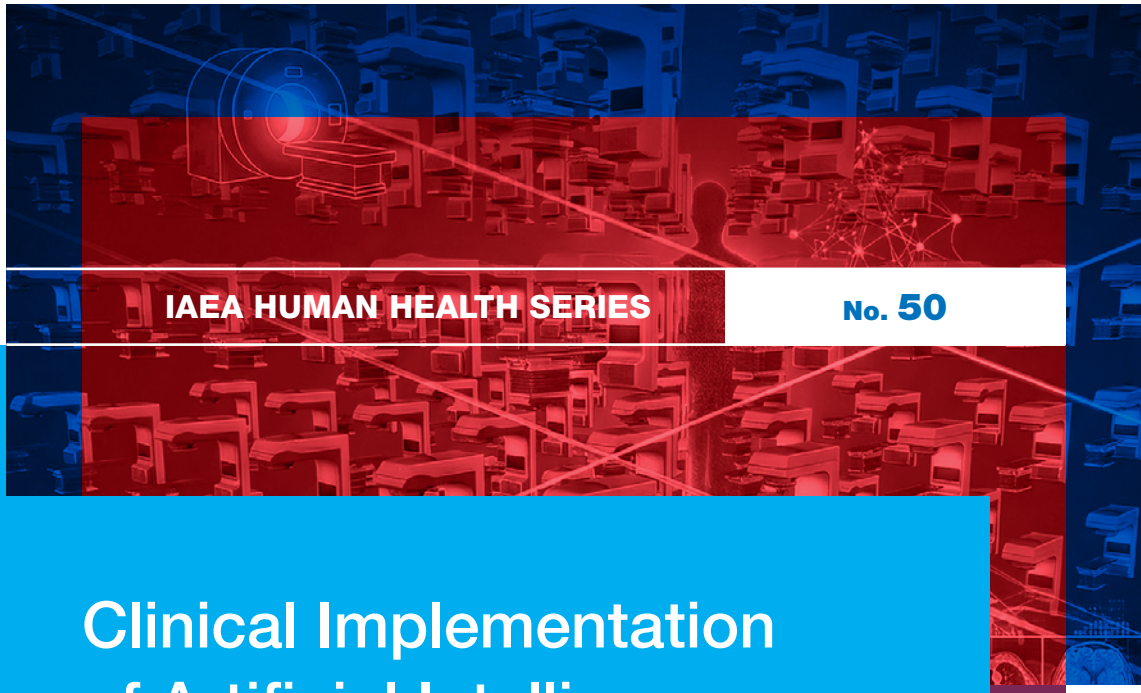
IAEA
International Atomic Energy Agency



IAEA HUMAN HEALTH SERIES No. 50

Clinical Implementation of Artificial Intelligence Systems in Medical Imaging and Radiotherapy

Guidelines for Medical Physicists

CLINICAL IMPLEMENTATION OF ARTIFICIAL INTELLIGENCE SYSTEMS IN MEDICAL IMAGING AND RADIOTHERAPY

The following States are Members of the International Atomic Energy Agency:

AFGHANISTAN
ALBANIA
ALGERIA
ANGOLA
ANTIGUA AND BARBUDA
ARGENTINA
ARMENIA
AUSTRALIA
AUSTRIA
AZERBAIJAN
BAHAMAS, THE
BAHRAIN
BANGLADESH
BARBADOS
BELARUS
BELGIUM
BELIZE
BENIN
BOLIVIA, PLURINATIONAL STATE OF
BOSNIA AND HERZEGOVINA
BOTSWANA
BRAZIL
BRUNEI DARUSSALAM
BULGARIA
BURKINA FASO
BURUNDI
CABO VERDE
CAMBODIA
CAMEROON
CANADA
CENTRAL AFRICAN REPUBLIC
CHAD
CHILE
CHINA
COLOMBIA
COMOROS
CONGO
COOK ISLANDS
COSTA RICA
CÔTE D'IVOIRE
CROATIA
CUBA
CYPRUS
CZECH REPUBLIC
DEMOCRATIC REPUBLIC OF THE CONGO
DENMARK
DJIBOUTI
DOMINICA
DOMINICAN REPUBLIC
ECUADOR
EGYPT
EL SALVADOR
ERITREA
ESTONIA
ESWATINI
ETHIOPIA
FIJI
FINLAND
FRANCE
GABON
GAMBIA, THE
GEORGIA
GERMANY
GHANA
GREECE
GRENADA
GUATEMALA
GUINEA
GUYANA
HAITI
HOLY SEE
HONDURAS
HUNGARY
ICELAND
INDIA
INDONESIA
IRAN, ISLAMIC REPUBLIC OF
IRAQ
IRELAND
ISRAEL
ITALY
JAMAICA
JAPAN
JORDAN
KAZAKHSTAN
KENYA
KOREA, REPUBLIC OF
KUWAIT
KYRGYZSTAN
LAO PEOPLE'S DEMOCRATIC REPUBLIC
LATVIA
LEBANON
LESOTHO
LIBERIA
LIBYA
LIECHTENSTEIN
LITHUANIA
LUXEMBOURG
MADAGASCAR
MALAWI
MALAYSIA
MALDIVES
MALI
MALTA
MARSHALL ISLANDS
MAURITANIA
MAURITIUS
MEXICO
MONACO
MONGOLIA
MONTENEGRO
MOROCCO
MOZAMBIQUE
MYANMAR
NAMIBIA
NEPAL
NETHERLANDS, KINGDOM OF THE
NEW ZEALAND
NICARAGUA
NIGER
NIGERIA
NORTH MACEDONIA
NORWAY
OMAN
PAKISTAN
PALAU
PANAMA
PAPUA NEW GUINEA
PARAGUAY
PERU
PHILIPPINES
POLAND
PORTUGAL
QATAR
REPUBLIC OF MOLDOVA
ROMANIA
RUSSIAN FEDERATION
RWANDA
SAINT KITTS AND NEVIS
SAINT LUCIA
SAINT VINCENT AND THE GRENADINES
SAMOA
SAN MARINO
SAUDI ARABIA
SENEGAL
SERBIA
SEYCHELLES
SIERRA LEONE
SINGAPORE
SLOVAKIA
SLOVENIA
SOMALIA
SOUTH AFRICA
SPAIN
SRI LANKA
SUDAN
SWEDEN
SWITZERLAND
SYRIAN ARAB REPUBLIC
TAJIKISTAN
THAILAND
TOGO
TONGA
TRINIDAD AND TOBAGO
TUNISIA
TÜRKİYE
TURKMENISTAN
UGANDA
UKRAINE
UNITED ARAB EMIRATES
UNITED KINGDOM OF GREAT BRITAIN AND NORTHERN IRELAND
UNITED REPUBLIC OF TANZANIA
UNITED STATES OF AMERICA
URUGUAY
UZBEKISTAN
VANUATU
VENEZUELA, BOLIVARIAN REPUBLIC OF
VIET NAM
YEMEN
ZAMBIA
ZIMBABWE

The Agency's Statute was approved on 23 October 1956 by the Conference on the Statute of the IAEA held at United Nations Headquarters, New York; it entered into force on 29 July 1957. The Headquarters of the Agency are situated in Vienna. Its principal objective is "to accelerate and enlarge the contribution of atomic energy to peace, health and prosperity throughout the world".

