BMJ TAG

Ethical implications of the use of AI-based technologies for medical image classification systems in screening

Protocol

September 2024

1 Title of the project

Ethical implications of the use of AI-based technologies for medical image classification systems in screening.

2 Name of TAR team and project 'lead'

BMJ Technology Assessment Group (BMJ-TAG)

Lead: Steve Edwards

Director of Health Technology Assessment

BMA House, Tavistock Square

London, WC1H 9JP

Telephone: +44 (0)20 3655 5203

Email: sedwards@bmj.com

3 Plain English Summary

Artificial intelligence (AI) is a technology that enables computers and machines to mimic human abilities like learning, understanding, making decisions, and acting independently.[1] AI can identify objects, understand and respond to human language, and learn from new information and experiences. AI can make recommendations and sometimes act on its own, reducing the need for human intervention.

AI has many potential uses, including in public health. For example, AI can help screen for medical conditions like breast cancer and eye problems in people with diabetes. The use of AI could improve medical imaging and diagnosis.

However, it is important to consider the ethical issues raised by the use of AI in screening. This project aims to identify these ethical concerns, particularly in the context of screening for breast cancer and diabetic eye disease. If we don't find sufficient information in these areas, we will broaden our approach to look at screening for all health conditions.

To do this, we will review existing studies that include information on the ethics of using AI in screening. The findings will be assessed to identify common ethical issues, which will then be compared with the ethical framework developed by the UK National Screening Committee (UK NSC). We will also identify any gaps in the current evidence and make recommendations for further research.

4 Decision problem

4.1 Purpose

AI algorithms are built using different approaches and techniques. Machine learning, a subset of AI has been a popular approach in current AI healthcare applications as it allows computational systems to learn from data and improve their performance without being explicitly programmed.[2] Deep learning, a subset of machine learning, employs artificial neural networks with multiple layers to identify patterns in very large datasets.[2] In recent years, there has been a rapid development of AI across different fields, including healthcare. Currently, AI makes use of novel machine learning techniques such as deep learning, which allow algorithms to independently classify data. Through exposure to data, these algorithms can develop the ability to recognise patterns in the data and are not programmed to pay attention to specific attributes or variables.[3] For example, a goal is set, e.g. to distinguish 'cancer' from 'no cancer' on mammographic images, and the algorithm is exposed to a large quantity of training data which may or may not be heterogenous, e.g. image data only or image data and clinical outcome data. Overtime, independent of human instruction, the algorithm 'learns' to identify relevant attributes from the data to achieve the set goal.[3] The increasingly enhanced performance of AI has been responsible to a great extent for recent advances in healthcare applications of AI but due to its capacity for unsupervised learning, the results produced can be difficult to interpret and their operation is much less transparent.

Breast cancer is the most common type of cancer in women in the UK, accounting for 15% of all new cancer cases (2017–2019) and is the 4th most common cause of cancer death in the UK, accounting for 7% of all cancer deaths.[4] Early detection of breast cancers through screening is shown to reduce overall breast cancer mortality.[5] For example, the NHS Breast Screening Programme (NHSBSP), that invites all UK-based women aged 50–70 to attend for screening every three years, is estimated to save a woman's life for every 400 screened, reducing breast cancer deaths by 25%.[5] Such screening programmes represent opportunities for the application of AI. Mammograms, which would constitute the AI algorithm's input data, have a standardised format, while the initial decision a radiologist makes, which would correspond to the output that the algorithm must decide, is a "yes" or "no" decision: does this person need to undergo further assessment?[5]

Screening programmes such as the NHSBSP, generate a large amount of data, which presents an ideal situation to train and test AI algorithms based on data sets of medical images. Providing effective screening for the UK requires significant time, costs and resources, and represents

opportunities for AI-based technologies to make substantial efficiencies. In addition, breast cancer detection is a challenging task as breast tissue has features that are confounding to the human eye in terms of what may or may not look like a tumour, requiring radiologists or radiographers with significant experience to interpret images.[5] In UK clinical practice, to achieve maximal accuracy, every mammogram is assessed by two expert radiologists or specialist trained clinical staff. Breast screening services undertake arbitration to make definitive decisions with each service using its own local protocol, for example using a third mammography reader or a small panel of readers. Usually, arbitration takes place in cases where there is disagreement between the first two readers, they both report an abnormality or when women have reported a symptom during their screening appointment.[6] Within this framework, there is the potential for an automated system such as AI to reduce costs and save processing time by replacing human radiologists.[5]

