Guidance for Procurement, Assessment and Commissioning of AI Tools in Irish Healthcare Settings

The Report of IAPM AI-SIG Work Group 1

Date: August 2025

The IAPM does not endorse any products, manufacturers, or suppliers. Manufacturers and products named herein are for the purpose of providing examples and should not be interpreted as an endorsement.

Disclaimer

The information provided in this document is for general informational purposes only. While we strive to ensure accuracy and reliability, we make no guarantees regarding the completeness, suitability, or validity of the information provided. Use this information at your own risk. This advice is not intended to replace professional advice. Always consult a qualified professional for specific concerns.

We assume no responsibility for any errors, omissions, or results obtained from using this information. By using this resource, you agree to the terms outlined in this disclaimer.

Clinical AI is a multidisciplinary space, bringing together computer science, radiology, radiography and medical physics

There are many challenges of AI adoption, from regulations, ethical and legal concerns, interpretability and explainability to validation.

Authors

Ronan Coleman[1], Elizabeth Keavey[2], Peter Conneely[3], Luke Oonan[4], Ciaran Malone[5], Dara O'Gallchobhair[6], Michael O'Neill[6]

Disclosures

Nothing to declare.

Affiliations

[1] St. James's Hospital
[2] Breastcheck National Screening Service
[3] Galway Clinic
[4] Mater Misericordiae University Hospital
[5] St. Luke's Radiation Oncology Network
[6] Beaumont Hospital

Contents

1. Introduction

The intent of this document is to provide general guidance for commissioning Artificial Intelligence tools that are becoming increasingly available for radiology and radiotherapy clinical tasks. The focus will be on practical recommendations for procurement, acceptance testing and quality assurance within the Irish context. Large Language Models (e.g. ChatGPT, CoPilot, Gemini) and other AI productivity tools are *not* within the scope of this document.

Today, AI is pervasive, with the most widely recognised models being general-purpose rather than task specific. Examples like Chat-GPT or CoPilot are large language models that generate text by predicting the most probable next word.

In medicine, AI typically consists of specialised models designed to perform a single task, with limited ability to generalise to other functions.

AI is most commonly defined as computer models that learn from data. Traditional machine learning models often depend on features crafted by humans. Neural networks, on the other hand, are specialised models with a structure loosely inspired by the human brain. The advantage of neural networks in imaging lies in its ability to learn features directly from the images themselves. Deep learning, the most advanced and widely used form of AI, refers to models that utilise extremely large neural networks.

AI has the capacity to profoundly impact radiology, with possible positive and negative consequences. The integration of AI in radiology has the potential to transform healthcare by progressing diagnosis, quantification, and management. However, the expanding availability of AI tools in radiology underscores the urgent need to critically assess claims about their effectiveness and to distinguish safe products from those that may be harmful or ineffective.[1,2]

This document serves as a guide to support the Medical Physics Expert (MPE) in leading the implementation of AI solutions within the healthcare setting, particularly during procurement, acceptance testing, commissioning, and QA. The MPE is uniquely positioned to function as a vital link between the manufacturer and the clinical team.

2. Identifying the AI Tool

2.1 AI vs Algorithms – Is the "smart system" AI or just a deterministic algorithm?

The Artificial Intelligence (AI) Act of 2024 defines an AI tool as

"*a machine-based system that is designed to operate with varying levels of autonomy that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments*".[3]

The primary objective of this subsection is to highlight some key differences for users in distinguishing between AI systems and general computer algorithms. In the below paragraphs three key differences are discussed.

2.1.1 Hyperparameters vs Parameters

For us to understand the key differences between AI models and traditional computer algorithms, we must first understand the difference between hyperparameters and parameters. A hyperparameter is an external variable that cannot be estimated from the data provided to the model. In terms of an AI algorithm, these are specified during the learning or training phase to help the model estimate the parameters it needs to produce the best results from that specific dataset. Parameters in comparison are generated from the data provided to the model and are not provided by the programme's author. For AI systems a combination of both Hyperparameters and parameters are used with the developer specifying hundreds of hyperparameters and the model itself specifying millions of parameters.[3,4]

2.1.2 Data Driven vs Rule-based Decision Making

One of the key differences between AI tools and traditional algorithms is how they make decisions. AI algorithms require huge databases of "training" data to learn patterns and features contained within the information. The quantity of these learned parameters can range from hundreds of thousands to millions, and almost all (except the hyperparameters) are set without the explicit input of the user. The AI algorithm's performance is dependent on the quality of the training dataset it is provided. These AI algorithms are also specific in their training and are generally not generalisable in their use, i.e., a noise reduction algorithm for Magnetic Resonance Imaging (MRI) cervical spines will not work on brain imaging. Traditional algorithms, in comparison, have explicit rulesets and constraints outlined by the programmer, allowing the algorithm to know exactly how to process each new piece of information. These algorithms cannot learn from the data they

are processing and any improvements to the algorithm must come directly from the user or programmer.[6]

2.1.3 Deterministic vs Probabilistic Results

Traditional algorithms require clear and consistent data to produce good-quality results and do not function well with ambiguity. The results of these algorithms are generally expected as they are developed around relationships or equations known to the designer of the application. These types of results are often referred to as "deterministic" in their nature. AI algorithms, in comparison, can make connections between information in ways humans cannot. This is due to their ability to learn from data and extract relationships between multiple different parameters. AI Algorithm can also deal with data that is less streamlined and ambiguous than traditional algorithms. These data types are often more effectively presented using probabilities and confidence intervals. These are known as probabilistic results.[6]

2.1.4 Dynamic vs Static Learning

Another key difference between an AI tool and a traditional algorithm is its ability to learn from the data it is provided with. Traditional algorithms are designed with specified parameters and inputs predefined by the programmer. These algorithms cannot learn from the data they are processing, however, they can be updated and further revised by human operators. AI Algorithms in comparison can learn from the data they are tasked with analysing. In the European Union while these "continuously learning" tools are not explicitly prohibited under the AI Act and Medical Devices Regulation (MDR), strong human oversight is mandated to ensure continuity in their outputs. Traditionally, the EU's regulatory framework has favoured AI tools which have ceased their learning phase (static) before being released to the public. The reasoning for this is that it is difficult to determine if a continuous learning tool is the same product that has originally received CE marking, over the lifetime of the tool. This landscape, however, may change and we could expect to see these kinds of tools being CE-marked for medical use in the future.[7]

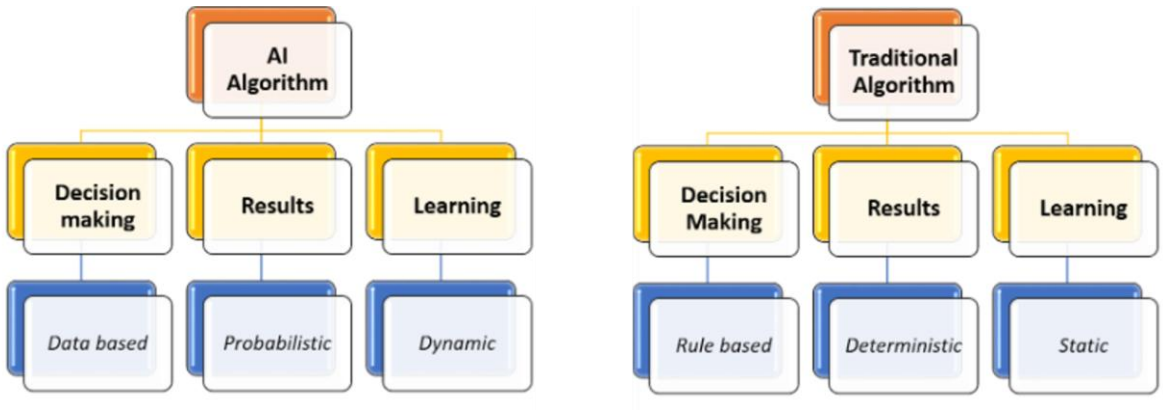

Figure 1: Key differences between AI algorithms and traditional deterministic algorithms.[4]

2.2 Decision Support vs Decision Making Tools.

2.2.1 Decision Support Tools

Decision support tools allow for shared decision making between the clinician and the tool in question. The AI or Machine Learning (ML) tool provides clinicians with an additional review of imaging or treatment progress; however, it is important to note that clinicians still hold full responsibility for the patient's diagnosis.[8] AI tools used in Medical Imaging examinations generally utilise terms such as “probability” or “likelihood” when conveying information about patient imaging or treatment results. This essentially shifts the responsibility to the clinician to make the final patient diagnosis. Examples of decision support tools already in clinical use for radiology include bounding boxes for perceived lesions on mammography images or general x-ray fracture detection. It is important to note that these tools are not just limited to Radiology/Radiotherapy and can be implemented in all aspects of patient care from assessment to support options.[9] A key component of decision support tools is the requirement of the “Human in the loop” i.e. there is always a human observer or clinician monitoring and deciding on the final clinical outcome for the patient.