IAEA HUMAN HEALTH SERIES No. 50

CLINICAL IMPLEMENTATION OF ARTIFICIAL INTELLIGENCE SYSTEMS IN MEDICAL IMAGING AND RADIOTHERAPY

GUIDELINES FOR MEDICAL PHYSICISTS

ENDORSED BY
THE AMERICAN ASSOCIATION OF PHYSICISTS IN MEDICINE,
THE AUSTRALASIAN COLLEGE OF PHYSICAL SCIENTISTS AND ENGINEERS IN MEDICINE,
THE EUROPEAN FEDERATION OF ORGANISATIONS FOR MEDICAL PHYSICS,
THE EUROPEAN SOCIETY FOR RADIOTHERAPY AND ONCOLOGY,
THE FEDERATION OF AFRICAN MEDICAL PHYSICS ORGANIZATIONS,
THE INTERNATIONAL ORGANIZATION FOR MEDICAL PHYSICS
AND THE LATIN AMERICAN ASSOCIATION OF MEDICAL PHYSICS

INTERNATIONAL ATOMIC ENERGY AGENCY
VIENNA, 2026

COPYRIGHT NOTICE

Publishing Section
International Atomic Energy Agency
Vienna International Centre
PO Box 100
1400 Vienna, Austria
tel.: +43 1 2600 22529 or 22530
email: sales.publications@iaea.org
www.iaea.org/publications

Printed by the IAEA in Austria
May 2026
STI/PUB/2135
https://doi.org/10.61092/iaea.9olq-aqa9

IAEA Library Cataloguing in Publication Data

Names: International Atomic Energy Agency.
Title: Clinical implementation of artificial intelligence systems in medical imaging and radiotherapy : guidelines for medical physicists / International Atomic Energy Agency.
Description: Vienna : International Atomic Energy Agency, 2026. | Series: IAEA human health series, ISSN 2075-3772 ; no. 50 | Includes bibliographical references.
Identifiers: IAEAL 26-01811 | ISBN 978-92-0-130026-3 (paperback : alk. paper) | ISBN 978-92-0-130126-0 (pdf) | ISBN 978-92-0-130226-7 (epub)
Subjects: LCSH: Diagnostic imaging — Technological innovations. | Artificial intelligence. | Radiology.
Classification: UDC 615.849:004.8 | STI/PUB/2135

FOREWORD

Artificial intelligence (AI) has significant potential to impact processes in science and technology, including in the area of human health. While bringing potential benefits to healthcare, the application of AI systems also introduces new challenges and potential risks. The clinical deployment of AI systems requires an adequate regulatory framework and highly educated, well trained health professionals to ensure their safe, ethical and beneficial use.

The use of AI in human health was a topic of the Technical Meeting on Artificial Intelligence for Nuclear Technology and Applications hosted by the IAEA in 2021, an international, cross-cutting event that helped to establish the IAEA's overall position with regard to the application of AI. During this meeting, experts organized in several working groups discussed opportunities and challenges associated with the implementation of AI systems in medical imaging and radiotherapy, and they identified a need for guidelines addressing the use of AI in the field of medical physics.

This conclusion was consistent with needs identified earlier. In 2020, the IAEA hosted the 19th Biennial Meeting of the Secondary Standards Dosimetry Laboratory (SSDL) Scientific Committee (SSC-19) for the Evaluation of and Recommendations on the Dosimetry Programme and the IAEA/WHO SSDL Network. The committee identified two related needs: first, the need for a new curriculum to update the knowledge and academic education of medical radiation physicists in the areas of data science, regression analysis, statistical learning and deep learning, and second, the need for guidelines for standardization and quality assurance of quantitative image analysis tools used in radiology, considering their final use in methods such as radiomics.

Training Course Series No. 83, Artificial Intelligence in Medical Physics: Roles, Responsibilities, Education and Training of Clinically Qualified Medical Physicists, published in 2023, addresses the first need, framing the roles and responsibilities of medical physicists in the implementation and application of AI in the medical uses of radiation. The present publication addresses the second need, providing practical information on the implementation of imaging based AI systems in clinical settings, developed in a series of consultancy meetings held in 2022, 2023 and 2024. The publication is aimed at clinically qualified medical physicists, who are health professionals uniquely positioned to bridge the gap between complex AI systems and practical clinical applications.

This publication has been endorsed by the American Association of Physicists in Medicine (AAPM), the Australasian College of Physical Scientists and Engineers in Medicine (ACPSEM), the European Federation of Organisations for Medical Physics (EFOMP), the European Society for Radiotherapy and Oncology (ESTRO), the Federation of African Medical Physics Organizations

(FAMPO), the International Organization for Medical Physics (IOMP) and the Latin American Association of Medical Physics (ALFIM). The IAEA is grateful to all the experts who contributed to this publication, in particular A. Dekker (Kingdom of the Netherlands), M.L. Giger (United States of America) and A. Zwanenburg (Germany) for their substantial efforts in drafting and technical review, as well as B. Haibe-Kains (Canada) and A.K. Jha (India) for their important contributions. The IAEA officers responsible for this publication were O. Ciraj-Bjelac and E. Titovich of the Division of Human Health.

CONTENTS

1. INTRODUCTION

1.1. BACKGROUND

The rapid integration of artificial intelligence (AI) into healthcare systems has the potential to significantly transform medical practice and provide new ways to improve diagnosis, treatment and patient care.

