# Supplementary SI

**M**odel for **AS**sessment of **A**rtificial **I**ntelligence (MAS-AI) guideline

## Background and target group

Artificial intelligence (AI) is seen as one of the major disrupting forces in the future healthcare system. However, the best way to assess the value of these new technologies is unclear. Therefore, a multidisciplinary group of experts and patient representatives from Denmark developed a Model for ASsessing the value of AI (MAS-AI) in medical imaging (see Fasterholdt et al. (2022)). The area of medical imaging was chosen because it is a relatively mature area in healthcare with widely adopted data standards developed which provides a good foundation for the development of AI and thus many studies have been published.

The main aim of MAS-AI is to support decision-makers when deciding whether a mature AI application should be implemented into clinical practice or not, i.e. support adoption decisions. Adoption decisions become relevant when reaching the light-gray steps in figure S1 for a particular technology. Hence, it is important to identify at which step on the development ladder the AI application is currently placed (figure S1), i.e., how mature is the AI? Are we still under retrospective development, or have we gained regulatory approval and are ready for deployment and prospective evaluation?

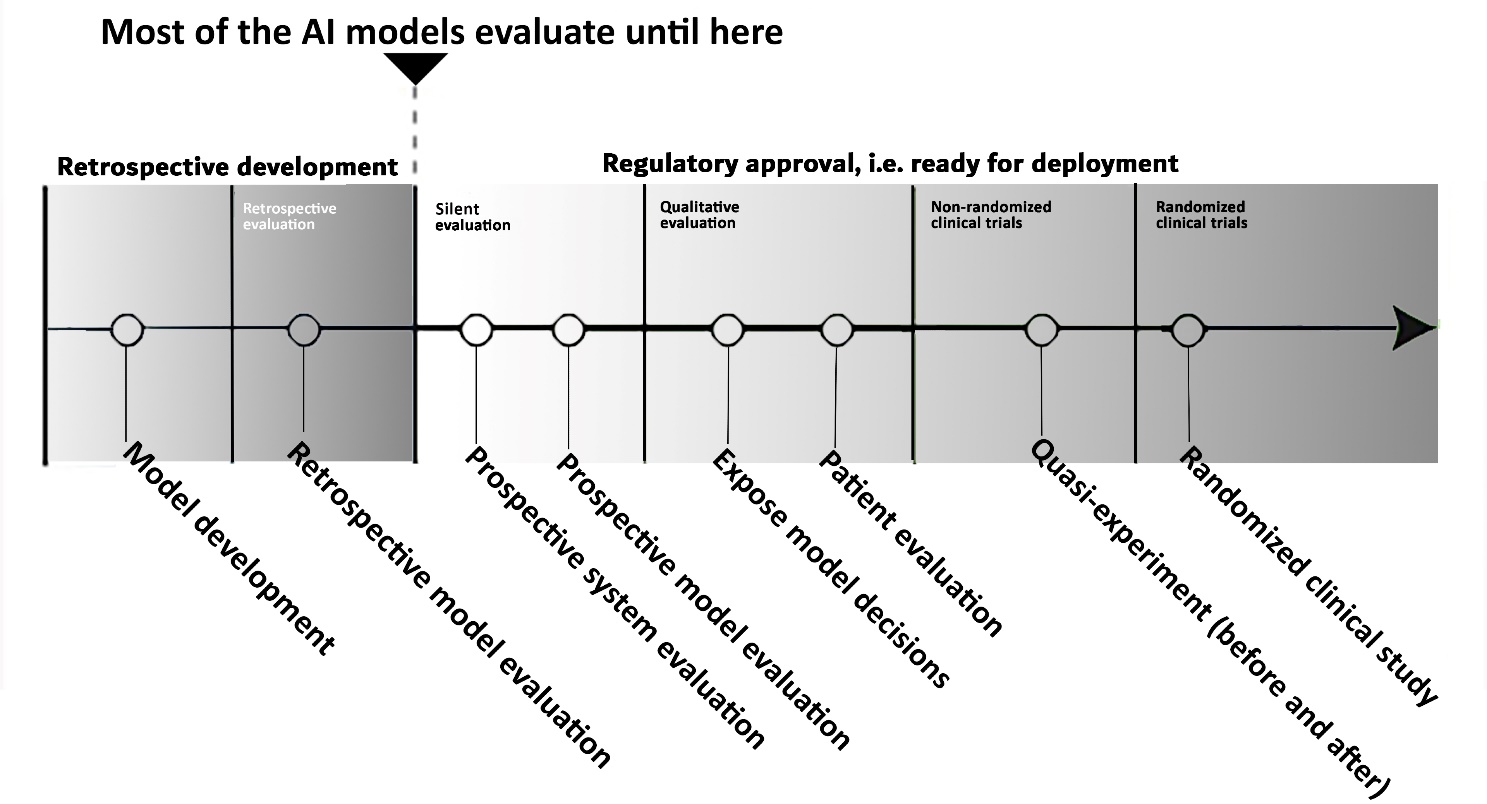


Figure S1. Pipeline for evaluating AI technologies (inspired by an unpublished figure by Simon Meyer Lauritsen)

MAS-AI is primarily an assessment model. The main target group is healthcare decision-makers, e.g. medical directors, heads of hospital departments, local or national treatment councils, procurement organizations, etc. However, developers, researchers and clinicians could also use the MAS-AI to guide the development, data collection or research process. Lastly, the regulatory side may also find parts of MAS-AI helpful, e.g. policymakers from HTA organizations or other regional and national authorities. Thus, MAS-AI covers evaluation in the whole life-span of an AI technology, i.e. all the phases in figure S1.

## Overview of MAS-AI

To accommodate both of the uses mentioned above (adoption or development decisions), the MAS-AI guideline has three parts and figure S2 gives an overview of these parts' content. Two steps cover nine domains, and then there are process factors for a MAS-AI assessment. Note that the order of domains has no particular significance.

Step one contains a description of patients, how the AI model was developed, and initial ethical and legal considerations. Finishing the first four domains in step one is a prerequisite or indication that we are ready to move to step two. In step two, a multidisciplinary assessment of outcomes of the AI application is done for the five remaining domains. The last part is five factors to facilitate a good evaluation process.

Finishing both steps is a full MAS-AI assessment. Finishing only the first step is considered an "early MAS-AI," i.e. an initial assessment in the stage when only limited data are available in a few domains. Hence, step one can be seen as a pre-screening, and if step one turns out positive, the second step may proceed. It is important to underline that MAS-AI is not intended as a "one-size-fits-all"-evaluation model. If the AI application is not very patient-critical, less rigorous evaluation may be appropriate.