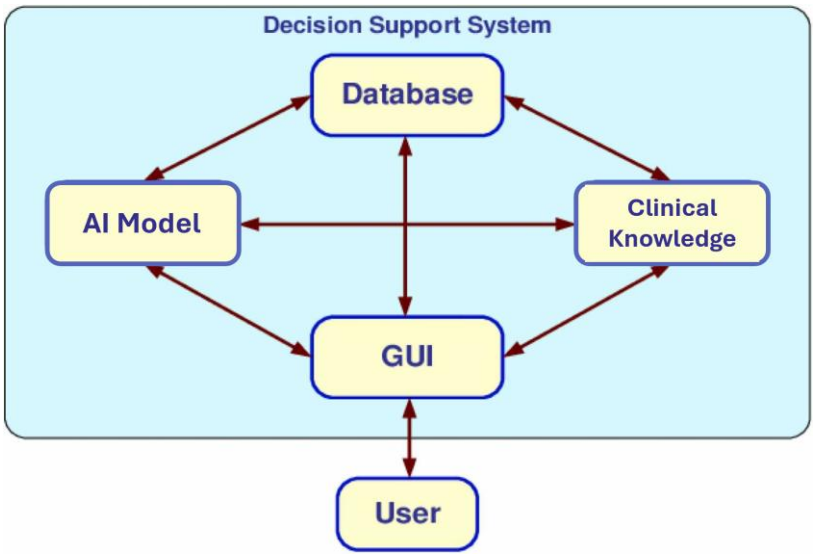

Figure 2: Components of a general decision support system [10]

2.2.2 Decision Making Tools

Decision making tools, in comparison to decision support tools, imply that they complete their task or diagnosis without explicit input or oversight from a clinician or human observer. In terms of AI products this would be considered a high to unacceptable risk due to their lack of oversight, which is a key obligation of the EU AI Act of 2024.[11] Human oversight in the decision-making process of AI tools is a major factor in risk management and from a patient welfare perspective. A recent Citizens Jury produced by the Irish Platform for Patient Organisations, Science and Industry (IPPOSI) in 2024 outlined 25 recommendations presented to the Minister for Health and the Minister for Enterprise, Tourism and Employment. The first of these recommendations was related to human oversight and the importance of maintaining the primary focus of healthcare between the provider and the patient.[12] In the future we may see decision-making tools more regularly in our healthcare institutions performing various tasks; however, at present, these are highly controlled and regulated forms of AI tools and are limited in their output on the market.

2.3 Classification of the AI Tool

2.3.1 EU AI Act

The final part of Section 2 seeks to inform the reader on how best to categorise the various AI tools that may be present in their medical institution based on the risk they pose to the patient's treatment or diagnosis. The EU AI Act categorises a large percentage of Artificial intelligence tools used in a healthcare setting as high-risk and outlines the necessary steps the institution and/or provider must follow to comply with the regulations.[10] If you are unsure of which risk category your AI tool falls into, a useful compliance tool is provided by the European Union here:

EU AI Act Compliance Checker (Website)

This compliance tool will determine if your AI tool falls within the scope of the EU AI Act and which category of tool you have at your medical institution.

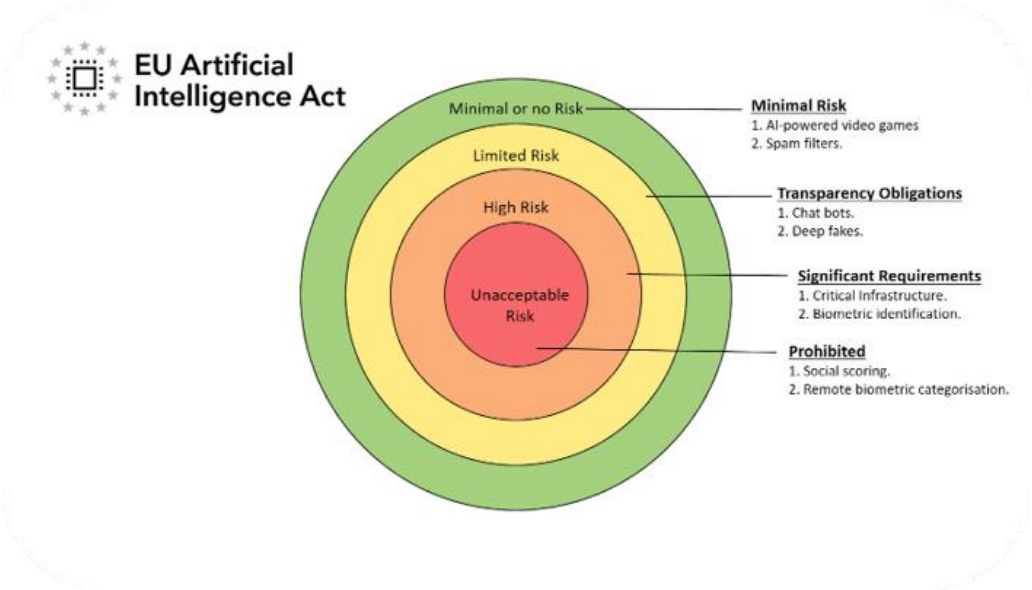

Figure 3: Risk categories as outlined by the EU AI act of 2024.

2.3.2 IAPM AI Tool Risk Assessment

For greater resolution and categorisation of these tools, we have decided to create subcategories to better stratify the risks each tool presents to the patient. To complete this sub-categorisation please download the accompanying risk template that follows these guidelines. The user should answer the questions to the best of their ability and will then be presented with a risk categorisation for the tool in question. It is to be noted that this sub-categorisation does not seek to replace the initial assessment provided by the EU AI Act but seeks to be employed alongside it. The additional sub-categorisation divides the high-risk category further into four additional subgroups based on the scores obtained by completing the risk assessment tool. If the user is unsure of the answer to one of the questions or is not provided with the required information to accurately complete it, we recommend assigning the question the highest score possible. A user guide is also provided alongside the risk assessment template to help you progress through the questionnaire. Table 1 below outlines the four subgroups and how they are divided.

Table 1: IAPM AI tool risk assessment sub-categorisation

Risk assessment score	Class	Risk categorisation
0 - 25%	I	High Risk Class I

25 - 50%	II	High Risk Class II
50 - 75%	III	High Risk Class III
75 - 100%	IV	High Risk Class IV

2.3.3 Sample Risk Assessment

Consider the following scenario:

A consultant radiologist approaches you regarding the implementation of a fracture detection tool within the emergency department of your hospital. The consultant provides you with the name of the software along with a brief description of what the tool does. While searching for information about the tool, you come across the website radiology.healthregister.com which has a breakdown of the product and its features.

The AI tool aims to highlight fractures within general x-ray extremity images. The tool provides three levels of confidence regarding fracture assessment (Positive – above 90%, Doubt – between 50 and 90%, Negative – lower than 50%). The fracture detection tool has 17 peer-reviewed papers linked on the radiology health register and is a subscription-based service. The tool assesses the DICOM files allowing it to assess images from different vendors. The tool can be installed locally on dedicated software or can be cloud-based. The tool is CE marked, FDA approved and meets the Medical Device Regulations (MDR).

Your task is to determine the risk category of this tool within the confines of the European AI Act.

Below is the risk assessment calculator completed using the information provided above. If you find that information related to one of the questions cannot be found through product research, please score it the maximum of 5 points.

Criteria	Relative Risk Score
What type of tool is in use?	
Identification tools	1
Number of Independent Statistically Significant Studies	
>10	1
Are Effective Metrics Provided for this tool?	
Yes	1

Confidence levels

Criticality of the Clinical Context	
Important but not life-threatening	3
Interoperability	
Agnostic	1
Part of the patient pathway using AI	
Detection	4
Role in patient pathway	
Supportive tool	1
At what point is the data being affected	
DICOM Data	3
How often is the AI tool being used?	
Very Frequent Daily or Multiple/day	5
Does the tool provide the user with any indication of what it is doing?	
Bounding Boxes	2
Timeframe of action	
Real Time	5
Model Assessment	
Internal audit	3
Human in the Loop	
Assisted Decision-Making	3
Percentage of patients' procedure impacted by AI	
>75%	5
Total Risk Score	38

Consultant or Registrar always reviewing imaging

Arguably greater use should provide lower score however greater risk if fault is found

Based on end-user assessment

Stated as intended use according to CE Marking certification

Based on the output of the above risk assessment calculator, we can see we get an overall risk rating of 38 for this AI tool. This falls into Class 3 (Score between 33 – 51) of our Risk Rating, determining that the overall classification for the tool based on the European AI act is:

High Risk Class 3

It is important to note the above questions are subjective and different assessors may record different risk ratings however these should be within a similar range of values.

Table 2: Score Thresholds for Risk Classes

Risk Categories	Score Ranges	
High Risk Class 1	0	17
High Risk Class 2	18	34
High Risk Class 3	35	51
High Risk Class 4	52	70

3. Procurement

3.1 The Need or Use Case

Before an AI solution is implemented in a clinical environment, it must first be established what problem the solution is intended to solve. Such a problem might be identified organically or through workshop-based exercises such as workflow and process mapping.

To explore how an AI solution might address the problem, it is recommended that a multi-disciplinary team (MDT) of relevant stakeholders be established as early as possible. The MDT must have a clear goal for the AI solution and an understanding of how it will address the clinical need or problem that has been identified. For example, if radiologists are dealing with low staffing levels and an increased volume of patient images to report, a solution that increases efficiency in reporting would be more beneficial than a solution that would solely improve diagnostic accuracy. As outlined in the ECLAIR guidelines, the relevance of the AI solution to the clinical need or use case must be a central focus during procurement.

Considerations when identifying the need and use case include[13]:

- What is the function of the solution (image quality enhancement, second reader, segmentation, etc.)?
- Is the solution a decision-making tool, decision support tool, or workflow support/automation tool?

- Who are the intended users of the solution (radiologists, radiographers, radiotherapy physicists, etc.)?
- What medical conditions can be treated, diagnosed, and/or monitored with this solution?
- Are there patient selection criteria, indications, contra-indications, or warnings?
- What output is the solution producing (a diagnosis, prognosis, or quantitative data)?
- Is the solution providing new and useful information that was previously unavailable?
- What are the potential benefits of the solution and for whom?

Please see Table 2 and Table 3 below for some sample use cases in diagnostic radiology and radiotherapy.[14]

Table 3: Medical imaging use cases.