Diabetes is estimated to affect more than 5.6 million people in the UK.[7, 8] One of the eye conditions associated with diabetes is called diabetic retinopathy; it can lead to sight loss if left untreated and is one of the leading causes of blindness in the working-age population in the UK.[9] Eye screening tests have the potential to detect eye problems before they start impacting sight. In the UK, all patients with diabetes aged 12 years and over are invited to attend the UK Diabetic Eye Screening Programme (DESP). This occurs once every year in England, Wales and Northern Ireland, and once every two years following two consecutive negative screens in Scotland. During the visit, images of each eye are taken by healthcare professionals which are then studies by appropriately trained people called level 1 graders. Subsequently, all images of patients with suspected diabetic retinopathy are sent for further assessment by level 2 graders, with level 3 graders required to make a final decision if there is disagreement between level 1 and level 2 graders. With the number of people with diabetes increasing annually leading to a greater need for trained health professionals, diabetic eye screening presents an opportunity for the application of AI technology to read images, with potential savings made similar to breast cancer screening.

In recent years, there has been a significant development in medical AI, particularly with respect to breast cancer detection and screening.[3] For example, the Digital Mammography DREAM Challenge, aimed to generate algorithms that can reduce false positives without impeding cancer detection.[10] Algorithms have reported 80.3–80.4% accuracy, while algorithms that can fully match the accuracy of expert radiologists are being developed.[3] Moreover, Google Deepmind Health, NHS Trusts, Cancer Research UK and universities have been developing machine learning technologies for mammogram reading.[5] Furthermore, AI systems using digital fundus photography instruments have been

developed for diabetic retinopathy screening. For example, the US Food and Drug Administration (FDA) approved, fully autonomous, EyeArt system (Eyenuk, Inc) has been shown to detect eyes with 'more than mild' diabetic retinopathy with 96% sensitivity and 88% specificity.[11]

Overall, AI has the potential to profoundly change the screening and diagnosis of medical conditions including breast cancer and diabetic retinopathy, speeding up processing times, reducing medical costs and potentially eradicating human errors.[5] AI may become widely implemented and an integral part of medical-image classification and screening for medical conditions. There has been a wealth of high-level governmental, professional and industry statements on the Ethical, Legal and Social Implications (ELSI) of AI, conveying both excitement about the potential benefits of AI but also highlighting concerns about the potential risks and harms.[3] However, currently, there are no ethical frameworks applicable to developing and using AI specifically in a screening context.

The purpose of this research is to identify the ethical issues related to the application of AI-based technologies for medical image classification in screening to inform stakeholders about the advancement of ethically responsible innovation in AI.

4.2 Intervention

The intervention of interest for this review is any AI-based technology applied to image classification and relevant in a screening context. The primary focus of the review is AI in medical image classification relevant to screening interventions for breast cancer and diabetic retinopathy with the scope broadened to the implementation of AI-based technology in screening programmes for any medical condition if insufficient evidence on the aforementioned target conditions is identified.

For this systematic literature review (SLR), the researchers will only search and include papers where the application of AI is medical imaging and include any ethical issues or principles from these papers that would generate evidence transferable or relevant to medical imaging in screening. Papers relevant to any healthcare context will be searched for, to allow the researchers to draw evidence on the application of AI in screening across medical conditions if a wealth of evidence relevant to breast cancer and diabetic retinopathy is not identified.

4.3 Place of the intervention in the treatment pathway(s)

Primary studies evaluating the ethical impact of AI-based technology applied in medical screening programs or similar settings will be sought. The ethical implications of AI-based technology across

the screening pathway of medical conditions, including the identification of a potential abnormality to diagnosis, differential diagnosis, grading/staging, will be examined. Different ethical considerations may be associated with different stages of the screening pathway, and this is to be determined by the evidence identified.

4.4 Relevant comparators

N/A; There are no comparators listed in the commissioning brief. As the application of AI in medical image classification in screening will not be compared to any other interventions within the context of this review, there are no relevant comparators.

4.5 Population and relevant subgroups

The population relevant to this review is the 'screening population', i.e. people undergoing screening for any health condition, with the primary focus being on people undergoing screening to detect breast cancer and diabetic retinopathy. The views of health professionals, care providers, patients, the users of screening programs and the general public on the AI implementation in population screening is also to be sought.

No population subgroups are relevant for this review and similar views emerging from health professionals, patients and the general public will be synthesised where possible.