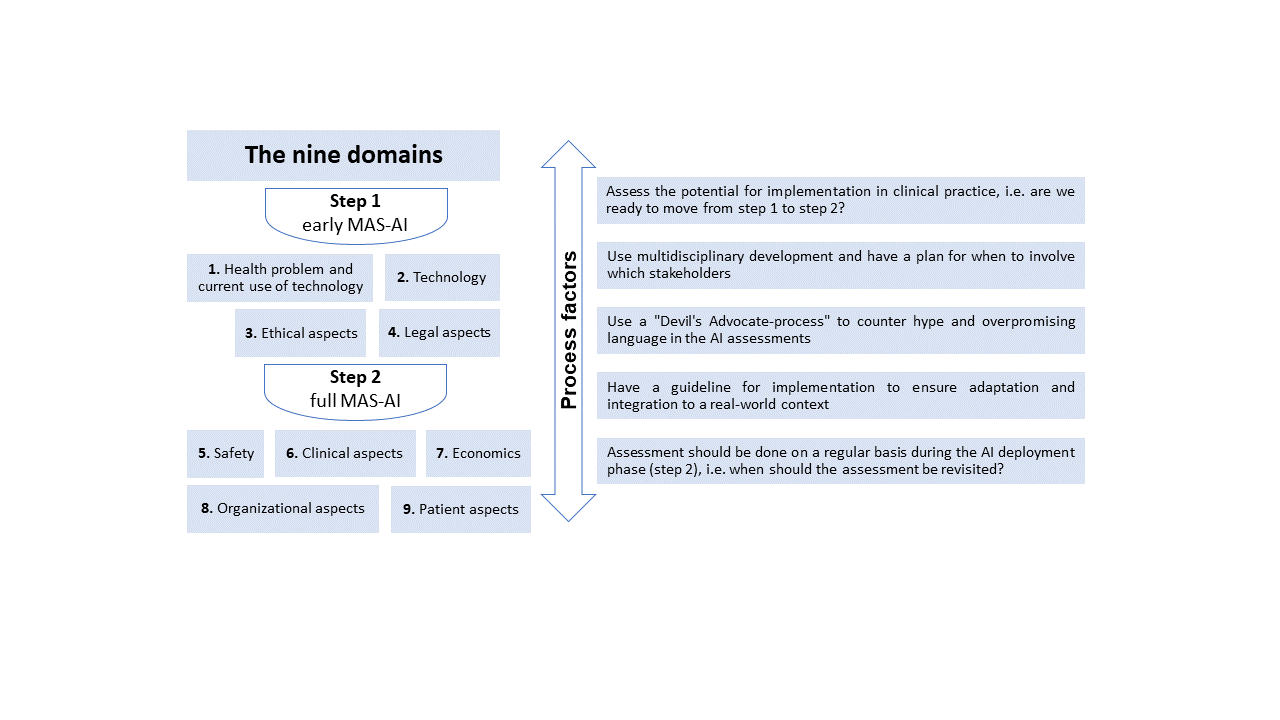


Figure S2. Overview of MAS-AI

The following paragraph presents the content of each of the nine domains. Lastly, the five process factors for a MAS-AI assessment are explained.

## Elaboration of domains

## Domain 1: The health problem and current use of technology

This domain includes a description of the health problem of the patient and a description of the current use of technology and the AI application being assessed. It is important to underline that the description needs to be thorough due to the content of this domain serving as a description of the background for the assessment. Further, the MAS-AI guideline suggests using a separate checklist e.g., "Checklist for Artificial Intelligence in Medical Imaging (CLAIM)", see Mongan et al. (2020). The CLAIM guideline has 42 items, and items 3-6 are relevant for domain 1. Thus, the following information should be provided:

1. Description of the health problem of the patient/health problem and current use of technology and transferability of previous data/results.
2. A description of the application, i.e., intended use and clinical role of the AI approach (CLAIM #3)
   * When in the diagnostic process will AI be used?
   * Indicate the type of AI in the clinic:
     + AI applications as alternative triage
     + Replacement
     + Add-on in clinical workflows
3. With regards to current and previous research, it is important to notice: study objectives (hypotheses), the study design of the model evaluation, and the aim/goal of the study (CLAIM #4,5,6)

## Domain 2: Technology

The technology aspects of the assessment can be split into two areas of concern; 1) issues relating mainly to the development of the model, and 2) issues relating to the deployment of the model into the specific clinical context being evaluated.

The *model concerns* mainly address issues that influence how the model came into being. As such, these issues are general to all model installations across all deployment sites. We need to ensure that the AI model's development, performance, and validation are described, i.e. is the CLAIM guideline filled out? These issues establish the prerequisites for the application.

While both *model* and *application* issues can be thought of as malleable – adjustment of the actual model is less likely to occur post-deployment given the Medical Device Regulation (etc.). Consequently, model aspects of the technology domain can be thought of as a guiding checklist for the assessment committee to position and compare multiple models.

In contrast, the *application* assessment aspects focus on the fit between a given candidate AI application and the organization with a well-defined need. These aspects draw inspiration from the notion of *quality attributes* utilized in systems engineering to describe and quantify desired traits of a given solution.

### Model concerns

If the scope of the evaluation is solely oriented towards the actual AI model, the assessment can utilize the existing CLAIM guideline or use the following slightly simplified list aggregated from CLAIM:

#### Data

* Which data sources were utilized? (CLAIM #7)
* Subject eligibility criteria and selection (CLAIM #8+10)
* Data pre-processing & handling of missing data (CLAIM #9+13)
* Data element definition (CLAIM #11)

#### Ground truth

* Definition of a ground truth reference standard to enable replication (CLAIM #14)
* The rationale for selection of ground truth (CLAIM #15)
* Ground truth annotation process & tools (CLAIM #16+17)
* Inter- and intra-rater variability analysis and mitigation (CLAIM #18)

#### Model development

* Description of the model selection process, design, input/output (CLAIM #22)
* Use of software libraries, frameworks, and packages (CLAIM #23)
* Initialization of model parameters (CLAIM #24)
* Model structure (e.g. ensemble learning) (CLAIM #27)

#### Model training

* Data partitions – sample size and determination of same (CLAIM #19-21)
* Details of training approach (number models, hyperparameters, etc.) (CLAIM #25)
* Method & grounds for the selection of a final model (CLAIM #26)

#### Model evaluation

* Metrics used for evaluation (CLAIM #28-30)
* Methods for explainability or interpretability and validation of same (CLAIM #31)
* External validation (CLAIM #32)

#### Results

* Data flow (CLAIM #33-34)
* Model performance (CLAIM #35-37)
* Evaluation of limitations (CLAIM #38)

### Application concerns

Given that the AI model is mature enough to be provided as a complete application, an assessment of the product should consider aspects that address the organization's technical compatibility and long-term technical perspectives, such as maintainability and operational flexibility.

#### Maturity

Although the product may be fully MDR compliant, the evaluation should also take a history of prior use into account by considering:

* Any known past incidents
* Time on market
* List of other customers/installation sites
* History of customer collaboration/satisfaction & experiences
* Presence of third party vendor collaborators

#### Compatibility & Adaptability

To what extent is the product directly compatible with the existing equipment and infrastructure? Is the product able to adapt to changing circumstances?

* Adherence to standardized interfaces – can the application as a whole, or even its parts, be interfaced with other components?
* Data interoperability constraints, such as formatting requirements, device source, image quality. Are any of these constraints compatible with the existing IT systems present in the organization?
* Mode of deployment. How is the AI application to be deployed into the organization? E.g., as a standalone application, an integrated plug-in for other systems, or as a service (internal/external) that can be requested at need?
* Does the AI application provide an API for exporting current and past recommendations?

#### Manageability

What level of control is provided to the operator?

* Does the system support parameter tuning post-deployment?
* How easy is it to recalibrate the system after it has been deployed?
* Does the system provide markers of algorithmic drift?
* Does the system adhere to the organization's requirements (speed, bandwidth, storage, etc.)?

#### Security

What additional cybersecurity challenges and concerns does the new system impose on the organization?

* Does the system depend on external resources to function?
* Does the system require the organization to expose services or interfaces externally?
* To what extent can the system be operated in an internal siloed environment?
* How important is the integrity of dependent resources such as validation datasets?

#### Usability

To what extent are the human-computer-interaction perspectives addressed by the application?

* Are there formative and summative evaluation results available describing the usability and acceptance of the product? Is AI application readily accessible and easy to use?
* Adaptation: To what extent can the user interface and workflow constraints be adjusted to local circumstances?
* How is the AI guidance provided? E.g., actively intervening in the clinical decision process, as a service that is summoned by the end-user, or as a process that monitors human decision making and intervenes given some threshold of disagreement.
* When is the guidance provided? In real-time or through retrospective post-processing.
* Needed level of end-user clinical and technical expertise required to interpret and utilize the provided guidance.