Process	Application
Imaging workflow optimization	Prioritization of cases; selection of imaging modality and personalised protocols.
Patient dose estimation in radiological imaging	Tracking of relevant dose metrics; optimization of the imaging procedures by comparing delivered with reference doses.
Image acquisition and reconstruction	CT image acquisition parameter selection; optimal image reconstruction technique for real-time MRI; patient-specific image reconstruction to improve image quality at lower radiation dose; artefact reduction; fully automated data-driven respiratory signal extraction; synthetic image generation; virtual contrast enhanced imaging; iterative image reconstruction.
Image registration and fusion	Mono-modal and multi-modal image registration; deformable registration; 2D-3D image registration
Disease identification and characterisation	CAD (detection/diagnosis) of breast cancer, lung cancer, prostate cancer, coronary artery disease, COVID-19 and other pulmonary diseases.
Risk assessment	Breast cancer risk assessment, breast density estimation, imaging-based risk models
Disease monitoring and response assessment	Monitoring of chronic disease, prediction of response to therapy, assessment of risk of recurrence

Table 4: Radiation oncology use cases.

Process	Application
Treatment decision support	Personalized treatment approach (e.g. proton vs photon); pre-treatment patient risk stratification; pre-treatment prediction of tumour response and RT toxicity; individualised RT dose prescription.
Target localization and segmentation	Automated gross tumour detection and segmentation; involved/elective nodal level segmentation, CTV segmentation considering patient-specific microscopic tumour spread, and resection cavity delineation.
OAR volume segmentation	Automated OAR volume segmentation for many sites (e.g., head and neck, breast, pelvis).
Dose prediction and automated planning	Decision support tools for IMRT/VMAT planning; anatomy based optimal dose prediction; planning process automation to improve efficiency and plan quality; multi-criteria treatment plan optimization.
IGRT and motion management	Imaging-based pre-treatment 2D or 3D target localization; automated fiducials or anatomical structure recognition and alignment; real-time tumour tracking; real-time fiducial-based patient motion monitoring; markerless tracking; motion management in MRIgRT; patient motion prediction.
Treatment plan QA	Patient-specific quality assurance (e.g. prediction of passing rates in IMRT/VMAT pre-treatment patient-specific QA).
Equipment QA	Machine-specific QA; performance monitoring over time; prediction of faults to schedule maintenance and QA, additional testing procedures and maintenance.

3.2 MDT with Involved Parties and Core Affected Staff

Representatives from all relevant stakeholder groups should be invited to form a multi-disciplinary team (MDT). The MDT should be drawn from the core affected staff members in the hospital at the earliest possible stage of the process. This will ensure all affected parties can participate in the procurement process[2] and raise concerns or support from a variety of perspectives.

The team members may include radiologists, radiographers, nurses, and managers from finance/procurement departments, clinical engineers, medical physicists, IT managers, hospital administration staff and the final user. This is not an exhaustive list and will vary depending on the nature of the problem the AI solution is intended to address. In all cases, it is expected that early IT involvement is of particular importance due to the data transfer, privacy and potential demanding hardware infrastructure requirements of some AI solutions.

This team should produce a clear and unambiguous problem statement. The problem(s) should be specific and measurable. Once this statement has been agreed, the team should define key performance indicators (KPIs) that will be used to compare the various solutions available to them. The KPIs should address the problem statement, and these metrics will be used when defining technical specifications that are to be added to the tender.

Some of the benefits to patients, radiologists, referring physicians, the healthcare institution and society are discussed.[13] The MDT should consider such potential benefits and weigh them against the risks associated with AI solutions. The procurement team should request a copy of the vendors risk assessment matrix and risk-benefit analysis that have been submitted in a regulatory technical file to inform this discussion.

3.3 Creating the Tender

It is advisable to perform market research before compiling the tender document, which may include a trial of several solutions using in-house clinical data/workflows. This research will inform the MDT, or team, in charge of procurement, of the number of vendors in the market, the level of performance currently available, considerations for its use on local data/workflows, scope of requirements, licence volume requirements (if charged on a per read/use case), and to review relevant publications.

When conducting market research, vendors may be willing to share datasheets, model cards and system cards. Model cards detail the declared purpose, model architecture,

training methodology, performance metrics, limitations and ethical considerations. Datasheets will summarise how data was collected, cleaned, and may also outline the sources, quality and reliability of the data used for training, testing and validating the model. When reviewing the datasets used, it is important to assess whether the data is unbiased, representative and contains sufficient information to achieve acceptable performance in practice. System cards will detail security, hardware and integration considerations. Access to this information may also be listed as desirable in the technical specifications of the tender document. As some of this information may be proprietary, members of the procurement team may indicate a willingness to sign a Non-Disclosure Agreement before the above documentation is shared.

When creating the tender, mandatory requirements must include compliance with EU laws such as the General Data Protection Regulations (GDPR), the Medical Device Regulations (MDR), and later the Artificial Intelligence Act. Additional functionality that is considered essential (i.e. product can interface with local existing infrastructure and scanners, be vendor neutral, run locally or be cloud-based, or report on decision certainty/explainability) can also be included in the technical specifications.

The KPIs defined by the MDT should also be included in the technical specifications. These KPIs and other metrics should not be restricted to the safe use of the product but should also verify that the solution meets the specified performance level when local data is used. Accuracy, sensitivity, specificity, and AUC ROC scores should be considered for diagnostic tools or time saving estimates should be considered for workflow related products. Interoperability and interpretability of results may also be considered in the technical specifications. Can the results of the model be exported, and in what format, and how are the solutions outputs displayed to the end user?

The tender should also outline the support and maintenance requirements that are expected from the vendor. It may be necessary to have on-site support when the AI solution goes live. During this time, the model may be running in parallel with the standard process while the vendors' staff refine the model based on user feedback or model underperformance when using local data. Information regarding updates to model architecture or changes to the training datasets should be requested and the expected impact on model performance should be communicated. The handling of malfunctions should also be taken into consideration. The team may request if reporting pathways are available in the case of malfunction or adverse events or if the system is being monitored remotely to automatically collect this data. Other considerations include estimates of downtime in case of repair and whether the product can be operated in an offline mode or whether it is possible to continue the service by another means if the product is down.

Training requirements for the intended users, and other staff that may come into contact with the AI solution, should also be outlined in the tender document. Training of IT staff regarding troubleshooting or local hardware requirements should also be detailed. The IT department should also be consulted regarding cyber-security and deployment models (e.g. is the solution cloud based, running on local hardware or in a hybrid model?).

The tender will also assess the pricing model of each vendor. Software is typically sold with a pay-per-use, subscription or once-off fee pricing model depending on the nature of the product. A clear breakdown of maintenance fees, update fees, install costs, etc. is required to assess the full cost of each product over the full life cycle of the AI solution.

3.4 Assessing the Available Options

The primary consideration when assessing the available options are how they score on the various KPIs, and other metrics defined at the tender stage. Meeting these KPIs should ensure that the product is useful and fits with the local needs that have been identified by the MDT.

Some products may be unsuitable if they are required to run on local hardware that does not meet the required minimum level of performance. The cost of additional hardware that can meet these specifications should be considered, if necessary, and a quotation can often be provided by the vendor for the additional hardware costs.

It is important to compare how the tools integrate into the clinical workflow. If the AI products require manual interaction, the extent of the interaction, e.g. the number of clicks / keystrokes for tools that streamline reporting, should be compared. The time for the solution to produce a result after initiation is also a significant consideration. Solutions should ideally be fully integrated into the user's workstation / desktop to facilitate easy access, and each product should be assessed with this in mind. This may require complete integration of the solution with PACS / RIS and the hospital network. The expertise of the IT staff and/or PACS manager will be essential when comparing solutions in this regard. Poor integration can result in a loss of efficiency or an avoidance of the new tool altogether.

Any documentation provided by the manufacturers should be examined and risks or limitations arising from how the product was designed and trained should be considered.

3.5 Engaging with the Vendor

It is important to engage with vendors when evaluating the suitability of a particular AI solution for your health institution. Detailed descriptions of the product, the intended use, documented compliance with EU regulations, hardware and other technical requirements

for local deployment should be sought, if such information is not available in product brochures or company whitepapers. Vendors should also be able to direct interested parties to scientific literature where data related to the product's performance, reliability and effectiveness has been published. Additionally, information regarding model architecture, granular detail about how the model was trained and tested, and fail-safes and safety measures incorporated to mitigate and manage risks should be requested from the vendor. Emphasis should be placed on the quality of the training data and how training was approached. Typical questions that may be asked include the demographics of the population training data, the inclusion of phantom or anthropomorphic data, the use of data augmentation techniques and the total number of independent samples used.

The vendor may also be engaged to provide case studies and examples of successful deployment in other health institutions. Policies and procedures for software updates, including frequency and delivery methods. There should also be an impact assessment of updates on the system's performance. Clarity surrounding the delineation of responsibilities should also be provided by the vendor, as the issue of liability as it relates to AI products may be a particular concern for the health institution, the patient and the clinician or operator using the product.

Since the introduction of the Medical Device Regulations (MDR), software that has been designed and developed for medical use is considered a medical device and therefore must bear a CE mark. The class of the medical device depends on the level of risk associated with the use of the software. Note, if the software influences or drives a medical device it automatically assumes the class of that device.

If the product you are purchasing is CE marked, the vendor will have undergone a conformity assessment. As part of this assessment, the vendor must prepare technical documentation, demonstrate that they have implemented a Quality Management System (QMS) and pass the audits / examinations performed by a Notified Body. Under the MDR, the software must also have undergone a clinical evaluation. Performance must be demonstrated under the normal conditions of the intended use of the product. Undesirable side-effects, and benefit risk-ratio will be determined based on clinical data. This clinical data must be statistically relevant for the intended purpose, intended users, and intended patient population. Copies of the technical documentation and clinical evaluation may be requested from the vendor; however, vendors are not obliged to share this documentation as it may contain proprietary information.