The term ‘artificial intelligence’ was first introduced to describe the concept of using computers to simulate intelligent behaviour and critical thinking in the 1950s [1]. The application of the concept in medicine began to emerge in the 1980s, with the introduction of personal computers and early techniques involving machine learning and deep learning, including convolutional neural networks [2–8]. Advances in information technologies and deep learning in the 2000s significantly contributed to the development of sophisticated AI systems that incorporate multiple inputs and complex algorithms, which are now capable of self-learning — that is, learning directly from data without the need for expert provided reference standard (sometimes referred to as ‘ground truth’) labels. With these developments, AI has assumed an increasingly important role in healthcare, becoming integral to various processes and tasks in medical imaging and radiation oncology [6, 7, 9–11].

Along with the potential benefits, the deployment of AI in healthcare also introduces new challenges and risks [4, 11, 12]. The integration of AI systems in clinical settings is currently constrained by concerns regarding trustworthiness and transparency, limited generalizability, fragility in real world scenarios, inadequate interoperability with other clinical systems and potential bias related to patient populations, as well as unresolved legal and ethical issues related to accountability and human autonomy. These challenges are emphasized in the World Health Organization (WHO) guidelines on ethics and governance of AI for health [11].

The safe and effective clinical deployment of AI systems requires an appropriate regulatory framework and a highly educated, well trained healthcare workforce [11, 13, 14]. Clearly defined roles and responsibilities for these professionals are essential to ensure safe, ethical and beneficial use. This publication focuses on imaging based AI systems, defined here as AI systems that utilize imaging data as input (referred to throughout this publication as ‘AI systems’). The multidisciplinary teams responsible for the clinical implementation and validation of such systems need to be co-led by clinically qualified medical physicists (CQMPs) [15] and radiological medical practitioners specializing in radiation oncology, nuclear medicine, diagnostic and interventional radiology or other relevant clinical disciplines. Additional essential team members include

information technology (IT) experts, data scientists, knowledge engineers and manufacturer representatives, as well as regulatory, data protection and ethics officers, whose collective expertise is crucial for successful implementation [16].

CQMPs play a central role in realizing the benefits of AI systems for patients and healthcare systems. Their involvement encompasses a spectrum of responsibilities, from ensuring seamless integration with existing medical infrastructure to optimizing AI systems for specific diagnostic or prognostic tasks. In short, CQMPs and the healthcare organizations in which they work have to consider the introduction of AI systems into clinical practice with the same rigor and deliberation as the introduction of new imaging devices, such as computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, positron emission tomography (PET) and single photon emission computed tomography (SPECT) systems, or new treatment modalities, such as intensity modulated radiotherapy, volumetric modulated arc therapy, image guided radiotherapy, hybrid linac based radiotherapy and magnetic resonance guided focused ultrasound.

1.2. OBJECTIVE

The roles and responsibilities of the CQMP are defined in IAEA Human Health Series No. 25 [15]. Considering the lack of comprehensive guidelines for the implementation of imaging based AI systems in clinical settings and the rapidly evolving nature of AI technology in healthcare, particularly in radiation medicine, there is a clear need for a structured framework to guide CQMPs. As leading members of multidisciplinary teams, CQMPs are uniquely positioned to bridge the gap between complex AI algorithms and practical clinical applications.

The key responsibilities of a CQMP in the clinical implementation of AI systems in medical imaging and radiotherapy can be grouped according to the phases of the implementation process, including identification of organizational needs; market research; preselection and demonstration; selection and purchasing; installation; development and implementation of quality assurance (QA) programmes, including acceptance, commissioning and routine performance testing; monitoring and evaluation of the impact of the AI system on users, decision making and the clinical workflow; and maintenance and decommissioning of the AI system [4].

This publication offers detailed guidance to CQMPs and their teams, covering all phases of AI system implementation. Guidance and recommendations provided here in relation to identified good practices represent expert opinion but are not made on the basis of a consensus of all Member States.

1.3. SCOPE

This publication provides comprehensive guidance for the clinical implementation of imaging based AI systems in medical imaging and radiotherapy. These AI systems use medical images as their primary input and can be applied across all medical disciplines that involve ionizing radiation, including diagnostic radiology, nuclear medicine and radiotherapy. The publication addresses the entire process, from the initial assessment of needs through selection, commissioning, ongoing (quality) management and eventual decommissioning. The AI system is assumed to be an already approved end product, for example approved as a medical device, such that the roles and responsibilities of the CQMP in the process of the clinical implementation of AI, as described in Refs [4, 15], are addressed throughout the process from the end user perspective. Although the primary focus is on imaging based AI systems, the guidance provided in this publication is broadly applicable to non-imaging based AI systems as well. The publication is intended primarily for manufacturer developed AI systems but could also be applicable to the implementation of locally developed systems.

1.4. STRUCTURE

This publication is structured to guide CQMPs through the selection, safe implementation and effective use of AI systems in clinical settings. Section 2 provides an overview of the multistep process of AI system development and validation, which is relevant for contextualizing the CQMP responsibilities described in the subsequent sections. Section 3 presents clinical and data science perspectives on the tasks performed by imaging based AI systems. Expanding on the roles and responsibilities relevant to applying AI in radiation medicine [4], Section 4 focuses on the roles of the CQMP in implementing medical imaging based AI systems within clinical environments, detailing the responsibilities in ensuring these systems are integrated effectively and safely. Section 5 provides practical, actionable information to CQMPs on leading and supporting the clinical implementation process as part of a multidisciplinary team, including initial evaluation, multidisciplinary coordination, continuous quality management and, ultimately, decommissioning. Section 6 offers concluding remarks on the rapid integration of AI into healthcare and the transformative role CQMPs play in this process. Appendix I provides a template model card for clinical AI systems based on medical imaging, outlining a standardized format by which manufacturers may provide information about their products. Appendix II summarizes the key tasks and considerations for CQMPs and their teams when acquiring and

implementing an imaging based AI system, with references to the relevant sections of this publication.

2. FUNDAMENTALS OF ARTIFICIAL INTELLIGENCE SYSTEM DEVELOPMENT

Although this publication focuses primarily on imaging based AI systems that are certified, approved or otherwise accepted as medical devices, a basic understanding of the AI system development cycle is important for understanding potential challenges in clinical practice.