Themes emerging for screening for different health conditions or that appear to be relevant to different aspects of screening or diagnosis, for example depending on whether AI is implemented in screening to identify potential abnormalities and the interpretation of a screening test or a diagnostic or confirmatory test that comes later in the screening pathway, will be explored separately where the evidence identified permits, depending on the type and wealth of the information identified and the themes emerging.

4.6 Outcomes to be addressed

The outcomes to be addressed in this review are:

Ethical issues

- To identify ethical issues specific to traditional machine learning and deep learning techniques, more specifically issues associated with non-continual learning and continual learning, respectively;

- Differentiation of ethical issues may be needed by use case, for example fully autonomous AI image interpretation where there is no human involvement versus different levels of human interaction or AI interpreted images double checked by a human.

Themes will be derived from the evidence identified for this review and not pre-specified. Examples of ethical issues may include but not be limited to safety and transparency, AI false negative and false positive cases, automation bias, data privacy, liability. Relevant quantitative data emerging from included studies will be extracted and presented alongside themes identified from qualitative analysis for illustrative purposes. Themes identified will be compared with the UK NSC ethics framework and mapped to the digital health technology product life cycle.

5 Report methods for synthesis of evidence of clinical effectiveness

A SLR of the evidence on the use of AI-based image classification systems in screening will be performed following the PRISMA statement.[12] A flow diagram illustrating the number of records identified, included and excluded at each stage of the SLR will be presented according to the PRISMA reporting guidelines.[12]

5.1 Search strategy

The researchers will replicate and update the search performed for Question 4: 'What are the social and ethical implications of implementing AI-based tools in screening programmes and would it be acceptable to health professionals and the public?' of the UK NSC 2021 evidence map from June 2020 onwards.[9] Papers identified in the 2021 UK NSC evidence map along with newly identified papers from the updated search will be assessed for relevance to the current review against the inclusion and exclusion criteria of this protocol. In line with the evidence map search, the researchers will perform systematic searches of Ovid MEDLINE, Embase, PsycINFO, CINAHL, to identify primary qualitative studies relevant to the evaluation of the ethical implications of AI-based technologies in medical image classification systems in screening for medical conditions. Review and opinion papers (summarising the views and opinions of the authors, clinicians or professional bodies or reviewing literature non-systematically without undertaking clinical research) will not be included as the current research will aim to determine ethical issues that the use of AI-based technologies in screening interventions may raise emerging from the qualitative views and opinions of health professionals, patients and the general public.

The search terms will include combinations of free text and subject headings grouped into the following categories:

- Intervention: diagnosis, early diagnosis, computer assisted, diagnostic test, screening, imaging, artificial intelligence, automation, machine learning, deep learning, neural network;
- Outcomes: attitudes, perception, accept, barriers, appropriate, experience, ethic, social;
- Study design: qualitative, interview, survey, questionnaire.

The searches conducted for the UK NSC 2021 evidence map[9] will be updated. Thus, searches will be limited to the period from the date of the previous search (30 June 2020) to the date of the current search. To update the UK NSC 2021 search, update terms from the University of South Australia guidelines on updating a search will be used for applying date restrictions.[13] Some additional search terms will be included to broaden the original search. All search terms are listed in Appendix 1. In addition, included studies derived from the original searches will be reviewed for inclusion in the current review. The research group's proposed search strategy for each database is presented in Appendix 1.

Table 1. Inclusion and exclusion criteria of the SLR

Factor	Inclusion and exclusion criteria
Population	Health professionals, providers and users (clinicians and patients) of screening programmes and the general public. Screening population (for breast cancer, diabetic retinopathy being the primary focus during inclusion/exclusion but searching for papers relevant to all target conditions)
Intervention	AI-based technology applied to medical image classification for screening of any medical condition. Full-text papers relevant to any medical condition will be acquired. Papers relevant to screening of breast cancer and diabetic retinopathy will be prioritised for inclusion in the review, with papers relevant to other medical conditions only included if no sufficient evidence emerges on the abovementioned target conditions.
Setting	Screening of medical conditions in UK clinical practice. Papers that appear to be irrelevant to UK practice, such as papers conducted in deprived countries will be considered for exclusion. The exclusion of such papers will be conducted on a case-by-case basis after discussion with the research group's clinical experts.
Comparators	n/a
Outcomes	Themes on ethical issues surrounding the use of AI in screening of medical conditions will be derived from

