bovaird, 2014). Add to this the novelty of a project that involves testing and validating new technology that is not yet regulated fully or commercially available, and the traditional approach to evaluating a programme theory of change becomes even more challenging.

The process of co-producing the programme theory of change and informing its development as the project progressed has enabled us to draw out some emerging impact pathways from the data.

The process of conducting the HRA approved research using Mia™ has had some limited impact on the confidence with which the clinical workforce in test sites perceive the AI tool. However, it is worth noting that similar patterns of increased confidence were noted at the control sites which may be indicative the increased profile of the use of AI in healthcare and breast screening specifically during the period of the project. The process of conducting the discovery work with non-clinical staff for the service optimisation tool did have a small effect on levels of engagement with staff and positive perceptions of the possible impact of and AI tool but this was not sustained through the delayed design and development of the tool. Overall, there were no significant differences across test and control sites that can be attributed to the objective of the test bed.

None of the predicted impacts in relation to numbers of women screened, workforce productivity or experience of care have been evidenced in the real world at this stage. The most likely immediate impact of workforce productivity has been modelled as part of the budget impact analysis but remains to be tested in real-world deployment.

[5] https://www.computerweekly.com/news/252488719/Matt-Hancock-announces-50m-for-healthcare-AI-projects

Recommendations

Based on the findings of this evaluation and what is known from research and evaluation in this area to date, we have highlighted the following recommendations that are specific to the UK context but may be generalisable to other contexts:

For policy makers and regulators

1. Continue to evolve the system of regulation in collaboration with interested groups shifting the focus from AI technology firms to healthcare professionals and the wider public including protected groups as adopters of the technology.
2. Continue to consider the role of data governance and ethics in the application of AI with a consideration of the impact on power relationships in the context of person-centred integrated care starting with focusing on the role of informed consent involving the public in the design and monitoring of these approaches beyond user testing.

For the national breast screening programme

3. Set out the expectations for the evidence threshold to be generated as part of future retrospective and prospective clinical trials of AI as a second reader of mammograms undertaken in the UK population.
4. Clarify the requirements and priorities for wider socio-political research on the impact of implementing AI technology in the breast screening programme.

For breast screening units and the NHS trusts that run them

5. If considering adopting AI as part of the clinical or non-clinical workflow, understand the level of readiness of your workforce, workflow and organisation to take up the new technology and ensure that the appropriate information governance and change management support is in place from the beginning to deliver the change.
6. Apply the principles set out in NHSX's *A Buyers Guide to AI in Healthcare*[6] to the procurement and implementation of service optimisation AI tools for use in the breast screening programme.
7. Share learning on the developments in AI across the breast screening unit team (clinical and non-clinical staff) and open a forum as part of team professional development that discusses critically developments in technology including AI in breast screening.

[6] https://www.nhsx.nhs.uk/ai-lab/explore-all-resources/adopt-ai/a-buyers-guide-to-ai-in-health-and-care/

For radiologists, radiographers and other clinical staff

8. Provide support and incentives for staff to learn from (and if possible engage in) research on AI in breast imaging as part of the CPD requirements in your work place.
9. Consider the likely effects of adopting AI as a second reader in the clinical pathway as part of a multi-disciplinary team in terms of professional accreditation and ongoing development, productivity, simplifying working practice and improving the experience of care.

For women of and under screening age and society more widely

10. Ask for information about the results of clinical and real-world research on the effectiveness, safety and ethics of AI in breast screening and other healthcare applications in ways that are clear and understandable to the layperson.

Dr Niamh Lennox-Chhugani
Research Director, TaoHealth Research & Implementation
nlennoxchhugani@taohealth.co.uk

Sanjeev Chhugani
Financial analyst, TaoHealth Research & Implementation
schhugani@taohealth.co.uk

For more information

T: +44 (0)7983 458 733
www.taohealth.co.uk
http://twitter.com/taohealth2