3.6 Product Demonstration

Vendors are often happy to provide demonstrations of their product prior to receiving a commitment to purchase. These demonstrations can take place in the customer's institution or at one of the vendor's partner institutions, where the product is already in use. It is advisable to bring members from the MDT to these sessions to explore the challenges and advantages of the product from a variety of perspectives. Experts employed by the vendor or local users who have successfully implemented the product are typically available for detailed questions at these demonstrations.

Some AI products can be evaluated in the form of a free trial license for a limited period. While the demonstration is taking place, it is important to establish what functionality that is being demonstrated comes as standard or if some elements come with an additional cost.

If you are seeking, or are offered, a demonstration of a product during the tender process, it is recommended to seek the advice of your local procurement department to discuss any policies or rules that are in place before proceeding. All communications should go through your procurement department while the tender is live.

3.7 Assessing Local Infrastructure and Workflow

A critical step in the procurement of AI products is the thorough assessment of the hospital's existing infrastructure and workflows to ensure seamless integration and optimal performance of the new solution. This assessment should begin with a comprehensive evaluation of the current IT infrastructure, including server capacity, storage solutions, network bandwidth, and cybersecurity measures. Depending on whether the AI solution is based locally on servers within the hospital environment, or externally using a cloud service, it is essential to determine whether the existing hardware and software environments can support the computational demands and data handling requirements of the proposed AI system. Additionally, the compatibility of the AI product with existing hospital information systems, such as Electronic Health Records (EHR), Picture Archiving and Communication Systems (PACS), and Radiology Information Systems (RIS), and medical image storage standards such as Digital Imaging and Communications in Medicine (DICOM) must be scrutinised to facilitate smooth data exchange and interoperability.

Understanding the current clinical workflows is equally important to identify potential areas where the AI solution may directly or in-directly impact patient care. Engaging with frontline staff during this phase can provide valuable insights into practical workflow challenges and opportunities for improvement. Moreover, assessing the readiness of staff

to adopt new technologies, including their proficiency with digital tools and openness to change, will inform the training and support strategies required for successful implementation.

Documenting the findings from the infrastructure and workflow assessment will aid the MDT in refining the technical specifications and identifying any necessary upgrades or modifications needed to support the AI solution. By evaluating the local infrastructure and workflows prior to defining specifications, the procurement team increase the likelihood of successful integration and ensures that the AI product delivers its intended benefits within the hospital setting.

3.8 Commissioning

A crucial phase in the procurement process is the assessment of AI products using the hospital's own data. This ensures that the selected AI solution performs reliably and effectively within the specific context and diversity of the patient population served by the institution, which may have both clear and unexpected differences from the data used for training the AI model. Robust testing with local data helps identify whether the AI product maintains high performance across a wide range of patient types and clinical scenarios, thereby ensuring its suitability and safety for actual clinical use. AI models often experience a decline in performance when deployed outside the original training institution or domain due to variations in data characteristics and clinical practices. It is imperative to assess how well the AI product generalizes to the local environment by comparing its performance on local data against the results reported by the vendor.

3.8.1 Model Cards and Technical Documentation

Each AI solution should be accompanied by a model card, which is essentially a specification sheet that provides comprehensive details about the model, including its architecture, training process, evaluation methods, AI techniques employed, and the datasets used for training. Reviewing the model card is essential as the first step in the assessment process to ensure that the AI model aligns with clinical requirements and has been trained on data comparable to the intended clinical use case. The model card can reveal potential limitations or edge cases associated with the AI technique used; for example, a UNET architecture for segmentation may have specific constraints or performance characteristics that need to be considered. By evaluating the model card, the procurement team can better understand the capabilities and constraints of each AI solution, facilitating a more informed decision-making process to determine which AI solution is the best suited before their clinical environment.

3.8.2 Robust Testing Across Diverse Patient Types

To validate the AI product's efficacy, it is essential to conduct comprehensive testing that encompasses the diversity of the hospital's patient population. This involves evaluating the AI system on various patient demographics, including age groups, genders, and those with specific medical conditions or implants. For instance, in radiological applications, the presence of metal implants on CT images can pose challenges for AI algorithms. Identifying such edge cases where the AI may underperform is vital to understanding the limitations of the solution and developing strategies to mitigate these issues.

3.8.3 Identifying and Addressing Edge Cases

Edge cases, such as atypical imaging scenarios or rare pathologies, must be thoroughly examined to determine the AI product's robustness. For example, if an AI model was trained predominantly on CT scans without contrast, its performance may decline when applied to scans with contrast. In such instances, it may be necessary to adjust clinical workflows or standard operating procedures to align the data acquisition processes with the conditions under which the AI model was trained. This alignment ensures that the data fed into the AI system closely matches its training data, which helps maintain high levels of performance and reliability.

3.8.4 Evaluation Metrics and Performance Indicators

Selecting appropriate evaluation metrics is fundamental to accurately assess the AI product's performance. For applications such as AI segmentation in radiotherapy, metrics like DICE coefficient, Surface Dice, and Hausdorff distance are valuable as they correlate with segmentation correction time. However, it is important to recognise that these metrics may not directly correlate with dosimetric impact, which is critical for radiotherapy treatment planning. Additionally, establishing ground truth can be challenging due to inherent intra- and inter-human variability in annotations. To address this, it is beneficial to measure intra- and inter-user variability, as well as sensitivity and specificity metrics, prior to purchasing the AI tool. This comprehensive evaluation provides a clearer understanding of the AI system's performance relative to the current standard of practice.

3.8.5 Impact of Machine Parameters on AI Performance

The performance of AI models can be significantly influenced by machine parameters and imaging protocols. For example, variations in kilovoltage (kV) and milliamperage (mA) settings during imaging can affect the quality and characteristics of the acquired data. It is essential to assess how the AI model performs across the range of protocols and custom scan variables routinely used in the radiology unit. Understanding the model's sensitivity to these parameters ensures that it can reliably handle the variability inherent in clinical imaging practices without compromising performance.

By meticulously assessing the AI product with local data, the procurement team can ensure that the selected solution not only meets the technical and clinical requirements but also integrates seamlessly into the hospital's unique operational environment. This thorough evaluation mitigates risks, enhances the reliability of the AI system, and ultimately contributes to improved patient outcomes and operational efficiency within the healthcare institution.

3.9 QA Schedule

Quality Assurance is required to assess whether the AI application is operating as expected over time.[2] As input data changes, with the introduction of new equipment or patient populations, the performance of an AI solution must be monitored.

A set of KPIs should be established and tracked allowing institutions to identify and communicate performance issues back to the vendor if necessary. KPIs will vary significantly depending on the intended use of the AI product but general rules for KPIs apply, i.e. KPIs should be interpretable, measurable and relevant to the task being tracked. Tolerance levels for these KPIs should also be established.

The success of the QA schedule will depend on a clear allocation of roles and responsibilities. Who will extract the data? Who will perform the analysis? How often are these results to be analysed?

4. Assessing the AI Tool

4.1 Datasets

A dataset is a collection of structured data used to train, validate and test AI models. The features and format of the dataset will be determined by the intended use of the trained model. AI models with similar goals, e.g. classifying whether a patient has COVID-19 or not, may choose to approach this task differently. One approach may rely only on the pixel data, whereas another may make use of multi-modal data.

Datasets are split into Training, Validation and Testing subsets. The training dataset will be the largest of the three and is used, as the name suggests, to train the model. Size requirements for this dataset have only grown with the maturation of the field, with modern neural networks demanding huge datasets that scales further with the complexity of the task. When assessing a vendor's AI model, it is important to investigate the characteristics of the training dataset used. If the dataset is too small, the model may perform significantly worse in a live implementation. The dataset may be sufficient size, but the data used may not adequately represent the local population and result in inferior performance. The validation dataset is often, but not always, kept separate from the training dataset and is used to more finely-tune the performance of a model. Knowledge of the splits between training and validation is of less importance to the Medical Physicist.

The testing dataset is a subsection of the dataset kept entirely separate from the training and validation sets. It is used to see gauge the performance of the model in a "real world" scenario and to ensure the model has not been over-fitted during the training stage. If you have direct model access, you can create your own test dataset and personally assess the performance of a model using data that reflects the demographic of your hospital. Otherwise, the vendor should be transparent about the breakdown of their test dataset as it is from this that their accuracy and precision scores are obtained.

4.2 Performance Assessment

Adequate Performance assessments in the context of AI systems in healthcare must relate to the intended use of tool, relevant performance metrics and reproducibility.

These key areas should be addressed during rigorous testing conducted by the manufacturer at the development stage. However, it is equally important to continually evaluate how the AI model performs when interacting with input data originating from a specific clinical institute to ensure its effectiveness and reliability in practical applications.

The use of error estimation, assessment of statistical significance and power calculations are encouraged and should be an integral part of any study design involving AI deployment/assessment.

4.2.1 Intended Use

The intended use of the system must match the clinical environment it is deployed in and is determined by the patient population, the image acquisition device, the stage of diagnostic intervention, and the diagnostic category.[31] These factors can contribute to lower-than-expected AI performance and are further detailed below

1. Patient population represented by the data used to train/develop should match the intended population of clinical use.
2. The AI model must be developed and tested on data from multiple vendors, particularly in cases where AI systems are intended to be vendor neutral.
3. The intended use depends on the patient care stage that requires the diagnostic intervention.
4. The diagnostic category of the data should match the clinical task e.g. screening, detection, staging, treatment assessment or follow-up.