	the evidence identified for this review and will not be pre-specified. Differentiation of ethical issues may be needed by use case, for example autonomous AI versus different levels of human interaction or checking of AI interpreted images. If few qualitative studies are identified, quantitative data from questionnaires/surveys representing relevant information on the views of health professionals, providers, users/patients of screening programs on ethical issues associated with the use of AI in medical screening will also be explored to help substantiate the themes identified from the limited qualitative data.
Type of studies to be included	Qualitative studies such as interview and focus group studies (including studies using thematic analysis, grounded theory or other appropriate qualitative approaches); if few qualitative studies are identified then relevant quantitative data from questionnaires/surveys will also be considered. Any relevant SLRs identified will be assessed as full texts and the individual studies included in the SLR will be cross checked to ensure they have been picked up by the systematic search and separately assessed for inclusion/exclusion in the current review.
Other exclusion criteria	Non-English language studies Conference abstracts Quantitative studies with no relevant data Studies published before 2000 Non-systematic narrative reviews and opinion papers where no analysis has taken place SLRs will be excluded but after their inclusion lists have been checked for inclusion/exclusion in the current review.

5.2 Review process

The following review process will be followed:

1) After removing duplicates, the records identified from the search will be imported to EndNote. Each paper abstract will be reviewed against the inclusion/exclusion criteria by a single reviewer, with a second independent reviewer providing input in cases of uncertainty and validating 20% of the first reviewer's decisions. Any disagreements will be resolved by discussion with consensus reached consulting a third reviewer if necessary to resolve any outstanding conflicts.
2) Full-text papers of the above records selected for inclusion will be acquired.
3) Each full-text paper will be reviewed against the inclusion/exclusion criteria of the protocol by one reviewer who will determine whether the paper is relevant to the review question. A

second reviewer will provide input in cases of uncertainty and validate 20% of the first reviewer's decisions. Any disagreements will be resolved by discussion until consensus is reached with the involvement of a third reviewer to resolve any outstanding conflicts if necessary.

4) Throughout the reviewing process, the 20% rule will be implemented as a first step in validating all the review process steps, as long as there is coherence in the decision of both reviewers. If there are inconsistencies in the reviewers' decisions, discussion between reviewers will take place, with the involvement of a third reviewer if necessary. If as a result of these discussions, the view of the first reviewer is changed, then 100% of the work will be validated as a second step. In addition, if there are no time constrains and the amount of evidence identified permits, 100% of the work will be validated even if there are no inconsistencies between reviewers in the first step of validation.

5.3 Data extraction strategy

Data will be extracted independently by one reviewer using a standardised data extraction form (see Appendix 2) for each study separately and checked by another reviewer. Discrepancies will be resolved by discussion, with involvement of a third reviewer when necessary. The Draft data extraction form is provided in Appendix 2. Extracted data will be validated by a second reviewer and discrepancies will be resolved by discussion, with involvement of a third reviewer when necessary.

Qualitative data/information on ethics surrounding the use of AI emerging from each paper selected for inclusion will be extracted and summarised as themes for each paper in its corresponding data extraction form. Prior to the thematic analysis described in Section 5.5 below, as part of the data extraction process, the reviewer will first identify information/qualitative data that are relevant to the topic of this research on an individual study level and summarise them as themes emerging from each study in the corresponding data extraction form. Within this framework the reviewer will not be constrained by potential themes already identified by the authors of each paper (whose aim may have differed to that of the present review) but can utilise information emerging from across the paper to compose new themes that are relevant to the focus of the current review. These will be validated by a second reviewer with discrepancies resolved by discussion and the involvement of a third reviewer if necessary, as described above. Themes extracted across papers will then be reviewed and combined/further synthesised into overarching themes where possible as described in the thematic analysis section below.

Critical Appraisal Skills Programme (CASP) checklist (see details in 5.4.1 below) will be used to assess included qualitative studies and a statement about limitations and applicability will be included in the data extraction form for each study.[14]

If few qualitative studies are identified, quantitative data from surveys reporting relevant information/views of patients or clinicians on the ethical implications of implementing AI-based technologies in screening will be reported narratively for illustrative purposes and presented alongside thematic analysis to help substantiate the findings from the limited qualitative data. Risk of bias will be assessed to ascertain outcome quality as specified in Section 5.4.

The following strategy will be followed for the inclusion of papers in the review:

1. Qualitative papers relevant to the use of AI in screening in the context of breast cancer and diabetic retinopathy will be prioritised for full-text assessment and inclusion;
2. If only a few relevant papers are identified relating to breast cancer and diabetic retinopathy, qualitative papers relevant to the use of AI in screening of any medical condition will be assessed;
3. If few qualitative papers are identified irrespective of medical condition, studies involving surveys/questionnaires will be assessed.