As shown in Fig. 1, AI system development is a multistep process. Careful execution of each step is essential, and multiple iterations are often necessary to arrive at a satisfactory result. Many AI systems have contributed to the automatic quantification of patterns in medical imaging data. These capabilities have been progressively enhanced by advances in the AI field, such as machine learning, deep learning and foundation models, whose relationship is illustrated in Fig. 2. Definitions of AI, machine learning, deep learning and other relevant terms are provided in IAEA Training Course Series No. 83 [4]. An overview of the steps involved in AI system development and validation is presented below. Other guidelines providing more in-depth information on AI system development are also available, such as CLAIM [17], STARD-AI [18] and TRIPOD-AI [19, 20]; these are further described in Section 2.2.

2.1. CLINICAL END POINT DEFINITION

The first step in AI system development is to clearly define the clinical problem (i.e. the use case) that the AI system is intended to address. Broad examples of clinical end points include detection or diagnosis of disease; disease classification; prognosis; treatment response; prediction of progression free survival, disease free survival, distant metastasis or overall survival; and supporting tasks such as image segmentation, image generation and interpretation, patient motion correction and machine performance monitoring. Such end points are always defined in conjunction with the specific task and the intended population. For example, ‘prediction of progression free survival’ is not an end point by itself, whereas ‘prediction of progression free survival based on T1 weighted pre-treatment MRI and clinical markers in patients with locally

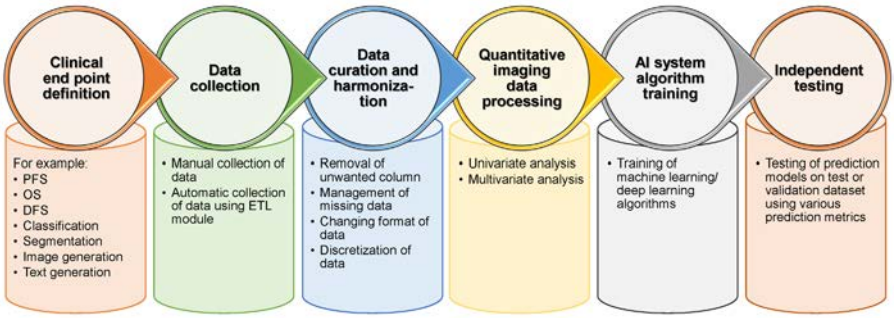


FIG. 1. Steps involved in the development of an imaging based artificial intelligence (AI) system. DFS — disease free survival; ETL — extract, transform, load; OS — overall survival; PFS — progression free survival.

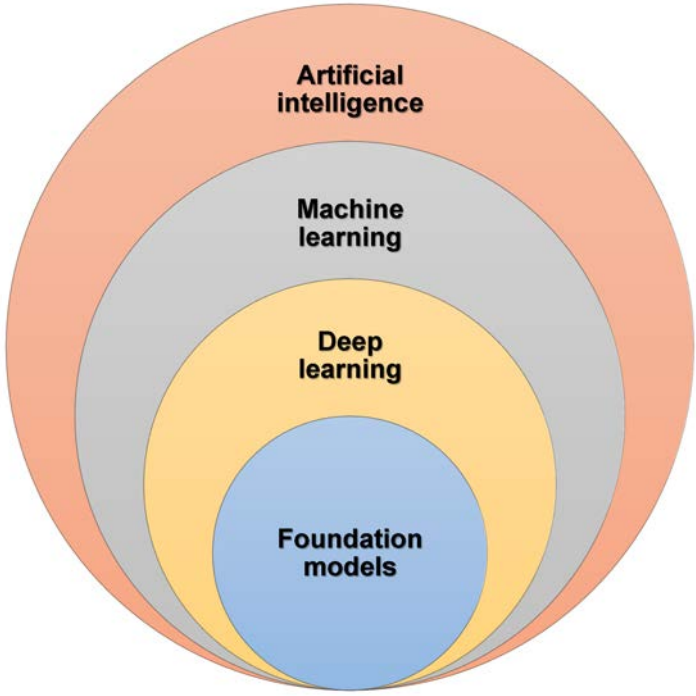


FIG. 2. Hierarchical relationship between artificial intelligence, machine learning, deep learning and foundation models, illustrating how each concept is a subset of the previous one.

advanced head and neck squamous cell carcinoma treated with neoadjuvant radiotherapy' constitutes a clearly defined end point [21].

Many currently regulatory approved imaging based AI systems have narrowly defined clinical end points. Therefore, any non-intended (off-label) use, in which the input imaging data or tasks differ from the intended use, population or claims, needs to be critically evaluated or avoided. Although a non-intended use of an AI system may still yield output, the output will most likely be inaccurate, unstable or clinically meaningless, thus requiring careful assessment.

Although this section focuses mainly on defined clinical end points, AI systems may also contribute to image reconstruction, image restoration (quality improvement), image quality monitoring, image management and organizational aspects.

2.2. DATA COLLECTION

Once the clinical end point has been defined, the corresponding data need to be collected. Several methods of data collection may be considered, including manual, semi-automated or fully computerized data collection and storage [22]. To support the development of a robust and generalizable AI system, the collected data have to reflect the variability observed in clinical scenarios, for example in terms of manufacturers, sites and protocols. Data collection is also the stage at which the data analysis strategy is decided. This includes specification of the intended population and, importantly, decisions regarding which data will be used to train and validate the AI system during development and which data will be set aside for final testing of the fully specified AI system. There are numerous guidelines for AI development in healthcare, many of which emphasize the (reporting of the) data collection process, including the following frameworks [23]:

— CLAIM (Checklist for Artificial Intelligence in Medical Imaging): Promotes transparent and reproducible reporting of AI research in medical imaging. It provides a structured framework for documenting the development, validation and implementation of AI in imaging [24].
— CONSORT-AI (Consolidated Standards of Reporting Trials — AI extension): Adapts the CONSORT guidelines for reporting randomized controlled trials involving AI. It addresses specific challenges related to AI interventions, such as algorithm updates and data handling [25].
— DECIDE-AI (Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence): Provides guidelines for the early stage clinical evaluation of AI based decision

support systems. It outlines best practices for assessing the feasibility, safety and effectiveness of AI tools in clinical settings [26].