## Domain 3: Ethical aspects

The ethics of artificial intelligence in healthcare includes attention to beneficence & patient integrity, privacy, equity, trust, autonomy, accountability & responsibility, and transparency. It is essential to draw attention to ethics:

* in the AI model development phase to proactively tackle bias and embed stakeholder values and ethical principles in the development of AI systems[[1]](#footnote-2).
* to clarify the benefits and risks of a specific AI application under scrutiny.

The above-mentioned ethical attention points are followed by specific outcome measures and recommendations.

**Beneficence and patient integrity:** Address the risk of overdiagnosis, and false-positive and negative results, which may raise a concern or unnecessary invasive procedures. Similarly, consider whether AI predictions indicating low-probability disease risks may cause undue concern for the patient. Further, assess the risk of misdiagnosis and patient harm in the model development phase by carefully curating training datasets and test datasets aligned with recommended standards. In the deployment phase, it is equally important to monitor the AI application's performance (these issues are handled in domain 6: "Clinical aspects").

Seek to enhance patients’ e-health literacy by empowering patients and raising awareness of the role of AI in healthcare.

**Privacy**: Consider how to protect patients’ right to privacy by using GDPR as a suitable instrument to regulate AI (in tandem with the upcoming EU AI Regulatory Framework), data ownership, consent models, data management, and data security. Furthermore, it is essential to establish guidelines and frameworks for using clinical data for research and development (also, see domain 4: "Legal").

**Equity** (fairness): Assess if the AI application promotes equity and diversity by facilitating equitable use and access to AI healthcare applications. In the model development and deployment phase, consider the risk of culturally biased health data, which may negatively impact patient groups (e.g., AI skin cancer diagnostic systems trained on white people may misdiagnose other skin colours). Likewise, consider the overall risk of stigmatization or discrimination toward specific patient groups due to encoded bias (concerning, e.g., gender, ethnicity, and religion).

**Trust**: It is essential that AI-enhanced clinical decision-making is trustworthy. To assess the interplay between clinical practice and AI concerning trust, consider the risk of lack of confidence in AI diagnostics by paying attention to **transparency, accountability and responsibility** issues. But also, consider **privacy**, e.g., that patient confidentiality is not threatened by AI, which may otherwise cause distrust in the healthcare system and healthcare professionals (also, see domain 9: "Patient aspects").

**Autonomy**:It is essential to ensure human oversight and control of AI applications. To do so includes further attention to:

* **Accountability & responsibility**:Assess if and how the AI application challenges the organizational chain of accountability (this is handled in domain 4: "Legal" and domain 8: "Organizational aspects"). To determine liability issues, a clear chain of accountability must be in place before deploying AI.
* **Transparency** is essential to validate results and to justify and legitimize decision-making. In the AI model development phase, assess the level of model interpretability, i.e., the degree to which a skilled AI developer understands the model's output at the technical level. In the AI deployment phase, assess AI explainability, i.e., does the AI system design have an explainable interface that conveys understandable information? (This is handled in domain 2, i.e., in the CLAIM guideline).

## Domain 4: Legal

In assessing AI, like other technology, it is important to map the legal landscape; otherwise, it is not possible to make a legal AI model. Figure S3 illustrates a general mapping (not exhaustive) of the legal landscape and shows the complexity of the legal rules with data as the starting point, see the green brick in the figure. The figure focuses on national laws with examples from Denmark and EU-law most of all because the EU is the front runner in many ways when it comes to regulation of AI. From a more international perspective OECD’s[[2]](#footnote-3) and Council of Europe’s[[3]](#footnote-4) works on AI are other examples. Further, some countries outside the EU also have laws regulating AI. For example Canada[[4]](#footnote-5), Singapore[[5]](#footnote-6) and Japan[[6]](#footnote-7).



Figure S3. The legal map with cooperation from a supplier and/or another partner (it is not an exhaustive list)[[7]](#footnote-8) [[8]](#footnote-9) [[9]](#footnote-10) [[10]](#footnote-11) [[11]](#footnote-12)

The figure is developed on the assumption that data contains personal data otherwise, this has an impact on the legal landscape (for instance, the GDPR does not apply). The project phase is the blue bricks, while the red bricks are the operation phase. Figure S3 is a puzzle because even if you divide the development of the AI application into these two steps (the project phase and the operation phase), you still have to consider the legal landscape for the whole lifecycle. Further, figure S3 has many repetitions between the blue and red bricks. Even if it is the same law that applies, you will activate different sections depending on where you are in the development of the AI application.

**The project phase:**

In the project phase (the blue bricks in figure S3), the following considerations and documents are relevant:

* Many different legal competencies are required when working with AI. Therefore, it is important to consult legal assistance early.
* It is important to contemplate the operation phase here because you have to ensure that the AI model is compliant with the law in its whole lifecycle.
* Is a Risk Impact Assessment performed?
* Is a Data Protection Impact Assessment (DPIA) performed?
* Is there a map of the legal landscape?
* Is there a dataflow description (handled in domain 2)?
* Is a technical description (handled in domain 2) made? Especially GDPR, MDR, and the proposal to an AI regulation that fuses law and IT security.
* Is the application process for regulatory approvals started? (if relevant)
* Is the AI model developed so it is easy to adapt to new legislation, as an example proposal about an AI regulation?
* Is there a process for compliance with the surrounding legal requirements (has a direct impact on the AI application), such as competition law, procurement law, the right to compensation and intellectual property rights

If there is a supplier involved in this phase, then these documents are also relevant to take into consideration:

* IT contract with the supplier incl. data processor agreement (if the supplier is a data processor) and the contract should have a focus on the regulating of;
  + the ownership/patents of the AI model, such as the financial standing. This can have a big impact on the business case of the AI application (handled in domain 7).
  + damages incl. responsibility for misdiagnosis
  + requirements based on the map of the legal landscape, including data security and privacy
  + technical requirements (handled in domain 2)

If there is a partnership in this phase, then this document is also relevant to make;

* A cooperation agreement which should focus on the regulating of;
  + The same points as in the IT contract

A supplier can also be in a partnership, but you will still make a cooperation agreement, often called private-public cooperation.

**The operation phase:**

In the operation phase (the red bricks in figure S3), the following considerations and documents are relevant:

* Is the Risk Impact Assessment done?
* Is the AI model MDR compliant – look after the CE-mark?
* Is the Data Protection Impact Assessment (DPIA) done?
* Is the mapping of the legal landscape in place?
* Is the dataflow in place?
* Is the technical description (see domain 2) in place?
* Are the regulatory approvals in place? (if relevant)
* Is there a plan for the ongoing work with legal compliance?

If there is a supplier involved in this phase, then this document is also relevant to take into consideration

* Is the IT contract with the supplier incl. data processor agreement (if the supplier is a data processor) clear?

In the IT contract, it is important to focus on the regulation of;

* + the ownership/patents of the AI application, such as the financial standing. This can have a big impact on the business case of the AI application (handled in domain 7).
  + damages incl. responsibility for misdiagnosis
  + requirements based on the map of the legal landscape, including data security and privacy
  + technical requirements (see domain 2)

If there is a partnership in this phase, then this document is also relevant to make;

* Cooperation agreement. In the cooperation agreement, it is very important to regulate the following;
  + The same points as in the IT contract
  + A supplier can also be in a partnership, but you will still make a cooperation agreement, often called private-public cooperation.

The black brick "the binding brick" in figure S3 consists of basic principles and best practices.