4.2.2 Performance Metrics

The most appropriate performance metric(s) will depend on the task and the reference standard. In many cases, employing multiple performance metrics is both appropriate and desirable. Performance metrics that are adopted should be applicable to the intended use of the AI Model. These metrics should be applied to the model while it is being trained, but they can also be applied to closed (non-continual learning) models when variations of input data are being analysed.

4.2.3 Specific Applications

Object Detection when applied to medical imaging is defined as the combination of localization (identifying the object location) and some level of classification (identifying a broad object category e.g. “Benign or Malignancy status”).[16]

The most common metrics for assessment in relation to detection include area under the ROC curve[21], Accuracy, sensitivity (recall), specificity (precision)[22], balanced accuracy (mean of the sensitivity and specificity), Youden index[23], and the prevalence dependent factors positive predictive value (or precision), negative predictive value[24], and F1 score.[23, 18-20]

Image Segmentation in image processing is the partition of a clinical image into multiple parts.[25] Classical segmentation methods can include examples such as Manual tracing, Region Based, Edge Based, Threshold based, Feature-based clustering and Coregistration

Atlas based approaches.[26] It is recommended to compare segments/contours produced by an AI model with those from the above classical approaches. Metrics of comparison are Dice coefficient, Jaccard index, and Hausdorff distance.

Regression analysis examines the ability of one or more factors, known as independent variables, to predict a patient's status in regard to the target or dependent variable. Independent and dependent variables may be continuous (taking a wide range of values) or binary (dichotomous, yielding yes-or-no results). Regression models can be used to construct clinical prediction rules that help to guide clinical decisions.[27] Metrics for comparisons with classical regression methods are mean squared error and coefficient of determination.[28]

4.2.4 Reporting Results

Error estimates must be included when the assessments carried out yield quantified results, such as standard deviations or 95% confidence intervals. Error estimates describe the uncertainty in the reported performance metric values and give insight into the sufficiency of the training sample size, soundness of the training/testing approach and generalisability.[17]

The evaluation of a CAD-AI algorithm includes both benchmarking algorithm performance during development and assessing the added value to the end user provided by the algorithm in improving clinical decision making.

4.2.5 Reproducibility

It is crucial to clearly define the conditions under which the results of a CAD-AI system are reproducible. AAPM Report No. 273[17] outlines three types of reproducibility in detail: technical, statistical, and inferential reproducibility. For AI deployment in healthcare, the first two are particularly relevant.

It is recommended to continuously monitor the variation in performance and results of an AI model using both the same input data and variations of input data from a clinically relevant population. This is especially critical when model updates are implemented, or scan sequence parameters are adjusted, as these changes can fundamentally alter the input data.

4.3 Retrospective Testing

The gold standard for assessing the accuracy and precision of an AI model for carrying out its intended task requires direct access to the model itself and uses a testing dataset. Having unfettered access to the model may become increasingly unlikely as the field matures as the model is a proprietary product, however, if the AI model is being developed

in-house or with close collaboration with a research partner this level of access can be expected. Vendors may provide a demo environment, but the access pipeline will be closer to the live implementation and limits the size of the testing dataset that can be effectively tested.

With full access to the model, a test dataset that is representative of the local population can be created and used to assess model performance. The procedure for carrying this out is similar to how the model's performance would be determined in production after model training and validation. Briefly, this involves running the dataset through the model and recording the true positives, true negatives, false positives and false negatives to determine the precision and accuracy of the model (see section 4.2).

If access to the model is limited through a demonstration environment prior to launch in a clinical setting, the physicist will be limited to manually sending test data through to the model using the vendor supplied interface. This will likely reduce the practical sample size used. An example process for obtaining these results for a simple dataset where the ground truth is known explicitly could be:

1. Dataset: A CSV file containing path to DICOM file in the first column and some classifier in the second column, e.g. "PNEUMOTHORAX PRESENT".
2. Model: Available as a function that takes the DICOM path, or object, and returns a prediction, e.g. "PNEUMOTHOROAX PRESENT: TRUE/FALSE".
3. Process: Loop through the dataset, obtaining a prediction for every file.
4. Analyse: Compare the predicted results to the known ground truth. Generate ROC curves and present results on accuracy. This is the performance of the model based on local data.

For many applications, the predictions and ground truth are not absolute values but a confidence rating or bounding region. In these cases, additional work is required to determine the agreement with clinicians by calculating the Dice coefficient or Hausdorff distance outlined in section 4.2.3.

Care should be taken in choosing or curating a dataset for retrospective testing, using patient data from a local hospital raises ethical and liability concerns if the AI model has a finding that may initially have been missed, see section 6, Ethics and Privacy in AI.

4.4 Prospective Validation & Monitoring

Implementing AI tools into clinical practice is a shared responsibility between manufacturers and end-users[37] that should mirror the QA programs required to install medical imaging devices.[38] The programs should include comprehensive acceptance

testing (AT) and continued, periodic quality control (QC) procedures. End-user training and a proper trial period with the local patient population should be required to ensure an understanding of the intended use and limitations of the AI tools before the AI recommendation may influence clinical decisions.[32]

AAPM group, TG 416, titled "Quality Assurance and User Training of CAD- AI Tools in Clinical Practice." has been established. The aim of this group is to develop best practices and guidelines for acceptance testing, quality assurance, and user training of CAD-AI systems and tools to facilitate their translation to the clinic and to ensure their reliability and reproducibility. At the time of writing this document, no guidance has been produced by this task group, but the reader is encouraged to refer to the task group website for updated guidance.

4.4.1 Quality Assurance Program

A QA program focused on AI should aim to address the following key areas.

1. Acceptance testing should involve more rigorous testing compared to routine QC and be used to determine Baseline Performance.
2. Routine Quality Control to ensure early detection of changes in performance.
    a. Homemade tools and manufacturer provided tools are encouraged for continuous monitoring where possible and where they comply with the MDR 2017.
3. Re-validation to verify performance after changes to the workflow (data/processes/software update) that could impact AI tool output.

The below identifies components of a comprehensive QA program.[32] An example of a QA workflow focused on auto-segmentation in radiotherapy can be found in Figure 4 below.

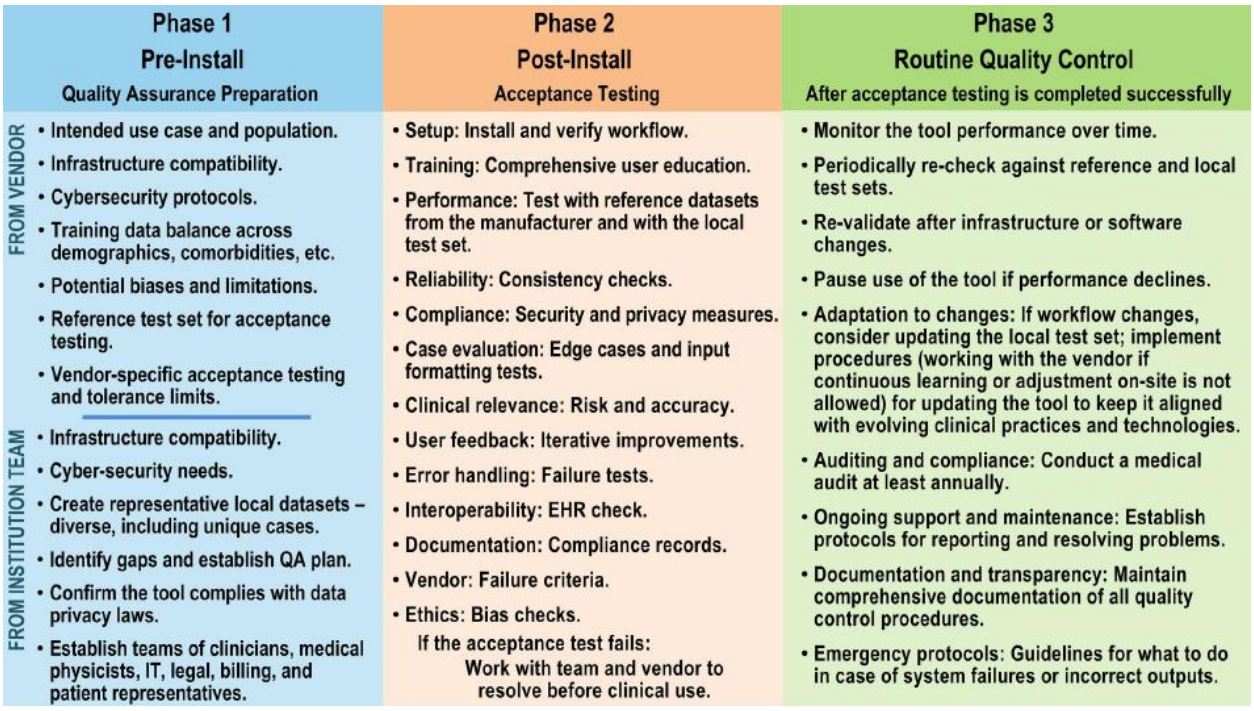

Figure 4: Suggested QA workflow for a new AI product.[32]

4.4.2 Live Testing

Live testing refers to the periodic QA program and general performance monitoring that should be put in place to detect when significant changes occur in the AI model output. Routine QA should be implemented (preferably by medical physicists in conjunction with routine QA testing of related medical imaging systems) to assess the following:

1. How variations in the imaging or data collection chain may affect the performance of the CAD-AI system.[33]
2. Performance drift associated with announced software updates provided by the manufacturer.

It should be noted that when referring to the AI model continual learning models are not considered at the time of writing this article, no such models are approved for use in healthcare.[36] For information on performance monitoring for continual learning systems, refer to AAPM Report no. 273.