— FUTURE-AI: Provides international consensus guidelines for trustworthy and deployable AI in healthcare. It covers technical, clinical, socioethical and legal dimensions, ensuring that AI systems are designed, developed, validated and monitored responsibly [27].
— PROBAST+AI (Prediction Model Risk of Bias Assessment Tool — AI extension): Used to assess the risk of bias in prediction model studies, including those involving AI. It helps identify potential sources of bias that could affect the validity of the model predictions [19, 28, 29].
— SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials — AI extension): Extends the original SPIRIT guidelines to include AI interventions in clinical trials. It ensures that AI components are adequately described in trial protocols, promoting transparency and reproducibility [30].
— STARD-AI (Standards for Reporting of Diagnostic Accuracy Studies — AI extension): Focuses on the reporting of diagnostic accuracy studies involving AI. It aims to improve the completeness and transparency of reporting, ensuring that AI based diagnostic tools are accurately evaluated [18].
— TRIPOD-AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis — AI extension): Extends the original TRIPOD guidelines to include AI and machine learning models. It emphasizes the need for clear reporting of model development, validation and performance to facilitate critical appraisal and replication [20].

2.3. DATA CURATION AND HARMONIZATION

The objective of data curation is to identify and remove or correct any errors, inconsistencies or missing values in the collected data. This step is commonly followed by data transformation, harmonization and/or homogenization, as appropriate, with the aim of converting the data into a consistent format for analysis [31]. Such processing can be achieved through, for example, normalization of imaging data, use of unified labelling schemes or discretization of data [5, 32].

2.4. QUANTITATIVE IMAGING DATA PROCESSING

Quantitative imaging data processing involves the computer based extraction of quantitative characteristics/features from curated and harmonized data. This step includes computation, preselection and transformation, resulting in image derived features suitable for subsequent analysis. Examples of quantitative imaging data processing include computation of volumetric, shape, intensity and texture features, such as radiomics features [5, 33–36], as well as imaging features extracted using deep learning and computed using deep neural networks [37, 38].

2.5. ARTIFICIAL INTELLIGENCE SYSTEM ALGORITHM TRAINING

At the core of any AI system is an algorithm (‘model’) that is designed to infer an output based on the data received as input. The choice of algorithm, its architecture and configuration, the intended population data used for its training and the training procedure are determined by the data science task corresponding to the defined clinical end point (see Section 3.2). To ensure that such algorithms yield output that relates to the clinical end point, AI systems undergo a training process. This process typically involves the use of input data with known reference standards (‘supervised learning’), allowing the algorithm to learn to recognize relevant patterns. Other algorithms learn patterns in data without an explicit reference standard (‘unsupervised learning’ or ‘self-supervised learning’). Self-supervised algorithms typically form the basis of foundation models. After learning the important patterns in a dataset, foundation models can either be used directly for inference (‘zero shot learning’) or undergo fine tuning for the specific task using supervised learning (‘transfer learning’). Depending on the type of algorithm employed, feature selection and dimensionality reduction techniques may be used to identify the most relevant features and patterns that contribute to the prediction task. These techniques help to reduce dimensionality and improve AI system interpretability and efficiency. Such tasks may be intrinsic to the algorithm (e.g. deep learning architectures) or extrinsic (e.g. generalized linear models). In addition, knowledge of the data characteristics and the representativeness of the data is necessary to reduce bias.

After training on the initial training data, an AI system is usually tested against a wider validation dataset. This validation step helps to determine whether the AI system needs further improvement, for example by changing parameters related to the algorithm. Depending on the results, the AI system may be trained again.

2.6. INDEPENDENT TESTING

Once an AI system demonstrates satisfactory performance during training and validation, it is then typically assessed using an independent test dataset. At this stage, the performance of the trained model(s) is assessed on independent data representing the full spectrum of known attributes of the intended population and clinical cases, using appropriate evaluation metrics, to ensure the generalizability and reliability of the model [39]. This test dataset could come from the same organization as the training data or from another organization. Testing with data from a different source is referred to as external validation. It is important to note that AI systems are influenced by the clinical standard of care at the time of their development and therefore need to be tested independently and repeatedly, ideally within an appropriate population.

The quality of an AI system is evaluated using metrics appropriate to the clinical task [40]. Examples related to various types of clinical task are provided in Section 3.2.

3. OVERVIEW AND USE OF MEDICAL IMAGING BASED ARTIFICIAL INTELLIGENCE SYSTEMS

Clinical AI systems have the potential to improve the efficacy or efficiency (or both) of various medical procedures that involve the use of medical imaging, for example through integration of AI systems in image acquisition, viewing protocols, diagnostic assessment, treatment guidance, outcome prediction and patient follow-up.

Two complementary perspectives need to be considered. The clinical tasks in which an AI system could potentially be involved are discussed in Section 3.1, and the data science perspective is addressed in Section 3.2.

3.1. CLINICAL TASK PERSPECTIVE

Clinical tasks can be broadly divided into two categories. The first category focuses on supporting or automating existing clinical workflow processes (i.e. improving healthcare efficiency), including improving the workflow itself. The second category focuses on improving the prediction of patient risk, diagnosis, treatment response or outcome, as well as detecting errors or suggesting interventions (i.e. improving healthcare efficacy). AI system algorithms are

optimized for specific performance metrics, such as sensitivity, specificity, false negative/positive rate or overall performance, depending on the clinical task. Whether the aim is efficiency, efficacy or both, the choice of approach depends on the specific use case and the needs of the healthcare organization and end user (see Section 5.1). AI systems can function in a variety of roles, including being used as secondary or concurrent readers/decision makers, functioning as primary readers/decision makers, working as rule-out or triage systems, assisting in or automating quality control (QC) activities, operating as monitoring systems or facilitating treatment planning. The clinical task perspective of an AI system is schematically represented in Fig. 3.