* Basic principles are the consistent foundation in the AI application
  + Basic principles, particular data ethics below mentioned bias and transparency for the data subjects (also, see domain 3: "Ethical aspects")
* The best practices are helpful when the law is not providing an answer
  + Be aware that not all law articles are written clearly, and you especially see that in the regulation of technology. Therefore you often have to look into best practices and work within other fields, such as technology (see domain 2: "Technology"). As an example, the GDPR says: *“[…] shall implement appropriate technical and organizational measures to ensure a level of security appropriate to the risk […]”[[12]](#footnote-13),* but there is no clear answer to this in the GDPR, but help and clarification can be found in "best practices," e.g. in the ISO standard 27005. This also shows the association between law and fields.

## Domain 5: Safety

This domain includes a description of the safety requirement as one of the main concerns in the implementation of AI models into clinical practice. Safety is known as a basic requirement for implementing any new technology into practice; therefore, AI models will not be legitimate unless they compile the safety standards. This requires assurance regarding preventing extra side effects and following the famous expression of "No Harm" to the patients for any new medical technology.

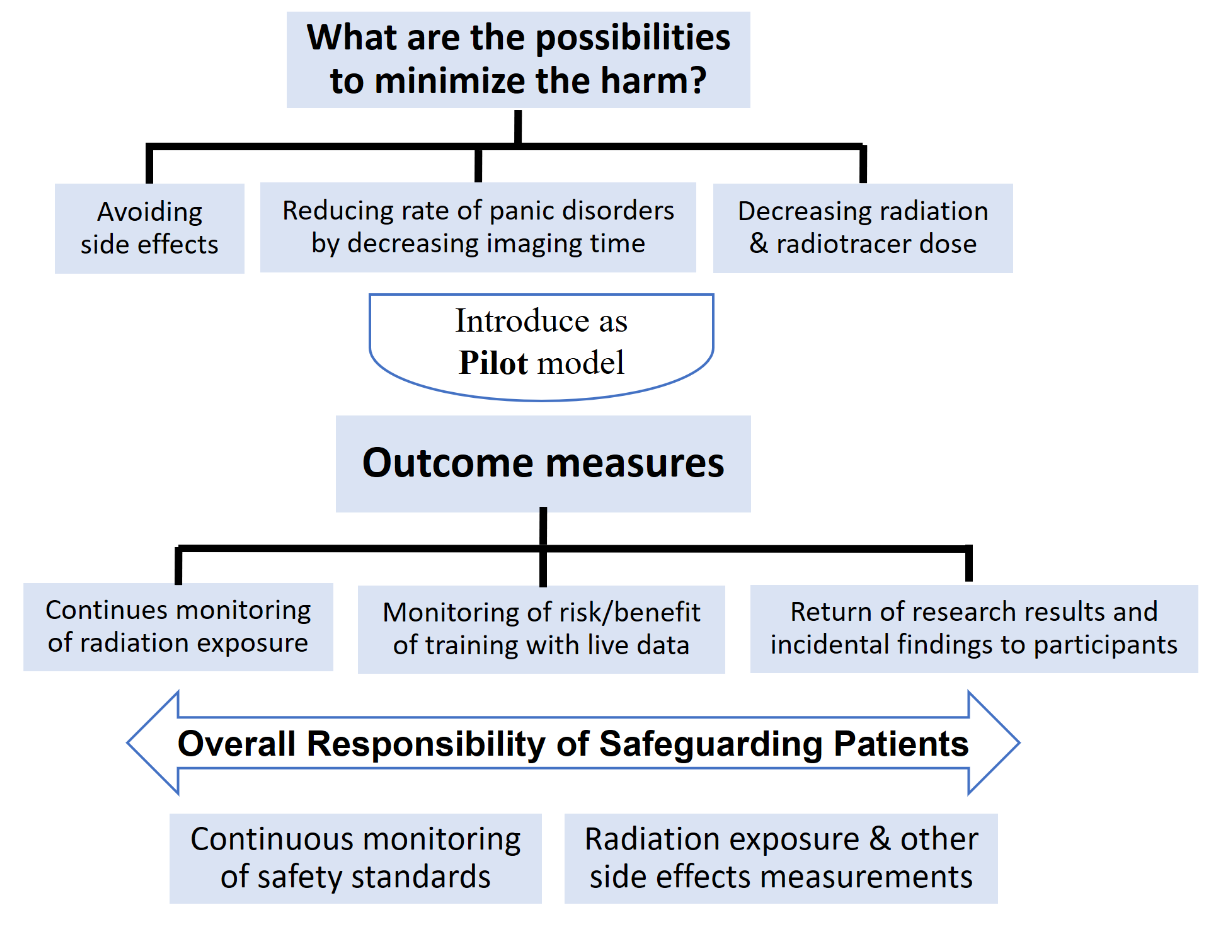


Figure S4. Summary of safety assessment, outcome measures, and main expectations through AI models

**Important outcome measures:**

* Adverse events
* Comparison of the level of received harm by patients (i.e. radiation exposure) with standard radiology methods
* Methods adopted for continuous monitoring of safety and new practice, i.e. establish and describe quality assurance (QA) program.
  + Constantly monitoring of the risk/benefit of training with live data for unintended consequences (failsafe system)
  + During the initial stages of deployment, the output of the automatic segmentation software should be treated as if a trainee had performed the contouring − careful review is essential.
  + Return of research results and incidental findings to research participants.
  + Does the application provide mechanisms for logging decisions and recommendations for later audits?
* Considering generalizability criteria and mechanisms to avoid patient harm in the rare case
* Considering platform safety criteria regarding data safety assurance

## Domain 6: Clinical aspects, e.g., clinical effectiveness

It is sensible to use validated clinical outcome measures where these are appropriate, as they can facilitate comparisons between the findings of different studies. However, suitable validated instruments are not always available. Further, which particular outcome measure is used in an evaluation depends on which diseases and populations (e.g. demography) are considered relevant to assess. Suggested validated outcome measures, however not limited to:

1. Sensitivity, specificity and receiver operating characteristic curve (ROC)
2. Mortality and morbidity
3. Time to an event, e.g., time to treat or decision
4. Quality of life

## Domain 7: Economic aspects

The economic analysis aims to estimate the impact of using AI technology in medical imaging on the use of resources per patient and contribute to a comparative assessment of costs and outcomes of implementing an AI technology, e.g., in a cost-effectiveness analysis. An economic analysis should include the following elements:

* Costs of using the AI technology:
  + Implementation costs (data collection including sorting and curating AI data in the electronic system, integration of IT systems, training of staff, changes to the workflow, hardware, and software)
  + Running costs of using AI (e.g., payment to the manufacturer of AI)
* Economic consequences of using the AI technology:
  + Changes in costs related to time, e.g., radiologist time to make classification and imaging interpretation per patient
  + Changes in costs related to the number of diagnostic examinations per patient
  + Changes in costs related to the medical treatment of the patients
  + Changes in patient time and costs related to transportation e.g., to the hospital for medical imaging.

Notice that the costs of the development of the AI technology will often be included in the running costs to the producer of the AI model. However, if further resources are needed for local adjustment of the AI model, these costs must be added.

**Perspective**

The perspective of the economic analysis is important. Ideally, the economic analysis should estimate the economic consequences of implementing an AI technology both from a broad societal perspective and from the narrow perspective of the health care institution, e.g., the hospital implementing the technology. Whereas the first will include all types of economic impacts in both primary and secondary care and costs to the patient and production costs (e.g., as in a cost-effectiveness analysis), the latter will focus on changes in expenditures and income or reimbursement for the hospital alone (e.g. as a business case).