5.4 Quality assessment strategy: per theme/ethical issue using GRADE CERQual and CASP checklist for each study

After quality assessment of individual studies has been performed, themes from the included qualitative studies will be evaluated and presented using the 'Confidence in the Evidence from Reviews of Qualitative Research' (CERQual) Approach developed by the GRADE-CERQual Project Group, a subgroup of the GRADE Working Group.[15] The CERQual Approach assesses the extent to which a review finding/theme is a reasonable representation of the phenomenon of interest (the focus of the review question). Each theme will be assessed for each of the 4 quality elements listed and defined in Table 2 below. GRADE CerQual will be used to assess the certainty of the evidence for each finding/theme.

Table 2. Descriptions of quality elements in GRADE-CERQual for qualitative studies

Quality element	Description
Methodological limitations	The extent of problems in the design or conduct of the included studies that could decrease the confidence that the review finding is a reasonable representation of the phenomenon of interest. Assessed at the study level using the CASP checklist.

Coherence	The extent to how clear and cogent the fit is between the data from the primary studies and the review finding.
Relevance	The extent to which the body of evidence from the included studies is applicable to the context (study population, phenomenon of interest, setting) specified in the protocol.
Adequacy	The degree of the confidence that the review finding is being supported by sufficient data. This is an overall determination of the richness (depth of analysis) and quantity of the evidence supporting a review finding or theme.

5.4.1 Methodological limitations

Each theme will have its methodological limitations assessed within each study first using the CASP checklist.[14] Based on the degree of methodological limitations, studies will be evaluated as having **minor**, **moderate** or **severe** limitations. A summary of the domains and questions covered is given below.

Table 3. Description of limitations assessed in the CASP checklist for qualitative studies

Domain	Aspects considered
Are the results valid?	• Was there a clear statement of the aims of the research? • Is qualitative methodology appropriate? • Was the research design appropriate to address the aims of the research? • Was the recruitment strategy appropriate to the aims of the research? • Was the data collected in a way that addressed the research issue? • Has the relationship between researcher and participants been adequately considered?
What are the results?	Have ethical issues been taken into consideration? Was the data analysis sufficiently rigorous? Is there a clear statement of findings?
Will the results help locally?	How valuable is the research?

For surveys reporting relevant quantitative data, methodological limitations will be assessed using the CEBMa checklist[16] listed in the NICE methods manual Appendix H.[17] The domains and questions covered can be found in Appendix 4. The overall assessment of the methodological limitations of the evidence will be based on the primary studies contributing to the theme. The relative contribution of each study to the overall review finding and the type of methodological limitation(s) identified will be taken into account when giving an overall rating of concerns for this component.

5.4.2 Coherence

Coherence is the extent to which the reviewer is able to identify a clear pattern across the studies included in the review, and if there is variation present (contrasting or disconfirming data) whether this variation is explained by the contributing study authors. For example, if a review finding in 1 study does not support the main finding and there is no plausible explanation for this variation, or if there is ambiguity in the descriptions in the primary data, then the confidence that the main finding reasonably reflects the phenomenon of interest is decreased.

5.4.3 Relevance

Relevance is the extent to which the body of evidence from the included studies is applicable to the context (study population, phenomenon of interest, setting) specified in the protocol. As such, relevance is dependent on the individual review and will be discussed with clinical experts.

5.4.4 Adequacy

The judgement of adequacy is based on the confidence of the finding being supported by sufficient data. This is an overall determination of the richness (and quantity of the evidence supporting a review finding or theme. Rich data provide sufficient detail to gain an understanding of the theme or review finding, whereas thin data do not provide enough detail for an adequate understanding. Quantity of data is the second pillar of the assessment of adequacy. For review findings that are only supported by 1 study or data from only a small number of participants, the confidence that the review finding reasonably represents the phenomenon of interest might be decreased because there is less confidence that studies undertaken in other settings or participants would have reported similar findings. As with richness of data, quantity of data is review dependent. Based on the overall judgement of adequacy, a rating of no concerns, minor concerns, or substantial concerns about adequacy will be given.

5.4.5 Overall judgment of the level of confidence for a review finding

GRADE-CERQual will be used to assess the body of evidence as a whole through a confidence rating representing the extent to which a review finding is a reasonable representation of the phenomenon of interest. For each of the above components, level of concern is categorised as either;

- no or very minor concerns;
- minor concerns;

- moderate concerns; or
- serious concerns.