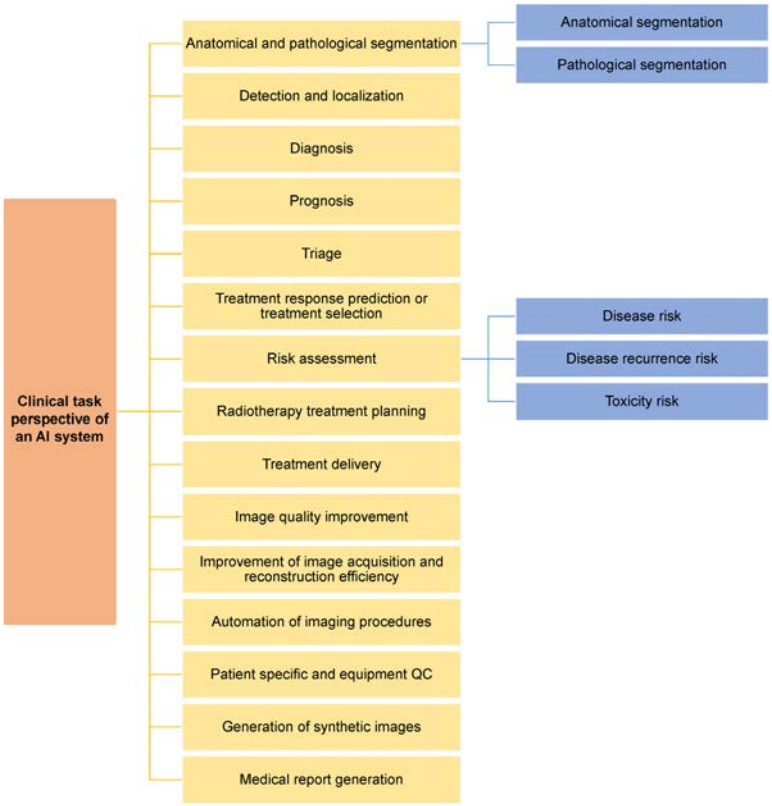


FIG. 3. Schematic representation of the clinical task perspective of an artificial intelligence (AI) system. QC — quality control.

Examples of different types of clinical task for which an AI system could be used include the following:

(a) Anatomical and pathological segmentation: AI systems can be used for segmentation or delineation of specific regions of interest in medical images, such as those corresponding to normal organs or to pathological conditions:
    (i) Anatomical segmentation: Delineates normal anatomical structures. In radiology, nuclear medicine and radiotherapy, AI systems for anatomical segmentation can be used to enhance treatment planning software and workflow through the automated creation of contours of organs at risk or target volumes.
    (ii) Pathological segmentation: Delineates pathological regions, such as tumours, epileptic foci or enlarged thyroids [41]. Although customized for each pathology, AI systems for pathological segmentation are broadly applicable, for example in automated delineation of suspicious regions, monitoring disease progression or streamlining radiotherapy treatment planning through automated creation of target contours. AI systems for pathology have recently been released, but their clinical utility and acceptability remain to be comprehensively assessed [42, 43].

    Both types of AI segmentation system can be integrated into other systems, for example for the characterization of anatomical and pathological tissues, for physiology and biology based radiotherapy treatment optimization, or for treatment staging and follow-up in radiology and nuclear medicine. For QC purposes, segmentation system outputs can be checked by comparison with manual segmentations or those created by different AI systems. The use of segmentation AI systems requires careful consideration of factors that may affect their performance, including the images used in the training dataset and the segmentations used as the reference standards, the imaging modality and its associated image quality, the choice of and adherence to clinical segmentation guidelines, and the acceptability of the segmentations for the particular downstream clinical care path.
(b) Detection and localization: AI systems for detection are used to localize and assess the presence or absence of specific abnormalities in images. They are typically employed in screening programmes for detection of potential abnormalities in asymptomatic patients, for example in breast cancer screening. In such cases, an AI system might be used as a secondary or concurrent reader, or as a primary reader to ‘rule out’ normal cases. Rule-out AI systems can reduce the workload of radiologists by identifying the ‘normal’ cases with least likelihood of abnormalities, which can be a large proportion of the cases in population screening programmes. It is anticipated that, in specific use cases, AI systems may function as primary

and potentially sole readers in the future. A detection AI system needs to be assessed both as a stand alone system and in comparison with the performance of experienced expert radiologists reading with and without AI support in the same clinical setting.

(c) Diagnosis: Diagnostic AI systems characterize specific findings observed in imaging. For example, such systems may assess the likelihood of malignancy of nodules or lesions detected in a lung screening CT or in a breast MRI, or they may perform automated tumour staging based on imaging, with or without the inclusion of additional clinical information.

(d) Prognosis: Whereas diagnostic AI systems characterize the current state of a specific finding on images, prognostic AI systems aim to predict a future state, such as the expected pattern of disease progression or the probability that a lesion will become malignant. For example, in breast cancer imaging, the aggressiveness of disease may be of interest, whereas in COVID-19 imaging, the likelihood of future admission to an intensive care unit may be an important predictor.

(e) Triage: AI systems can also be used to triage cases, especially in situations where a high workload prevents immediate assessment of imaging by clinical experts (e.g. radiologists in an emergency care department). In such circumstances, an AI system can be used to identify and rank which images need to be read most urgently, thereby expediting medical care for the patients most in need.

(f) Treatment response prediction or treatment selection: AI systems may integrate medical images with other clinical data to provide predictions of disease response to treatment, normal tissue toxicity and complications, and patient reported outcomes, as well as to support treatment selection and decision making. Such applications have the potential to improve patient outcomes and healthcare system efficacy.

(g) Risk assessment: The types of risk that can be assessed include the following:

    (i) Disease risk: AI systems can be used to predict the risk of future cancer or other diseases (e.g. cardiovascular disease) on the basis of features extracted from seemingly normal images, potentially in combination with other clinical data (including prior or planned diagnostic or therapeutic procedures) or demographic information. For example, in screening mammography, future breast cancer risk can be assessed using clinical data together with computer extracted characteristics of the breast parenchyma (specifically, breast density and texture) from mammography images.

    (ii) Disease recurrence risk: AI systems can be used to predict the risk of future local or regional recurrence and distant metastasis after a treatment has been concluded. The relevant features considered in these

models can include patient demographics, lifestyle factors, treatments received or ongoing, biopsy results and genetic information, as well as extracted features of tissues from medical images.

(iii) Toxicity risk: AI systems can also be used to predict the risk of toxicity after completed treatments such as radiotherapy or radiopharmaceutical therapy. The relevant features considered in these models include radiation dose to the organ at risk and normal tissues, the pre-existing condition of the organ, relevant genetic information and measures taken to mitigate toxicity.

(h) Radiotherapy treatment planning: AI systems may help in optimizing radiotherapy treatment workflow and plans. As described above, AI can initially be used to segment target volumes to be treated and critical structures to be avoided, potentially incorporating multimodality and synthetic images as discussed below. During the planning process, AI systems can be used to determine an optimal treatment plan and modality for a specific patient, for example by optimizing dose distribution given the geometry, size of structures and target volumes, or treatment beam characteristics based on the device options and hardware limitations that determine the conditions for the optimization problem (e.g. widths of multileaf collimator leaves, position limits). AI systems can also be used to model uncertainties in the treatment delivery process, such as uncertainties in target and organ segmentation, patient motion, set-up variability and changes in organ shape and size over time. Such systems could be used to assess treatment planning robustness, to develop a more robust treatment plan and to improve the consistency and quality of the treatment planning process (e.g. by reducing interuser variability). Finally, AI might be used for the radiation dose calculation itself, replacing or supporting current model based dose calculation algorithms.