Empirical outcome measures in a study of the costs of implementing AI technology in medical imaging could be:

* Number of minutes used by staff for interpretation of a medical image
* Number of diagnostic examinations per patient
* Number of patients needing biopsy
* Duration of a diagnostic examination per patient
* Number of hours and number of kilometers of transportation per patient for diagnostic examination

The economic evaluation should follow international guidelines for the conduct of health economic evaluation of health interventions as stated by CHEERS – Consolidated Health Economic Evaluation Reporting Standards, see Husereau et al. (2013).

## Domain 8: Organizational aspects

Regarding organizational aspects, the aim is to capture benefits in the form of reductions in workflow and tasks related to imaging for the staff because of AI. Further, the use of additional time related to implementation and training and the challenges of ensuring acceptability are important to capture. Specific outcome measures that may be relevant:

1. Changes in time for the health care professionals and patients:
   1. User time (e.g., the time required for AI application to make a diagnosis and treatment recommendation compared to the senior consultants)
   2. Time used on additional education or training needed to use AI application, e.g., knowledge of statistics and data science
2. Describe the user perspective, i.e., the interaction between user and the technology:
   1. Patient and clinician acceptability, reassurance/trust, satisfaction, the convenience of patients and physician
   2. Explore the clinical setting: What is the extent of "no-use" of AI among clinicians? Explore the cultural mindset or norms among the staff: Are there an incentive to want the changes? How is the employee trust in the AI model?
3. Staff (e.g. radiologist) burnout
4. Output (report) comprehensiveness

Further, some implementation, or organizational requirements and recommendations, are to:

1. Describe needed changes in workflow, i.e., describe workflow with and without the AI regarding:
   1. Frontstage, i.e., image interpretations leaving the clinical department
   2. Backstage, i.e., the processes that take place internally in the clinical department
2. Investigate if - and how - management anchoring is achieved?
3. Does the AI application change clinical decision making, i.e. elaborate on:
   1. When does the AI application propose an action? How and who will actually implement it? And is staff's approval needed for action proposed by the AI?
   2. Intended to replace physicians in a role they can already perform well or, alternatively, provide novel information that clinicians would not otherwise have
   3. Are people displaced by these technologies or role expansion? What is the risk of “de-skilling”, i.e., the clinician loses an important clinical skill due to the AI performing that task?

## Domain 9: Patient aspects

This domain contains the potential benefits, upcoming challenges, important outcome measures, and required research related to patients and social aspects. As the main customers of the healthcare system, patients have a key role in integrating AI-based practice. Developed AI models cannot be implemented into the healthcare system unless they receive the trust and satisfaction of the patient population. Patients' wishes and public acceptability of the new technology should receive high priority in further developing AI models at the level of research and clinical practice. The most considered parameters for this model are patients' comfortability and access to their own data/report in a safe and secured platform. Patient-based research must focus on patients' true wishes and perspectives.

* **Potential benefits of AI integration to the medical imaging field**

Patient benefits consisted of three main categories: Patients' willingness, technical improvement during the imaging process, and clinical-based patient benefits.

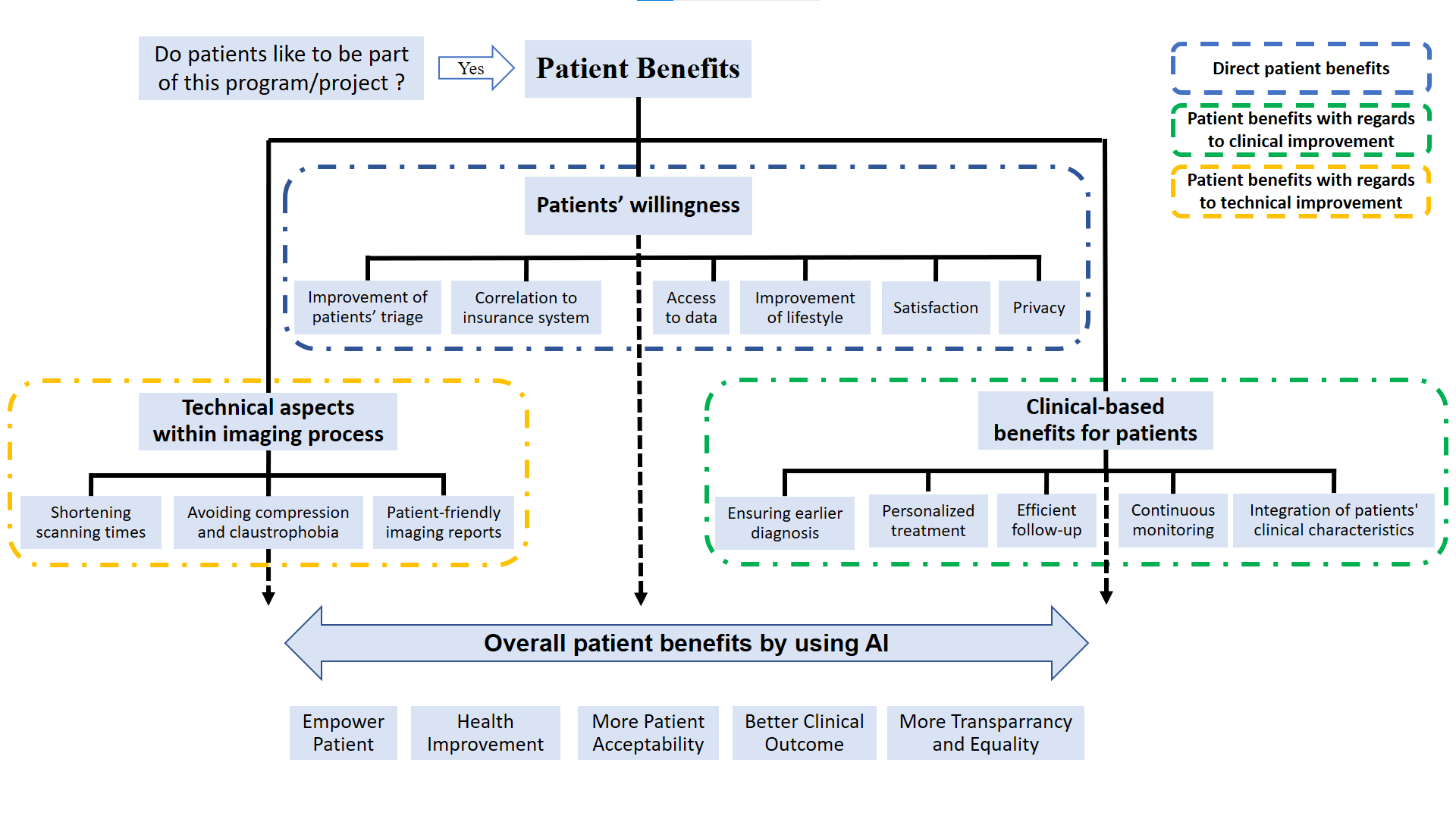


Figure S5. Summary of patient benefits through AI models which include patients' direct benefits, benefits regarding clinical improvement, and benefits with regards to technical improvement.