Testing regimes should ideally include the use of curated datasets (see section 4.1 & 4.3) to assess changes in performance, however as discussed earlier this isn't always practical to implement.

Clinical sites & manufacturers should develop tools to track performance levels of specific indices over time; many vendors now provide dashboards to track basic features of

performance. The exact features offered will vary from vendor to vendor, but filtering options for date ranges, confidence thresholds and total studies processed are often available.

The tolerance limits and corrective actions for any observed deviations should be established based on the CAD-AI application. The risk associated with any deviation will vary significantly for different diseases and tasks performed by the CAD-AI system. For example, if the system is an autonomous CAD-AI detection or decision tool for triaging or rule-out, immediate corrective actions are recommended, while tools designed only to provide a second opinion or supplementary information may be less urgent.

4.4.3 Clinical Reader Performance Assessments

Clinical reader performance assessments can be useful to estimate the clinical impact of a CAD-AI algorithm and would be ideally carried out as part of the acceptance testing stage. Clinicians' interactions with the AI tool should be reviewed to identify potential issues such as automation bias.

A common approach for assessing clinical performance is through a controlled reader study (either retrospective or prospective), directly comparing the performance of a human reader without and with output from the CAD-AI system. A disadvantage of this approach is that the estimated performances are unlikely to match those in the true clinical setting because of differences in the cases, physicians, and reading process. It is important to realize that both the population of patients undergoing the examination (cases) and the population of physicians interpreting the data (readers) are sources of substantial variability in clinical reader studies. Specialized statistical and methodological tools are needed for these analyses.

Prospective evaluation of CAD-AI can be carried out using randomised controlled trials and observational studies. These will help evaluate the AIs impact on workflow, efficiency, clinical performance, cost-effectiveness and patient outcomes.[33] Adopting these evaluation strategies is encouraged and facilitated by many companies introducing AI packages to healthcare settings. In the first instance it is recommended to assess the interaction of the institutes specific input data with the package. It may also be recommended to periodically evaluate the package over its lifetime in this fashion particularly when significant software updates are carried. Suppliers and manufacturers may have guidance for controlled randomised trials with suggested timeframes ranging from 1 to 12 months. Parallel testing schemes are also used where testing is carried out alongside clinical rollout.

Key elements for a performance assessment have been described in previous sections, and these can be implemented into the acceptance testing and routine QC stages mentioned above. The frequency of monitoring should be aligned with the risk level, existing regulations and operational experience with the AI tool.

4.4.4 Risk Categorisation

High risk tools are those involved in triage or medical diagnosis and should require annual assessment. Low risk tools are categorised outside of these areas and require testing less frequently. Testing should scrutinize for fairness, potential biases and error rates.

Testing may also be warranted for high-risk tools after significant changes in clinical workflow, technical updates, or unusual errors noticed by clinical users. The goal is to balance patient safety with operational efficiency.

4.4.5 Performance Drift

To assess the impact of detected performance drift for a particular AI model it may be useful to correlate the drift to a diagnostic efficacy level. Diagnostic efficacy is the ability to produce expected results from a diagnostic procedure. Sensitivity and specificity indicate the probability of a positive or negative test result, given that the patient either has or does not have, respectively, a specified disease, and are the most reported measures of test efficacy.[34]

Diagnostic efficacy levels are adopted here (table 5 below) to assess the potential impact of performance drift of the AI tool. These are categorised into a hierarchy table where each level relates to the severity of potential impact. 1 being low severity, and 6 being high severity.[35]

Table 5: Level 1 denotes low impact severity and incrementally increases to level 6 which is the maximum.

Level 1	Technical quality of the images.
Level 2	Diagnostic accuracy, sensitivity, and specificity associated with interpretation of the images.
Level 3	Information produces change in the referring physician's diagnostic thinking.
Level 4	Information affects the patient management plan.
Level 5	Information effects patient outcomes.
Level 6	Effects societal costs and compromises benefits of the diagnostic imaging technology.

4.5 Retesting

AI models used in the healthcare setting are not allowed to "learn" over time in a live environment[36] however, there are several external factors that can influence the model's performance over time.

1. Vendor updates the model
2. Change to modality
3. Change in demographic

The overall performance of the department (radiology/radiotherapy + AI model) may also change over time due to improved cohesion with the model (better performance) or increasing over-reliance (worse performance) on the model. Both cases are not within the scope of this section and need to be assessed on a case-by-case basis by having close involvement with the relevant clinical team.

Vendors consistently offer updates to software packages, during scheduled maintenance or in response to a user-detected fault. It will become vital to be aware of whether a software update includes an update to the model or just the surrounding interface. While an update to the interface *might* have an impact on the output if there is any change in how the data is input to the model, it is unlikely compared to the guaranteed change if the model is updated. The extent of the performance change is intended to be small but may have unexpected downstream effects in the user's local environment. The vendor may have improved the model for their dataset which may or may not be representative of the local demographic. A cited positive improvement from the vendor may not be reflected locally.

A change in the modality, either by the introduction of new imaging systems, or a decrease in image quality due to degradation of the equipment over time can impact the efficacy of an AI model, this can be known as "input drift". A new imaging system may feature novel post-processing or image formation that is sufficiently different from the images the model was trained on, lowering the efficacy of the model (e.g. synthetic 2D images in mammography). A degradation of the imaging system, for example, by a change to the texture or intensity of noise can impact model performance before it significantly impacts subjective image quality. It is recommended to closely monitor the performance of AI models on images produced by new systems to ensure there has been no unexpected change in performance. Furthermore, a noted deterioration in image quality found during annual quality assurance testing should also consider the potential impact on any AI models that use those images.

In a rarer scenario, the patient demographic served by your institution may change over time. On a long timescale, the average age of the patient is expected to increase. Diseases may change over time in how they manifest. Disasters, pandemics, or other unforeseen incidents that impact a substantial proportion of the population may result in a cohort presenting prominently to your institution who historically would rarely have been seen.

Section 4.4 outlines several key-moments to re-assess the CAD-AI, either as part of routine QC or performance drift noted by close observation of key performance indicators, however it is important to note the discussed factors that are external to the AI system that can also require a re-assessment of CAD-AI performance. Ongoing monitoring of performance is essential to flag to the MPE when a system has dropped in performance and such tools that facilitate this (either through an AI orchestrator, built-in, or home-grown) should be prioritised for safe implementation of AI tools.

5. Sample Case Study

5.1 Third Party Tool Assessment

AI Segmentation in Radiation Oncology: A Case Study at St Luke's Radiation Oncology Network (SLRON)

This implementation took place between 2022 and 2024 at SLRON, Dublin Ireland. The AI solution decided on was MVision AI Auto-contouring which featured a cloud-hosted "zero-click" pipeline.

5.1.1 Summary

St Luke's Radiation Oncology Network (SLRON) in Dublin, Ireland, provides radiotherapy treatment for over 5,500 external radiotherapy patients annually. In 2022, SLRON started a project aimed at introducing AI-driven segmentation to streamline the contouring process in radiation oncology, ultimately selecting MVision's AI auto-contouring as a preferred solution. The goal was to increase efficiency, reduce inter-observer variability, and improve the quality of service offered to our patients. This case study explores the project journey from initial proposal and business case justification, through risk management, implementation, and post-implementation monitoring and evaluation.

5.1.2 Project Proposal and Justification

In late 2022, SLRON recognised a growing need to address the significant workload and variability inherent in manual contouring processes. After a careful assessment, a proposal was developed outlining the justification for purchasing an AI-based segmentation technology to both standardise contouring procedures, and streamline the contouring process, thereby enhancing efficiency and clinical consistency. The proposal emphasised expected reductions in manual labour, variability among clinicians, and improvements in the workflow based on published literature. This was particularly important given the increasing complexity of radiotherapy techniques and the rising patient numbers, both requiring efficient solutions without compromising clinical quality.

5.1.3 Business Case and Strategic Alignment

Following a review of several commercial vendors, a business case for MVision's AI segmentation solution was written considering multiple factors including clinical efficiency gains, accuracy, availability of models to cover all clinical cohorts, and economic viability. The business case included both evidence from literature, and in-house pilot results of the solution in action. The pilot was run to assess contour accuracy and timesaving at an MDT level across each clinical cohort. The initial pilot analysis detailed potential time savings in contouring, estimated reductions in inter-observer

variability, and subsequent enhancement of workflow productivity. Financial considerations evaluated cost-benefit scenarios against current manual approaches, projecting significant annual savings over the operational life of the technology. Strategic alignment with national healthcare guidelines and the overarching mission of patient-centred care were integral components of the justification.

5.1.4 Risk Assessment and Management

A Failure Mode and Effect Analysis (FMEA) was conducted in May 2023 to systematically identify, quantify, and mitigate potential risks associated with implementing MVision AI segmentation. Risks were categorised into clinical, technical, operational, and data security domains. Key risks included the potential for complacency among clinicians relying too heavily on AI, inaccuracies in AI-generated contours, data security concerns, and integration challenges with existing IT infrastructure. Mitigation strategies involved rigorous staff training, structured clinical validation procedures, continuous monitoring, and cybersecurity measures to ensure patient data confidentiality.

5.1.5 Implementation and Change Management

Implementation commenced in late 2023, following a series of training and education sessions designed to facilitate seamless integration into existing workflows. The comprehensive training and change management program was designed to equip staff with necessary skills, manage expectations, and maintain clinical oversight during the transition. We decided to roll out the AI segmentation tool on one date, to ensure there was no confusion over which patient structures were AI generated or not, focusing on a single unified workflow change. This roll-out approach minimised disruption, ensure clarity on workflow processes, and helped encourage broad staff acceptance and confidence in using AI-assisted contouring. It also allowed staff to “know the model” as they were only presented with AI generated contours to amend, not a mix of AI generated and human generated contours.