(i) Treatment delivery: During treatment, AI systems can monitor patient positioning and physiological state, patient and tissue motion, radiation dose delivery and equipment performance; they can also provide warnings of malfunctions, help guide and adapt treatment or automate treatment procedures. Examples include real time monitoring of tumour ablation in interventional radiology and theranostics, patient monitoring under fluoroscopy in cardiovascular plaque removal and the monitoring of patient motion during radiotherapy treatment delivery.

(j) Image quality improvement: AI systems may be used to measure and monitor physical image quality or to improve image quality, for example by reducing artefacts and noise or by inferring aspects of an image from other data sources (e.g. by ‘stitching’ a limited field of view cone beam CT image using a full field of view CT image).

(k) Improvement of image acquisition and reconstruction efficiency: AI systems may improve the efficiency of image acquisition and reconstruction, for example by enabling faster image acquisition in MRI or CT or acquisition with a decreased radiation dose to the patient by optimizing scan parameters. AI based reconstruction methods may generate high quality images using low count acquisition, thereby reducing image noise.

(l) Automation of imaging procedures: AI systems may be used to automate one or more steps in the process from the request for imaging to reporting for use in decision making for patient care, with the potential to achieve a 'closed loop' workflow that minimizes the risk of human error. For example, an AI system may automate the referral procedure or select the appropriate imaging protocol based on the prescription and then communicate this information to the X ray unit directly. QC of such tools ensures that an optimized imaging protocol is applied.

(m) Patient specific and equipment QC: AI systems may assist in routine QC tests conducted by the CQMP or other professionals, for example by automating data collection from routine phantom studies or facilitating verification of dose delivery during patient radiation treatment. AI systems could lead to more consistent QC practices, which may have an indirect impact on improving care. AI systems could also be used during routine equipment QC to predict malfunctions before they occur, thus reducing downtime and providing valuable information for maintenance planning.

(n) Generation of synthetic images: Imaging based AI systems could generate synthetic images for subsequent clinical tasks, for example synthetic CT images derived from MRI for MRI-only workflow in radiation oncology, synthetic post-contrast MRI phases or synthetic mammograms generated from breast tomographic synthesis [40, 44]. This may reduce the need for additional image acquisition, which may involve radiation doses, for these clinical use cases. Other AI systems could be used to predict the need for additional imaging (e.g. a cone beam image for the next fraction of adaptive radiotherapy) or to generate hypothetical clinical cases with images for education and training purposes [45, 46].

(o) Medical report generation: AI systems may be able to generate text based outputs (e.g. radiology reports) from imaging data, ranging from simple structured reports to full image analysis capabilities [47].

3.2. DATA SCIENCE TASK PERSPECTIVE

Underlying each clinical task are one or more data science tasks performed by the imaging based AI system. The data science task dictates both the type of output produced by the machine learning algorithm contained in the AI system and the performance metrics used for its evaluation. System performance is assessed by comparing the system's output against reference data or ground truth.

The most common data science tasks, illustrated using the example of whole brain MRI of patients with brain lesions, are presented in Table 1. Examples of specific clinical tasks and their corresponding data science tasks are presented in Table 2. Further details of such tasks and the associated performance metrics can be found in Ref. [48].

TABLE 1. OVERVIEW OF DATA SCIENCE TASKS IN CLINICAL AI SYSTEMS, INCLUDING USE, OUTPUT AND EVALUATION METRICS FOR WHOLE BRAIN MRI IN PATIENTS WITH BRAIN LESIONS

Data science task	Description of use	Examples of output and reference data	Examples of metrics
Binary and multiclass classification	Used to predict whether a sample belongs to one of two (binary) or more than two (multiclass) classes; for example: — To predict the absence or presence of a brain lesion on brain MRI (binary) — To predict malignancy of a brain lesion (binary) — To predict a treatment-relevant genetic mutation (binary) — To predict tumour staging or lesion subtype (multiclass) — To predict an outcome of treatment of the brain lesion at a defined time point, such as survival at two years (binary) or acute toxicity within six months after treatment (multiclass) — To predict whether or not a treatment plan based on the MRI will pass QC (binary)	Output: — The probability of each class — The most probable class (with uncertainty estimates) Reference data: — The observed class	— AUROC — AUPR — (Balanced) accuracy — Brier score — Matthews correlation coefficient — Sensitivity — Specificity — PPV — NPV — F1 score

TABLE 1. OVERVIEW OF DATA SCIENCE TASKS IN CLINICAL AI SYSTEMS, INCLUDING USE, OUTPUT AND EVALUATION METRICS FOR WHOLE BRAIN MRI IN PATIENTS WITH BRAIN LESIONS (cont.)

Data science task	Description of use	Examples of output and reference data	Examples of metrics
Regression	Used to predict a single (continuous) value of interest; for example: — To estimate patient age from brain MRI — To estimate the extent of atrophy in specific brain regions — To predict a biomarker (e.g. inflammation, hypoxia, tumour-infiltrating lymphocytes) — To estimate dose–volume histogram parameters of a radiotherapy treatment plan based on the MRI	Output: — An estimation of the value of interest Reference data: — The observed (continuous) value of interest	— Mean squared error — Root mean square error — R2 score
Time to event analysis (survival analysis)	Used to estimate the probability and timing of an event occurring in the future; for example: — To estimate the time to a specified clinical event (e.g. death) after treatment — To estimate the probability of tumour progression over time — To estimate the probability of cognitive toxicity over time	Output: — Hazard ratios — The probability over time of an event occurring Reference data: — The time until the event is observed, or the follow-up time if the event is not observed	— Concordance index — Brier score — Time dependent ROC curve

TABLE 1. OVERVIEW OF DATA SCIENCE TASKS IN CLINICAL AI SYSTEMS, INCLUDING USE, OUTPUT AND EVALUATION METRICS FOR WHOLE BRAIN MRI IN PATIENTS WITH BRAIN LESIONS (cont.)