* **Important outcome measures which need to be considered:**
* Patient involvement in decision making
* Explainability of AI-based projects to the patients
* Using appropriate (preferably online) surveys for patients’ feedback on their experience
* Feasibility to reach patient-friendly imaging reports
* Patients’ access to imaging data
* Using standard patient satisfaction score for satisfaction criteria
* Considering patients with special medical conditions
* Considering imaging time and time required to receive imaging report
* Considering socioeconomic features of the target patient population
* Development of patient-friendly software/application for remote access
* Fulfill the patients' expectations regarding the use of the latest and greatest technology
* **Upcoming challenges:**
* When and at which level should patients be informed about the involvement of AI?
* Should it always be a free choice to be assessed by AI?
* Taking into account the different patient wishes
* Changing the patient's view to accept the use of AI instead of doctors

## Process factors for a MAS-AI assessment

Further, the following five factors should be considered during the process of assessing an AI technology:

1. Assess the maturity: Judge the potential for clinical practise implementation through classification in development phases, i.e., are we ready to move from step 1 to step 2?
   1. Use figure S1 to indicate the current stage of AI technology, i.e. project or operation (deployment) phase?
2. Assessment should be done on a regular basis during the AI deployment phase, so when should the assessment be revisited?
3. Use multidisciplinary development with active participation across all stakeholders – have a plan for when to involve which stakeholders.
4. Use a “Devil's Advocate-process” to counter hype and overpromising language in the assessments of AI, e.g., by having people in the assessment team who are skeptical towards the AI application.
5. The organization should have a guideline for implementation to ensure adaptation and integration into real-world existing workflows and context.

## References

Fasterholdt I, Naghavi-Behzad M, SB Rasmussen B, Kjølhede T, Skjøth M, Grubbe Hildebrandt M, et al. **Value assessment of artificial intelligence in medical imaging: a systematic scoping review** (in peer-review). 2022.

Fasterholdt I, Kjølhede T, Naghavi-Behzad M, Schmidt T, Rautalammi Q, Grubbe Hildebrandt M, Gerdes A, Barkler A, Kidholm K, E. Rac V, SB Rasmussen B. **Model for ASsessing the value of AI in medical imaging (MAS-AI)** (submitted for peer-review). 2022.

Husereau D, Drummond M, Petrou S, et al. **Consolidated health economic evaluation reporting standards**

**(CHEERS)–explanation and elaboration: a report of the ISPOR health economic evaluation publication**

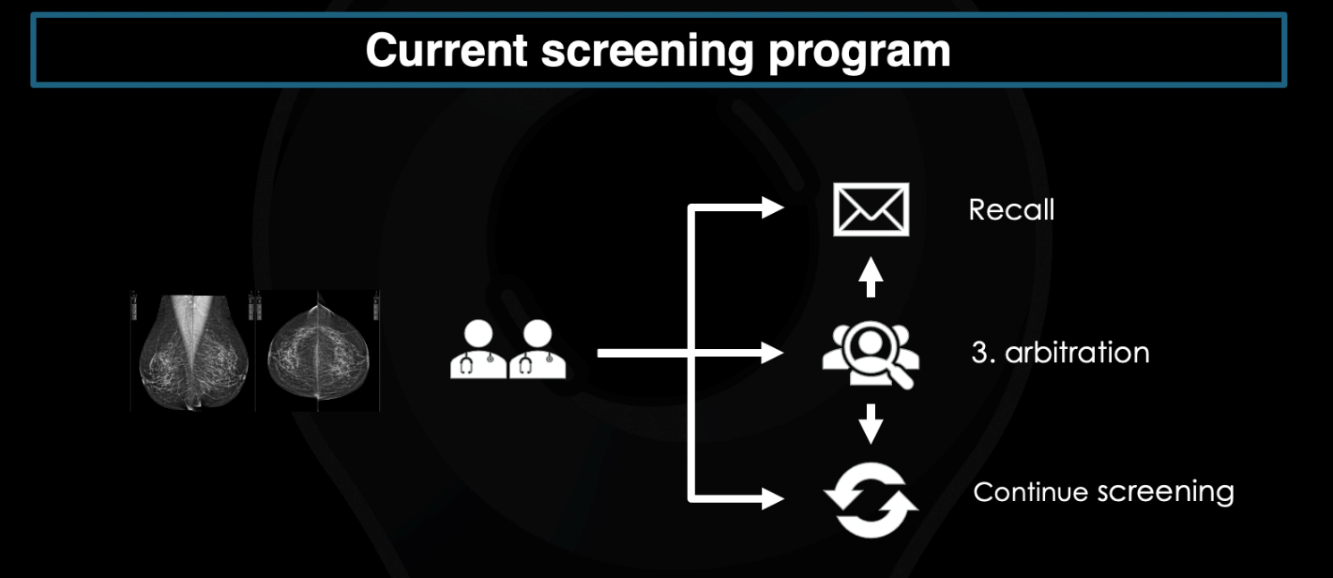
**guidelines good reporting practices task force.** Value Health 2013; 16: 231–250

Mongan J, Moy L, Kahn CE, Jr. **Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers.** Radiol Artif Intell. 2020;2(2):e200029.

# Supplementary S2

**Case: How to use the different parts in the MAS-AI guideline**

Here we provide cases as examples of how to use the parts in the MAS-AI guideline. Primarily we use a case from the area of breast cancer screening, i.e. the MAGIC project. Figure S6 shows how the patient voyage is today – and how it is altered with the use of AI. However, the MAGIC example is somewhat imperfect or hypothetical. Although CE marked and FDA cleared, no actual prospective assessment has been done yet, and MAGIC is still an immature technology. Thus, to demonstrate MAS-AI, we have filled out this case as if we were ready for implementation. A short video about the MAGIC project, see <https://www.linkedin.com/posts/radiology-research-unit_ai-mammography-magi-activity-6846730039178862592-jt1N/>



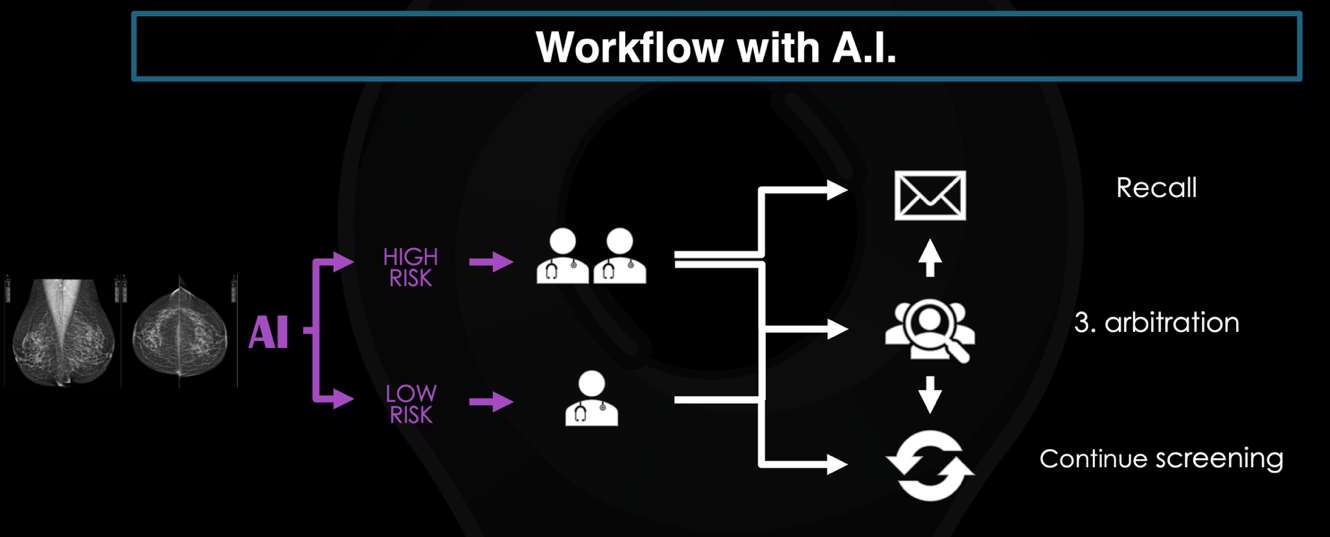


Figure S6. The patient voyage – today and with the use of AI

## Executive summary – January 2022

This report presents the (partly hypothetical) results of a MASAI assessment done in January 2022 on the MAGIC project. MAGIC is an artificial intelligence application used in breast cancer screening to potentially increase the quality and efficiency of breast cancer screening in Denmark. The assessment was done by a multidisciplinary group of experts and a patient representative. Next, we briefly present the results in nine domains.