The concerns from the 4 components (methodological limitations, coherence, relevance and adequacy) will be used in combination to form an overall judgement of confidence in the finding. GRADE-CERQual uses 4 levels of confidence: high, moderate, low and very low confidence. The significance of these overall ratings is explained in Table 4 below. Each review finding starts at a high level of confidence and is downgraded based on the concerns identified in any 1 or more of the 4 components. Quality assessment of qualitative reviews is a subjective judgement by the reviewer based on the concerns that have been noted. An explanation of how such a judgement had been made for each component will be included in the footnotes of the summary of evidence tables where the evidence will be summarised as well as narratively under each overarching review theme.

Table 4. GRADE CERQual levels of confidence for each finding

Level	Desription
High confidence	It is highly likely that the review finding is a reasonable representation of the phenomenon of interest.
Moderate confidence	It is likely that the review finding is a reasonable representation of the phenomenon of interest.
Low confidence	It is possible that the review finding is a reasonable representation of the phenomenon of interest.
Very low confidence	It is not clear whether the review finding is a reasonable representation of the phenomenon of interest.

5.5 Methods of analysis/synthesis: Thematic analysis

The synthesis of qualitative data will follow a thematic analysis approach. Thematic analysis is a method used to analyse qualitative data that involves the identification and reporting of patterns in data sets, which are then interpreted for their inherent meaning thus offering insights into the research question.[18] Information will be synthesised into the main report findings/themes as outlined below. Results will be presented in a detailed narrative and in table format and with summary statements of the main findings alongside their GRADE-CERQual confidence rating (see Appendix 3)

Information relevant to the research will be identified from each included paper and narratively summarised into themes for that paper. Themes emerging across studies will then be reviewed and thematic analysis methods will be used to identify common patterns of information which will be further synthesised into broader overarching themes. These will form the main review findings. The evidence will be presented in the form of a narrative summary detailing the evidence from the relevant papers and how this informed the overall theme plus a statement on the level of confidence for that theme. Considerable limitations and issues around relevance will be listed. A summary evidence table with the succinct summary statements for each theme will be produced including the associated quality assessment. If required, relevant quantitative data from surveys will be extracted in a narrative format and included in the qualitative synthesis of themes.

In more detail, papers selected for inclusion in this research will first be read by the reviewers to familiarise with the data. The data will be closely examined and information relevant to the outcome of interest, ethical issues surrounding the use of AI in screening of medical conditions, will be noted and given a code/title, e.g. data confidentiality. The code will be assigned to segments of data within the paper that capture the core message of the code. Identified information across the paper that is relevant to each code will be narratively summarised in the data extraction form for each study under a theme and given a theme title. Quotes from study participants will be extracted and used where appropriate to further illustrate the themes. Once this process is completed for each included study, derived study themes will be reviewed with recurring elements, i.e. common themes across studies and patterns in the qualitative data identified and further synthesised into overarching themes that link the review question, and the data identified.

6 Contribution of the research group:

Steve Edwards	Director of Health Technology Assessment, BMJ-TAG, London. Validation of the work of the research group; will provide feedback on all versions of the protocol and the report. Guarantor of the report.

Melina Vasileiou	Clinical Evidence Analyst, BMJ-TAG, London. She will be the main reviewer on this project and will maintain day-to-day running of the review. She has compiled the study protocol and will carry out the study selection, data extraction and synthesis. She will draft the methods, narratives for included trials, and part of the results and discussion of the final report.
Victoria Wakefield	Senior Clinical Evidence Analyst, BMJ-TAG, London. She will be the second reviewer on this project and will run the SLR; she will validate the selection of studies, data extraction and synthesis by the first reviewer and provide input in cases of uncertainty; she will draft sections of the final report.
Clare Dadswell	Clinical Evidence Manager, BMJ-TAG, London. She will provide feedback to resolve any outstanding conflict between the first two reviewers.