5.1.6 Post-implementation Monitoring and Audit

Upon completion of the phased implementation, a rigorous monitoring and auditing process was established to evaluate the real-world performance and impacts of MVision AI. This was overseen by an MDT comprised AI implementation group. Any feedback received, or issues were fed back to the implementation group to act or disseminate important information about expectations of model performance and limitations. A post-implementation monitoring process was initiated. Metrics collected included the accuracy of AI-generated contours compared to clinician-generated contours pre- vs. post-implementation to detect bias/complacency, and time taken at both a task level and workflow level per clinical cohort.

5.1.7 Outcomes and Impact Assessment

Initial findings from the AI-segmentation impact study (October 2024) indicated that MVision:

1. Reduced contouring time by up to 70 % for some clinical cohorts
2. Decreased clinician variability in 68 % of structures
3. Enabled roughly 10 % of treatment plans to be finalised two days earlier than before implementation

Clinicians reported high levels of satisfaction with AI-assisted contours; however, monitoring also revealed that about 50 % of structures exhibited a systematic shift post-implementation, underscoring the need for ongoing quality checks and vigilance against human-factor issues such as bias and complacency.

5.1.8 Multidisciplinary Engagement and Research Culture

From the outset, we considered it essential to embed the entire multidisciplinary team (MDT) in the development process, not merely as end-users but as co-investigators. By actively encouraging small, focused research and development projects, we included physicists, radiation therapists (RTTs) and radiation oncologists to test the model's strengths, expose its limitations and explore novel clinical applications. Projects examined site-specific contour accuracy, quantified task-level time savings, investigated changes in intra- and inter-observer variation, and investigated whether daily AI-generated contours could act as early-warning signals for anatomical adaptations. This collaborative approach helped demystify the technology, aligned expectations of model performance and limitations across professional groups and resulted in an immediate academic dividend: several abstracts were accepted at national and international conferences, helping us shift towards a shared responsibility to "know the model" and remain vigilant about bias and complacency.

5.1.9 Conclusion

SLRON's successful implementation of AI segmentation provides an example for other institutions considering similar AI implementations. Through careful planning, robust justification, proactive risk management, structured implementation, management of expectations and ongoing evaluation and monitoring, SLRON demonstrates that integrating AI into radiotherapy workflows can significantly enhance clinical practice, operational efficiency, and patient care quality, provided it is done thoughtfully and with competent human oversight at all stages.

6. Ethics & Privacy with AI

Ethics associated with the use of AI tools is a significant area of interest across many different industries. A Pub-Med search for articles containing the words “AI” and “Ethics” showed that over 6,000 articles have been published on the subject this year alone. In this section we aim to provide some guidance on ethical implementation of AI in a healthcare setting. However, it is by no means an exhaustive list and the dust is still settling on the correct path to navigate this issue. It is imperative to state that the IAPM takes no legal responsibility for any site policies or procedures pertaining to the use or validation of AI tools. This section only aims to offer guidance gained from practical experiences and study of the EU AI Act and GDPR. Over the course of this section, we aim to answer some of the key questions the user may wish to be aware of before engaging in implementation of an AI tool.

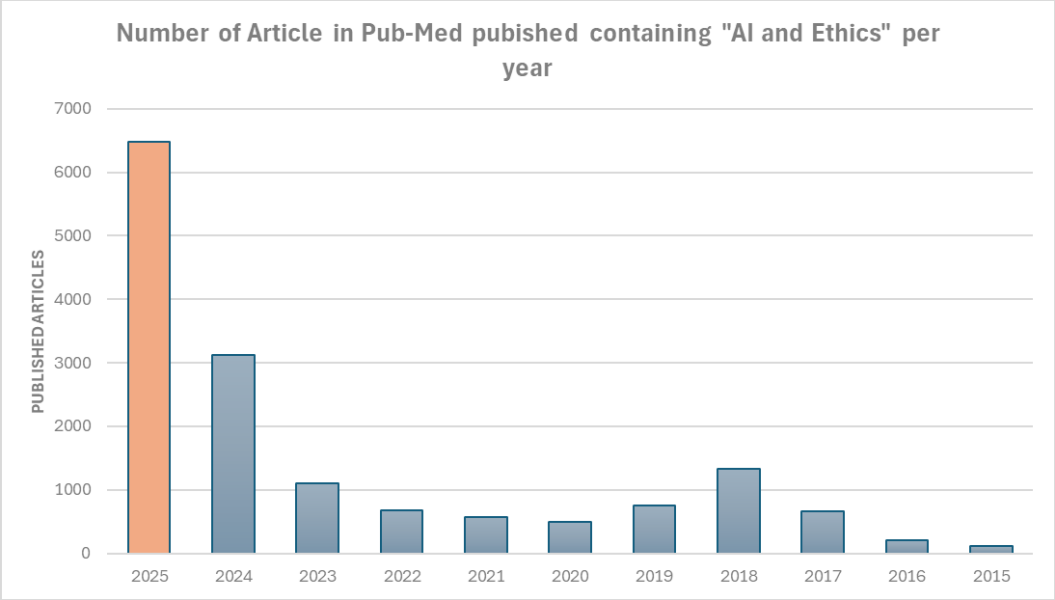

Figure 5: Number of PubMed Articles Published in the last 10 years associated with AI and Ethics[38]

6.1 Do patients need to be consented for their data to be used for on-site evaluation of AI models?

It is our understanding that patient imaging data can be used for on-site evaluation of AI tools without requiring additional explicit consent from the patient, once the data is fully compliant with GDPR regulations and for the direct benefit of patients, local service evaluation, audit, quality improvement or validation of tools used in direct care. However, if biopsy data is used to assist with AI validation, further patient consent may be required in this case if the scope of investigation is closer to research. Contact your hospitals Research Ethics Committee for feedback for your specific use-case. Fully anonymising the

data may exempt this consideration but raises further challenges in pairing the data to the appropriate imaging data or AI output. The multi-disciplinary team should consult with site management and the DPO to ensure all necessary stakeholders are fully aware and informed of the process being undertaken.[39]

6.2 What should we do regarding incidental findings found during retrospective studies?

Guidance is limited as to the procedures to be followed regarding incidental findings of AI tools. It is recommended that the MDT assessing the tool and establish procedures to follow if or when an incidental finding is identified. This should be completed before the assessment of the tool begins and is agreed upon by all parties and is compliant with hospital policies. Consider the impact of a false positive on the patient if recalled for additional imaging or treatment after previously been given the all-clear. Where AI is being used prospectively, the data should always be reviewed by a clinician to support their own assessment and not as the main assessment tool.

6.3 Who needs to know that AI is being used in the healthcare setting?

The deployer (the hospital) needs to ensure the workers who interact with, or rely on, the tool or work processes where the tool is implemented are aware of the AI tool and competent in its use and limitations. The obligations of a deployer are outlined in the EU AI Act in Article 26, however hospitals are not included in Annex III and so do not need to inform individual patients of the usage of AI tools.[40]

6.4 How should we approach the procurement process with these AI vendors?

For the procurement of AI tools for use in healthcare the standard HSE tender process should be followed. For hospitals directly under the control of the HSE, these tools may be procured at a national level without the involvement of local sites. For other not directly under the control of the HSE, the executive has provided guidance on the procurement pathway with their policy on procurement which can be found on their website.[41]

Additional literature has also been released by the Irish government in recent years in relation to public procurement. While this information is delivered on a much broader sense, there is useful information that would be beneficial to be aware of when designing a tender document for healthcare purposes.[42]

6.5 Who is responsible for incorrect diagnosis once AI tools are involved?

It is our belief that the clinician is always responsible for the patient's diagnosis and treatment. AI tools designed to assist the clinicians only and not be allowed to make sole decisions. Establishing liability for faults in AI tools can be complex, but it is possible if comprehensive ongoing evaluation of the tool is implemented and that staff are appropriately trained and understand the tool's limitations.

7. References

1. Brady A. P. et al, Developing, purchasing, implementing and monitoring AI tools in radiology. Radiology: Artificial Intelligence 2024;6(1):e230513 https://doi.org/10.1148/ryai.230513
2. Bosmans H, Zanca F, Gelaude F, Procurement, commissioning and QA of AI based solutions: An MPE's perspective on introducing AI in clinical practice. Physica Medica 83(2021) 257-263
3. Article 3: Definitions | EU Artificial Intelligence Act [Internet]. Available from: https://artificialintelligenceact.eu/article/3/
4. Kizito Nyuytiymbiy. Parameters and Hyperparameters in Machine Learning and Deep Learning | Towards Data Science [Internet]. Towards Data Science. 2020. Available from: https://towardsdatascience.com/parameters-and-hyperparameters-aa609601a9ac/
5. Brownlee J. What is the Difference Between a Parameter and a Hyperparameter? [Internet]. Machine Learning Mastery. 2017. Available from: https://machinelearningmastery.com/difference-between-a-parameter-and-a-hyperparameter/
6. What makes AI algorithms different from traditional computer algorithms? [Internet]. www.linkedin.com. Available from: https://www.linkedin.com/advice/3/what-makes-ai-algorithms-different-from-omp7f
7. https://www.vde.com/en/press/press-releases/vde-dgbmt-marktzugang-lernende-ki-systeme-medizin
8. Cestonaro C, Delicati A, Marcante B, Caenazzo L, Tozzo P. Defining medical liability when artificial intelligence is applied on diagnostic algorithms: a systematic review. Frontiers in Medicine. 2023 Nov 27;10(1305756).
9. NHS. NHS England» Decision support tools [Internet]. www.england.nhs.uk. 2024. Available from: https://www.england.nhs.uk/personalisedcare/shared-decision-making/decision-support-tools/
10. Soori M, Karimi F, Dastres R, Arezoo B. AI-Based Decision Support Systems in Industry 4.0, A Review. Journal of Economy and Technology. 2024 Aug 1;
11. Article 16: Obligations of Providers of High-Risk AI Systems | EU Artificial Intelligence Act [Internet]. Artificialintelligenceact.eu. 2014. Available from: https://artificialintelligenceact.eu/article/16/
12. IPPOSI launches Citizens' Jury report on AI in Healthcare | IPPOSI [Internet]. Ipposi.ie. 2024 [cited 2025 Feb 28]. Available from: https://ipposi.ie/aicitizensjury/
13. Omoumi P, Ducarouge A, Tournier A, Harvey H, Kahn CE Jr, Louvet-de Verchère F, Pinto Dos Santos D, Kober T, Richiardi J. To buy or not to buy-evaluating commercial AI solutions in radiology (the ECLAIR guidelines). Eur Radiol. 2021 Jun;31(6):3786-

3796. doi: 10.1007/s00330-020-07684-x. Epub 2021 Mar 5. PMID: 33666696; PMCID: PMC8128726.