**Domain 1: The health problem and current use of technology**

Breast cancer is the most common cancer and the second leading cause of cancer deaths globally. Breast cancer screening is implemented to reduce mortality through early detection. AI solution-based technology is proposed as a clinical tool that could potentially increase breast cancer screening quality and efficiency.  
  
**Domain 2: Technology**

The MAGIC AI tool has been designed, trained, and evaluated following the CLAIM guideline. Regarding deployment, no technical application issues have been identified.

**Domain 3-4: Ethical and legal aspects**

MAGIC is developed and tested on a large-scale validation study, which lowers the risk for encoded bias in MAGIC and provides reliable results. However, distrust in AI may be present among certain patient groups who are not comfortable with AI in general. If MAGIC is used for fully autonomous screening, there is a risk that failures (e.g., false-positive and negative) may go unnoticed and thereby harm patients. To avoid responsibility gaps, MAGIC should be used as a second reader of mammograms for low-risk groups. This might increase screening efficiency and ensure human oversight. Regarding legal issues, the project phase is adequately handled, despite the legal landscape being developed after the MACIG project's start.

**Domain 5: Safety**

In this study, the model is locked from continuous learning. Also, data storage and processing will be performed on-site, no data is shared outside of this location, and the software does not save any of the processed data. So currently, no safety issues are identified.

**Domain 6: Clinical aspects**MAGIC showed an improvement in detection accuracy in terms of the area under the receiver operating curve (AUROC) in this retrospective study.

**Domain 7-8: Economic and organizational aspects**

The use of AI in total reduces the average costs per patient by $156 and increases the number of quality adjusted life years (QALYs) by 0.0095 QALY. The AI changes clinical decision-making. Also, the radiologist's workload regarding interpretation is reduced while patient and clinician acceptability are unchanged (however, "no-use" poses a minor issue among the clinicians).

**Domain 9:** **Patient aspects**

There are no direct patient benefits in MAGIC. However, the findings of the current study may result in an improved clinical outcome which will be a gain for the patients in the long-term.

## Step 1

### Domain 1: The health problem and current use of technology

**Description of the health problem of the patient/health problem and current use of technology**

Breast cancer is the most common form of cancer amongst women in Denmark, with approximately 5000 new cases annually. Despite a 5-year relative survival rate of up to 89%, breast cancer has the second-highest mortality rate amongst Danish women across all cancers. Population-based organized breast cancer screening is implemented with the overall purpose of reducing mortality from breast cancer through early detection.

Danish women aged 50-69 years are offered biennial mammography screening. Following the European guidelines for quality assurance in breast cancer screening and diagnosis, all mammograms are read independently by two specialized breast radiologists. In cases of discordant opinions between the two readers, a third reading (arbitration) by an expert screening radiologist is required for a final decision. If a suspected lesion is detected, the woman gets recalled for assessment. Otherwise, she continues to the next round of screening. Recalled women are offered diagnostic mammography consisting of further imaging, a clinical breast examination and non-excisional biopsy if necessary. Double reading and arbitration have shown improvement of cancer detection compared to single reading; however, this might come at the expense of an increased false-positive recall rate. Thus, mammography screening is associated with a not so irrelevant number of false positives and, to a lesser extent, false negatives. Moreover, overdiagnosis and overtreatment pose an acknowledged risk of harmful physical and mental consequences.

**A description of the application assessed, i.e. intended use and clinical role of the AI approach**

In recent years, advancements in artificial intelligence (AI) applications based on novel deep learning (DL) methods have shown great potential to increase the quality and efficiency of breast cancer screening.

**Study objectives and hypotheses**

The latest studies on DL models for cancer detection provide comparative data on AI versus human interpretation of screening mammograms with quite encouraging results. Most studies show an improvement of detection accuracy in terms of the area under the receiver operating curve (AUROC) of the radiologists' performance with AI in support mode. Furthermore, a non-inferior performance of AI in standalone mode is shown in some studies, even surpassing human experts as a single reader in some cases. Yet, all existent studies have tested a DL-based AI system on either a retrospective dataset (mainly single centre studies) or in limited reader studies, thereby reducing the generalization of the results to a real-world clinical setting. Clinical validation studies and prospective trials are needed to provide evidence for the benefits and applicability of AI applications in Danish breast cancer screening practice.

**Study design of the model evaluation + aim and goal of the study**

A multicenter, non-inferiority simulation study of diagnostic accuracy and feasibility on a retrospective cohort.

### Domain 2: Technology

In its current state of maturity, the technical assessment of the MAGIC case would focus primarily on the modelling concerns. Drawing on the CLAIM guidelines, the MAS-AI framework can guide an evaluation committee through the necessary areas of concern – thus ensuring that the MAGIC AI tool has been designed, trained, and evaluated using industry best practices.

In a future scenario, where the AI diagnostic tool utilized in the MAGIC case has been thoroughly evaluated, achieved MDR compliance, and packaged in the marketable product – the technical domain of the MAS-AI is envisioned to support assessment of technical fit and long-term deployment within the organization.

E.g., a full technical assessment of the MAGIC diagnostic tool would entail a review of all of the application dimensions above, in addition to the model-oriented initial assessment.

### Domain 3: Ethical aspects

An ethical assessment of the MAGIC diagnostic tool for clinical decision support regarding breast cancer diagnostics implies attention to the below-mentioned points.

**Beneficence and patient integrity**

If MAGIC is used as an application for standalone reading in the triage mode of mammograms for low-risk groups, this might increase screening efficiency and thereby be beneficial. However, there is a risk that failures in fully autonomous screening (e.g., false-positive and negative) may slip under the radar and thereby harm patients. Consequently, MAGIC should preferably be used as a second reader of mammograms for low-risk groups; this might still increase screening efficiency and ensure human oversight.

**Privacy**

MAGIC is privacy compliant, and patients' confidentiality is not threatened.

**Equity**

The model behind MAGIC is developed and tested on the backdrop of a large-scale validation study, which includes a complete, unselected screening cohort spanning an entire population from the region of Southern Denmark. Thus, MAGIC is validated with local data in a local setting, which lowers the risk of encoded bias (concerning, e.g., patient demography) in the model. However, there is a risk that MAGIC may negatively impact equity if specific patient groups are not comfortable with AI solutions and therefore choose to drop out of screening programs.

**Trust in AI-enhanced clinical decision making**

Clinical practitioners may trust MAGIC as the DL model behind MAGIC is tested in a large-scale validation study, implying that image processing can be considered reliable. Moreover, MAGIC is transparent due to an explainable interface that provides understandable visualizations of tissue alterations. Hence, the validity and comprehensibility of outputs can be confirmed. On this background, it is possible to legitimize and justify decisions. Moreover, patients’ trust is also preserved as clinical staff can explain system output, and, furthermore, patient trust is facilitated because MAGIC is privacy compliant. However, as noted above, there still resides a risk that distrust may occur if AI gives rise to discomfort among certain patient groups.

**Accountability& responsibility**

If MAGIC is used for standalone readings without humans in or on the loop, it may give rise to responsibility gaps making it difficult to determine liability issues.