7 Timetable/Milestones

Table 5 Draft milestones table

Milestone	Start date	End date
Project start date	14/08/2024	14/08/2024
Initial meeting with NIHR/UK NSC	14/08/2024	14/08/2024
Draft protocol development	20/08/2024	10/09/2024
Submit draft protocol to UK NSC	10/09/2024	10/09/2024
Comments back from UK NSC on draft protocol	12/09/2024	12/09/2024
Submit final protocol to UK NSC	16/09/2024	16/09/2024
SLR and data analysis	17/09/2024	30/10/2024
Project pause for NICE projects	31/10/2024	06/11/2024
Clinical report writing	07/11/2024	20/01/2025

Project pause for NICE projects	18/11/2024	18/12/2024
Submit draft report to UK NSC	21/01/2025	21/01/2025
Comments from UK NSC on draft report	22/01/2025	28/01/2025
Final report work and internal peer review	22/01/2025	06/02/2025
Final report ready to submit to UK NSC	07/02/2025	07/02/2025

Appendix 1: Draft search strategy

Table 6 Researchers' strategy for MEDLINE via Ovid ALL

#	Searches
1	exp Diagnosis, Computer-Assisted/
2	exp Mass Screening/ or exp early diagnosis/
3	diagnostic test*.ti,ab.
4	(screening or imaging).ti,ab.
5	early diagnosis.ti,ab.
6	or/1-5
7	exp Artificial Intelligence/
8	Automation/
9	exp neural networks, computer/
10	artificial intelligence.ti,ab.
11	(automated adj2 (tool* or technique* or identification or detection or test* or screening)).ti,ab.
12	automation.ti,ab.
13	machine learning.ti,ab.
14	deep learning.ti,ab.
15	neural network*.ti,ab.
16	or/7-15
17	exp Ethics/ or exp Attitude/
18	(attitude* or perception* or acceptab* or barriers or appropriate* or experience* or views).ti,ab.
19	(ethic* or social*).ti,ab.
20	or/17-19
21	((("semi-structured" or semistructured or unstructured or informal or "in-depth" or indepth or "face-to-face" or structured or guide) adj2 (interview* or discussion* or questionnaire*)) or (focus group* or qualitative or ethnograph* or fieldwork or "field work" or "key informant")).tw,kw. or interviews as topic/ or focus groups/ or narration/ or qualitative research/
22	exp ''Surveys and Questionnaires''/
23	(qualitative or interview* or survey* or question* or focus group).ti,ab.

24	or/21-23
25	6 and 16 and 20 and 24
26	limit 25 to dt=20200630-20240917
Database(s): Ovid MEDLINE(R) and Epub Ahead of Print, In-Process, In-Data-Review & Other Non-Indexed Citations, Daily and Versions from June 30, 2020	

Appendix 2: Data extraction form

Study	Author Year (Reference)
Aim	[Aim of the research as stated in the paper]
Population	[As stated in the paper. See example clinical evidence table below for guidance on content] [Characteristics: n= ; male/female; mean age (SD); and other important characteristics as per review context.]
Setting	[As stated in the paper]
Study design	[As stated in the paper. For example: qualitative, mixed methods]

Methods and analysis	[Data collection method and analysis method. Add detail as described in the paper. See example clinical evidence table below] [For example: In-depth semi-structured telephone interview with thematic qualitative analysis. For example: Semi-structured interview and phenomenological (grounded theory) analysis]
Findings	[As many rows as needed to be used to give a clear succinct summary of the main findings (themes) emerging from the paper that are relevant to the current review. Information from the paper can be synthesised by the reviewer under a new theme and themes derived by the paper authors do not need to be extracted as reported in the paper, whose original aim/focus may have been different to that of the current review. See example clinical evidence table below for guidance on content]

Funding	
Limitations and applicability of evidence	[Note any limitations worth highlighting and comments about applicability and directness]

Appendix 3: Example of GRADE CERQual summary table

Table 7. Tabulated summary of evidence with GRADE CERQual confidence rating

Study design and sample size			Quality assessment		
Number of studies contributing to the finding	Design	Findings	Criteria	Rating	Overall assessment of confidence
Title of theme					
	e.g. Semi-structured interviews (1 study); quantitative questionnaire (1 study)	Short summary statement of the narrative synthesis of the theme	Limitations	No/ very minor, Moderate or serious concerns about methodological limitations	HIGH/ MODERATE/ LOW/VERY LOW
			Coherence	No/ very minor, Moderate or serious concerns	

Study design and sample size		Findings	Quality assessment		
Number of studies contributing to the finding	Design		Criteria	Rating	Overall assessment of confidence
				about coherence	
			Relevance	No/ very minor, Moderate or serious concerns about relevance	
			Adequacy	No/ very minor, Moderate or serious concerns about adequacy	

Appendix 4: Critical Appraisal checklist for cross-sectional studies/surveys

Table 8. CEBM checklist for critical appraisal of cross-sectional studies/surveys

Appraisal questions	Yes	Can't tell	No
Did the study address a clearly focused questions/issue?			
Is the research method (study design) appropriate for answering the research question?			
Is the method of selection of the subjects (employees, teams, divisions, organizations) clearly described?			
Could the way the sample was obtained introduce (selection)bias?			
Was the sample of subjects representative with regard to the population to which the findings will be referred?			
Was the sample size based on pre-study considerations of statistical power?			