14. International Atomic Energy Agency, Artificial Intelligence in Medical Physics, Training Course Series No.83, IAEA, Vienna (2023)
15. Yousefirizi, F., Decazes, P., Amyar, A., Ruan, S., Saboury, B., & Rahmim, A. (2022). AI-Based Detection, Classification and Prediction/Prognosis in Medical Imaging: Towards Radiophenomics. In PET Clinics (Vol. 17, Issue 1).
16. Hadjiiski, L., Cha, K., Chan, H. P., Drukker, K., Morra, L., Näppi, J. J., Sahiner, B., Yoshida, H., Chen, Q., Deserno, T. M., Greenspan, H., Huisman, H., Huo, Z., Mazurchuk, R., Petrick, N., Regge, D., Samala, R., Summers, R. M., Suzuki, K., … Armato, S. G. (2023). AAPM task group report 273: Recommendations on best practices for AI and machine learning for computer-aided diagnosis in medical imaging. Medical Physics, 50(2), e1–e24. https://doi.org/10.1002/mp.16188
17. Petrick, N., Sahiner, B., Iii, S. G. A., Bert, A., Freedman, M. T., Fryd, D., Gur, D., Hadjiiski, L., Huo, Z., Jiang, Y., Raykar, V., Samuelson, F., Summers, R. M., Tourassi, G., Yoshida, H., Zheng, B., Zhou, C., & Chan, H.-P. (2013). Evaluation of computer-aided detection and diagnosis systems. Med. Phys, 40(8), 87001–87002. https://doi.org/10.1118/1.4816310
18. Doi, K., MacMahon, H., Katsuragawa, S., Nishikawa, R. M., & Jiang, Y. (1999). Computer-aided diagnosis in radiology: potential and pitfalls. European Journal of Radiology, 31(2), 97–109. https://doi.org/10.1016/S0720-048X(99)00016-9
19. Gallas, B. D., Chan, H.-P., D'orsi, C. J., Dodd, L. E., Giger, M. L., Gur, D., Krupinski, E. A., Metz, C. E., Myers, K. J., Obuchowski, N. A., Sahiner, B., Toledano, A. Y., & Zuley, M. L. (2012). Evaluating Imaging and Computer-aided Detection and Diagnosis Devices at the FDA. 463–477. https://doi.org/10.1016/j.acra.2011.12.016
20. Melo, F. (2013). Area under the ROC Curve. In: Dubitzky, W., Wolkenhauer, O., Cho, KH., Yokota, H. (eds) Encyclopedia of Systems Biology. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-9863-7_209
21. Baratloo, A., Hosseini, M., Negida, A., & Ashal, G. el. (2015). Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity. Emergency, 3(2), 48. https://pmc.ncbi.nlm.nih.gov/articles/PMC4614595/
22. Berrar, D. (2019). Performance Measures for Binary Classification. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, 1–3, 546–560. https://doi.org/10.1016/B978-0-12-809633-8.20351-8
23. Shreffler J, Huecker MR. Diagnostic Testing Accuracy: Sensitivity, Specificity, Predictive Values and Likelihood Ratios. [Updated 2023 Mar 6]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2025 Jan. https://www.ncbi.nlm.nih.gov/books/NBK557491/
24. Zhang, H., & Li, D. (2014). Applications of computer vision techniques to cotton foreign matter inspection: A review. Computers and Electronics in Agriculture, 109, 59–70. https://doi.org/10.1016/J.COMPAG.2014.09.004

25. Anand, R., Mishra, R. K., & Khan, R. (2022). Plant diseases detection using artificial intelligence. Application of Machine Learning in Agriculture, 173–190. https://doi.org/10.1016/B978-0-323-90550-3.00007-2
26. Guyatt, G., Walter, S., Shannon, H., Cook, D., Jaeschke, R., & Heddle, N. (1995). Basic statistics for clinicians: 4. Correlation and regression. CMAJ: Canadian Medical Association Journal, 152(4), 497. https://pmc.ncbi.nlm.nih.gov/articles/PMC1337703/
27. Coefficient of Determination. (2008). The Concise Encyclopedia of Statistics, 88–91. https://doi.org/10.1007/978-0-387-32833-1_62
28. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. J Am Med Assoc. 1982;247:2543-2546.
29. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361-387.
30. Vergara, D., Armato, S. G., Hadjiiski, L., & Drukker, K. (2024). Best Practices for Artificial Intelligence and Machine Learning for Computer-Aided Diagnosis in Medical Imaging. In Journal of the American College of Radiology (Vol. 21, Issue 2, pp. 341–343). Elsevier B.V. https://doi.org/10.1016/j.jacr.2023.10.021
31. Mahmood, U., Shukla-Dave, A., Chan, H.-P., Drukker, K., Samala, R. K., Chen, Q., Vergara, D., Greenspan, H., Petrick, N., Sahiner, B., Huo, Z., Summers, R. M., Cha, K. H., Tourassi, G., Deserno, T. M., Grizzard, K. T., Näppi, J. J., Yoshida, H., Regge, D., … Hadjiiski, L. (2024). Artificial intelligence in medicine: mitigating risks and maximizing benefits via quality assurance, quality control, and acceptance testing. BJR|Artificial Intelligence, 1(1). https://doi.org/10.1093/bjrai/ubae003
32. Hadjiiski, L., Cha, K., Chan, H. P., Drukker, K., Morra, L., Näppi, J. J., Sahiner, B., Yoshida, H., Chen, Q., Deserno, T. M., Greenspan, H., Huisman, H., Huo, Z., Mazurchuk, R., Petrick, N., Regge, D., Samala, R., Summers, R. M., Suzuki, K., … Armato, S. G. (2023). AAPM task group report 273: Recommendations on best practices for AI and machine learning for computer-aided diagnosis in medical imaging. Medical Physics, 50(2), e1–e24. https://doi.org/10.1002/mp.16188
33. Replogle, W. H., Johnson, W. D., & Hoover, K. W. (2009). Using evidence to determine diagnostic test efficacy. Worldviews on Evidence-Based Nursing, 6(2), 87–92. https://doi.org/10.1111/J.1741-6787.2009.00148.X
34. Fryback, D. G., & Thornbury, J. R. (1991). The efficacy of diagnostic imaging. Medical Decision Making : An International Journal of the Society for Medical Decision Making, 11(2), 88–94. https://doi.org/10.1177/0272989X9101100203
35. Vokinger, K. N., Feuerriegel, S., & Kesselheim, A. S. (2021). Continual learning in medical devices: FDA's action plan and beyond. In The Lancet Digital Health (Vol. 3, Issue 6, pp. e337–e338). Elsevier Ltd. https://doi.org/10.1016/S2589-7500(21)00076-5

36. Raymond Geis, J., Brady, A. P., Wu, C. C., Spencer, J., Ranschaert, E., Jaremko, J. L., Langer, S. G., Kitts, A. B., Birch, J., Shields, W. F., van den Hoven van Genderen, R., Kotter, E., Gichoya, J. W., Cook, T. S., Morgan, M. B., Tang, A., Safdar, N. M., & Kohli, M. (2019). Ethics of artificial intelligence in radiology: Summary of the joint European and North American multisociety statement. Radiology, 293(2), 436–440. https://doi.org/10.1148/radiol.2019191586
37. el Naqa, I., Karolak, A., Luo, Y., Folio, L., Tarhini, A. A., Rollison, D., & Parodi, K. (2023). Translation of AI into oncology clinical practice. In Oncogene (Vol. 42, Issue 42, pp. 3089–3097). Springer Nature. https://doi.org/10.1038/s41388-023-02826-z
38. National Institutes of Health (2025) PubMed. Available at: https://pubmed.ncbi.nlm.nih.gov/ (Accessed: 12/08/2025).
39. Art. 6 GDPR – Lawfulness of processing, 2025. General Data Protection Regulation (GDPR). Available at: https://gdpr-info.eu/art-6-gdpr/ (Accessed 12/08/2025).
40. Article 26 EU AI Act, Obligations of Deployers of High-Risk AI Systems. Available at: https://artificialintelligenceact.eu/article/26/ (Accessed: 03/09/2025)
41. National Financial Regulations [WWW Document], n.d. HSE.ie. Available at: https://www.hse.ie/eng/about/who/finance/nfr/about-nfrs.html (Accessed 12/08/2025)
42. Public Procurement Guidelines for Goods and Services [WWW Document], 2023. Available at: https://www.gov.ie/en/office-of-government-procurement/publications/public-procurement-guidelines-for-goods-and-services/ (Accessed 12/08/2025)