### Domain 4: Legal

In the project phase (the blue bricks in figure S3), the MAGIC project has talked to the legal adviser in some areas early in the project. However, the whole legal landscape and the Risk Impact Assessment, including Data Protection Impact Assessment (DPIA), were first developed after the MACIG project. The MAGIC project is still in the project phase, and therefore the project has attention to new legal requirements in the operation phase.

## Step 2

### Domain 5: Safety

Due to the retrospective nature of the MAGIC study, the processing of screening mammograms by the deep learning model will not affect any of the women in the study population in terms of the standard quality and offers of screening, diagnostic assessment, follow-up and treatment, and none of the women will suffer any physical or mental harm. In addition, continuously adaptive AI algorithms raise concerns regarding their safety and consistency in performance, but in this study, the model is locked from continuous learning, thereby eliminating this risk. Finally, since data storage and processing will be performed on-site, no data is shared outside of this location, and the software does not save any of the processed data.

### Domain 6: Clinical aspects

The endpoint for diagnostic accuracy is the difference in cancer detection performance between the

index test and the reference test. The metrics of performance are the overall AUROC across the AI

Score cut-off points as well as cut-off point-based sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and recall rate. E.g, The AUROC of AI models for detecting cancer, ranged from 0.893-0.908, and the prediction accuracy ranged from 87%-100%.

### Domain 7: Economic aspects

The MAGIC case is not mature enough to present an economic analysis. Hence, an example of an AI technology for improved detection of intracranial large vessel occlusions (LVO) on computed tomography angiography in stroke is used, see Van Leeuwen et al. (2021).

The technology is expected to increase diagnostic sensitivity without a decrease in specificity. The estimation of costs using Markov modelling includes the price per patient of using the AI model ($40). In addition, the economic effects include estimation of changes in costs related to a reduction in the number of missed diagnoses of LVO and changes in the health care costs related to acute stroke and the long-term care of stroke patients. The results show that the use of AI, in total reduced the average costs per patient by $156 and increased the number of quality adjusted life years (QALYs) by 0.0095 QALY, similar to a national cost saving of $11 million.

### Domain 8: Organizational aspects

In the organizational analysis of the MAGIC, we measure the following outcomes: 1) workload of radiologist regarding interpretation + time used on additional education to use AI-application, 2) patient and clinician acceptability, 3) No-use + cultural mindset among staff are investigated by focus group. Results show that the workload of radiologist regarding interpretation is reduced by 10%. Patient and clinician acceptability is unchanged. However, a focus group indicates that "no-use" poses a minor issue among clinicians.

Next follows a description of how the workflow and decision-making is affected.

**Describe the workflow with and without the AI**

With-out: All mammograms are read independently by two specialized breast radiologists. In cases of discordant opinions between the two readers, a third reading (arbitration) by an expert screening radiologist is effectuated for a final decision. With AI: One of the possible applications of a DL model for breast cancer screening is autonomous standalone reading in triage mode, in which cases are sorted according to suspiciousness of findings. Several studies indicate a cost-benefit when only cancer suspicious examinations are triaged to a single reading. This approach would significantly reduce the number of normal or non-suspicious mammograms needing radiologist interpretation, thereby reducing workload.

**Does the AI application change clinical decision-making?**

Yes. The DL model will be used as a triaging tool, while the reference standard diagnostic test is double reading with the arbitration. Thus, double reading is replaced with single reading at a certain optimal triage threshold score. Mammograms assigned with a particular threshold score and below (≤ threshold score) are considered non-suspicious and will go only to single reading (low-risk group), while the rest (> threshold score) go to double reading (high-risk group). In the low-risk group, the DL model is counted as first reader in favour of continuing screening and the radiologist as second reader (the other radiologist's assessment is left out). In the high-risk group, the DL model's evaluation is not counted in the decision making, and the decision outcome of the standard radiologist double reading and arbitration is used.

When in single reading mode (the low-risk group), the AI application indicates the likelihood that a tissue alteration is malignant, and it is shown to the user at two different levels: region-level and exam-level.

• Region level: the most suspicious regions are marked and assigned a score between 1-100, with 100 indicating the highest likelihood that a finding is malignant. The marks are also colour-coded to represent this malignancy likelihood and indicate whether the finding corresponds to a calcification cluster or soft tissue lesion.

• Exam level: based on the local findings, the system assigns on a whole-examination basis a score from 1 to 10 denoting the likelihood of cancer (NONAME Score), with 10 indicating the highest likelihood of a cancer being present in the exam.

### Domain 9: Patient aspects

Study participants were not directly involved in the design of the MAGIC project, considering the retrospective nature of the work. Therefore, there are no direct patient benefits in MAGIC. However, the current study findings may result in improved clinical outcomes, which will be gainful for the patients in the long term.

## Process factors for a MAS-AI assessment

1. Assess the potential for implementation in clinical practice, i.e. are we ready to move from step 1 to step 2? **NO (but we have filled out this case as if we were ready for implementation!)**
2. Are you using multidisciplinary development with a plan for when to involve which stakeholders? **YES**
3. Are you using a "Devil's Advocate-process" to counter hype and overpromising language in the assessments of AI, e.g. by having people in the assessment team who are skeptical towards the AI application, etc.? **NO**
4. Do you have a guideline for implementation? **YES**
5. When is the assessment planned? **2022** When should the assessment be revisited? **Not decided yet.**

1. There are ISO standards and several value-based and humancentric design methods, frameworks, and guidelines. For example, the EU Ethics guidelines for trustworthy AI, Ethics by Design guidelines, agile consequential scanning for responsible innovators, the value-sensitive design method, and the IEEE Ethically Aligned Design. Moreover, the European proposal for harmonized rules on artificial intelligence (the Artificial Intelligence Act (April 20, 2021)) will increase the amount of AI Ethics by Design tools and methods. [↑](#footnote-ref-2)
2. OECD: <https://oecd.ai/en> [↑](#footnote-ref-3)
3. Council of Europe: <https://www.coe.int/en/web/artificial-intelligence> [↑](#footnote-ref-4)
4. Directive on Automated Decision-Making: <https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592#appA>. See also the comparing European and Canadian AI Regulation: [Regulating AI: A new report from the URC on Accountable AI and the LCO compares two leading international approaches | Faculty of Law – Civil Law Section | University of Ottawa (uottawa.ca)](https://droitcivil.uottawa.ca/en/news/regulating-ai-new-report-urc-accountable-ai-and-lco-compares-two-leading-international). [↑](#footnote-ref-5)
5. <https://www.globallegalinsights.com/practice-areas/ai-machine-learning-and-big-data-laws-and-regulations/singapore>. [↑](#footnote-ref-6)
6. <https://www.globallegalinsights.com/practice-areas/ai-machine-learning-and-big-data-laws-and-regulations/japan> [↑](#footnote-ref-7)
7. <https://eur-lex.europa.eu/legal-content/DA/TXT/?uri=celex%3A32016R0679> (GDPR). [↑](#footnote-ref-8)
8. <https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:52021PC0206> (proposal to an AI regulation). [↑](#footnote-ref-9)
9. <https://eur-lex.europa.eu/legal-content/DA/TXT/?uri=CELEX%3A32017R0745> (MDR). [↑](#footnote-ref-10)
10. <https://www.datatilsynet.dk/media/7753/danish-data-protection-act.pdf> (DBL). [↑](#footnote-ref-11)
11. The European Commission has in October 2021 initiated a consultation process prior to submitting a concrete feed for regulating product liability in relation to AI. <https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12979-Civil-liability-adapting-liability-rules-to-the-digital-age-and-artificial-intelligence/public-consultation_en> [↑](#footnote-ref-12)
12. Article 32 in the GDPR. [↑](#footnote-ref-13)