Was a satisfactory response rate achieved?			
Are the measurements (questionnaires) likely to be valid and reliable?			
Was the statistical significance assessed?			
Are confidence intervals given for the main results?			
Could there be confounding factors that haven't been accounted for?			
Can the results be applied to your organization?			

8 References

1. What is artificial intelligence (AI) ?, 2024. Available from: https://www.ibm.com/topics/artificial-intelligence. Date accessed: 29 Aug 2024.
2. Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare, 2020. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7332220/. Date accessed: 29 Aug 2024.
3. Carter S, Rogers W, Win K, Frazer H, Richards B, Houssami N. The ethical, legal and social implications of using artificial intelligence systems in breast cancer care. *Breast (Edinburgh, Scotland)* 2020; **49**: 25-32.
4. CRUK. Breast cancer statistics, 2024. Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer#heading-Two. Date accessed: 29 Aug 2024.
5. CRUK. Speaking a thousand words – how a cancer image collection is set to improve AI diagnosis, 2021. Available from: https://news.cancerresearchuk.org/2021/09/08/speaking-a-thousand-words-how-a-cancer-image-collection-is-set-to-improve-ai-diagnosis/. Date accessed: 29 Aug 2024.
6. Gov.UK. Breast screening: guidance for image reading, 2024. Available from: https://www.gov.uk/government/publications/breast-screening-guidance-for-image-reading/breast-screening-guidance-for-image-reading#:~:text=Arbitration%20maximises%20potential%20sensitivity%20for,to%20maximise%20specificity%20of%20recall. Date accessed: 16 Sept 2024.
7. NHS. Diabetes, 2023. Available from: https://www.nhs.uk/conditions/diabetes/. Date accessed: 30 Aug 2024.
8. UK D. How many people in the UK have diabetes, 2024. Available from: https://www.diabetes.org.uk/about-us/about-the-charity/our-strategy/statistics. Date accessed: 30 Aug 2024.
9. Committee UNS. Automated grading in the Diabetic Eye Screening Programme 2021. Available from: https://www.gov.uk/government/consultations/automated-grading-in-diabetic-eye-screening-rapid-review-and-evidence-map. Date accessed: 02 Sept 2024.
10. The Digital Mammography DREAM Challenge. Available from: https://www.synapse.org/Synapse:syn4224222/wiki/401743. Date accessed: 29 Aug 2024.
11. Lim J, Regillo C, Sadda S, Ipp E, Bhaskaranand M, Ramachandra C, et al. Artificial Intelligence Detection of Diabetic Retinopathy, 2022. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9636573/. Date accessed: 30 Aug 2024.

12. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021; **372**: n71.
13. UniSA. Updating a search. Available from: https://guides.library.unisa.edu.au/SystematicReviews/UpdateASearch Date accessed: 29 Aug 2024.
14. Programme CAS. CASP Checklist: 10 questions to help you make sense of a Qualitative research 2018. Available from: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://casp-uk.net/checklists/casp-qualitative-studies-checklist-fillable.pdf. Date accessed: 29 Aug 2024.
15. Colvin JC, Garside R, Wainwright M, Munthe-Kaas H, Glenton C, Bohren M. Applying GRADE-CERQual to qualitative evidence synthesis findings—paper 4: how to assess coherence, 2028. Available from: https://implementationscience.biomedcentral.com/articles/10.1186/s13012-017-0691-8#:~:text=GRADE%2DCERQual%20(hereafter%20referred%20to,(4)%20adequacy%20of%20data. Date accessed: 29 Aug 2024.
16. Management CfEB. Critical Appraisal Checklist for Cross-Sectional Study (Survey), 2014. Date accessed: 02 Sept 2024.
17. NICE. Developing NICE guidelines: the manual, 2014. Date accessed: 01 Sept 2024.
18. Naeem M, Ozuem W, Ranfagni S, Howell K. A step-by-step Process of Thematic Analysis to Develop a Conceptual Model in Qualitative Research, 2023. Available from: https://journals.sagepub.com/doi/10.1177/16094069231205789. Date accessed: 29 Aug 2024.