Data science task	Description of use	Examples of output and reference data	Examples of metrics
Segmentation (including localization and object detection)[a]	Used to predict the most likely label for a single pixel, a region of pixels (2-D) or a volume (voxels, 3-D); for example: — To localize the lesion — To segment the lesion in the brain MRI — To segment normal tissues in the brain MRI (e.g. as organs at risk, for radiotherapy planning)	Output: — A binary mask — An attention map — The probability of each label or the most probable label Reference data — Typically, a segmentation created by one or more experts or by a computer algorithm	— Sørensen–Dice similarity coefficient — Hausdorff distance — Average surface distance — Jaccard index (intersection over union)
Text synthesis and generation	Used to convert input data into human readable text; for example: — To automatically generate a neuroradiology report based on the brain MRI — To summarize a brain MRI radiology report by generating a structured report — To translate a radiology report	Output: — Text Reference data: — Typically, clinical acceptability of the produced text (e.g. by comparison with expert reports or by having the quality of the text scored by an expert)	— Acceptance rate — Acceptability score

TABLE 1. OVERVIEW OF DATA SCIENCE TASKS IN CLINICAL AI SYSTEMS, INCLUDING USE, OUTPUT AND EVALUATION METRICS FOR WHOLE BRAIN MRI IN PATIENTS WITH BRAIN LESIONS (cont.)

Data science task	Description of use	Examples of output and reference data	Examples of metrics
Image synthesis and generation	Used to convert input data into an image; for example: — To create a synthetic CT image from brain MRI for radiotherapy planning — To forecast growth of a brain lesion in the absence of intervention — To accelerate MRI acquisition and reconstruction by inferring points in k-space — To improve the quality of the MRI scan (e.g. by removing noise and artefacts) — To introduce artefacts and noise for testing purposes — To perform deformable registration to a reference image	Output: — An image Reference data: — A reference image for one-to-one comparison with the synthesized or generated image	— Mean squared error — SSIM — Perceptual loss based on pretrained networks
Structured data synthesis and generation	Used to generate structured, machine readable data from input data; for example: — To create rare cases — To generate an anonymized structured dataset with the same characteristics as the original dataset — To create test datasets — To correct class imbalance	Output: — Structured data Reference data: — The original (i.e. the input data)	— Statistical characteristics of the original dataset vs the synthetic dataset — Kolmogorov–Smirnov test

TABLE 1. OVERVIEW OF DATA SCIENCE TASKS IN CLINICAL AI SYSTEMS, INCLUDING USE, OUTPUT AND EVALUATION METRICS FOR WHOLE BRAIN MRI IN PATIENTS WITH BRAIN LESIONS (cont.)

Data science task	Description of use	Examples of output and reference data	Examples of metrics
Detection of outliers	Used to identify samples that do not represent a typical sample; for example: — To detect erroneous input data, such as a brain MRI sequence that differs substantially from those used to train the AI system — To detect anatomical abnormalities (e.g. prior surgery) that were not observed during training — To flag scans that might not be of sufficient quality (e.g. because of an artefact)	Output: — The probability or score of being an outlier — The label of being an outlier Reference data: — Samples used to train the AI system and expert opinion on normal cases	— Accuracy — False detection rate — Kolmogorov–Smirnov test
Detection of dataset differences	Used to detect whether the distribution of new samples fits the intended distribution of samples; for example: — To compare demographic and imaging protocol differences between two datasets, such as training vs testing datasets, or between healthcare organizations — To detect bias in cases, batches and algorithms	Output: — The probability of two datasets being different Reference data: — Known differences between datasets	— Jensen–Shannon distance — t-test — z-test — Kolmogorov–Smirnov test — Chi squared test

Note: 2-D: two dimensional; 3-D: three dimensional; AI: artificial intelligence; AUPR: area under the precision-recall curve; AUROC: area under the receiver operating characteristic curve; CT: computed tomography; MRI: magnetic resonance imaging; NPV: negative predictive value; PPV: positive predictive value; QC: quality control; ROC: receiver operating characteristic; SSIM: structural similarity index metric.

[a] Segmentation is in essence a binary or multiclass classification per pixel or voxel. Note that localization of a structure can be viewed as segmentation from a data science perspective.

TABLE 2. EXAMPLES OF SPECIFIC CLINICAL TASKS, THEIR ASSOCIATED DATA SCIENCE TASKS AND HOW PERFORMANCE OF THE AI SYSTEM MAY BE MEASURED

Specific clinical tasks (see Section 3.1)	Data science tasks	Examples of clinical use	Common metrics
Anatomical segmentation (organ segmentation)	— Segmentation	— Treatment planning	— Sørensen–Dice similarity coefficient — Hausdorff distance
Pathological segmentation (tumour segmentation)	— Segmentation	— Follow-up or response monitoring (e.g. RECIST)	— Sørensen–Dice similarity coefficient — Hausdorff distance
Detection and localization (CADe)	— Segmentation — Binary and multiclass classification	— Localization of suspected lesions in breast cancer screening	— Sensitivity — Specificity — FROC
Diagnosis (CADx, abnormality characterization)	— Binary and multiclass classification	— Classification of cancer vs non-cancer — Classification of COVID-19 vs bacterial pneumonitis	— Sensitivity — Specificity — ROC — AUROC
Prognosis: Prognostic prediction (further characterization of known disease)	— Binary and multiclass classification — Time to event analysis	— Classification of aggressive vs non-aggressive subtypes (e.g. in cancer or COVID-19)	— ROC — Concordance index — Log rank test for Kaplan–Meier survival curves
Prognosis: Prognostic regression (time to event analysis)	— Time to event analysis (right censored data)	— Risk prediction — Prediction of time to metastasis or recurrence	— Cox proportional hazard ratio — Concordance index — Log rank test for Kaplan–Meier survival curves
Triage: CADt	— Binary and multiclass classification	— Prioritization of urgent cases to top of reading list	— Correlation to clinical ranking

TABLE 2. EXAMPLES OF SPECIFIC CLINICAL TASKS, THEIR ASSOCIATED DATA SCIENCE TASKS AND HOW PERFORMANCE OF THE AI SYSTEM MAY BE MEASURED (cont.)

Specific clinical tasks (see Section 3.1)	Data science tasks	Examples of clinical use	Common metrics
Triage: Rule out	— Binary and multiclass classification	— Identification of normal samples in screening	— AUROC — Sensitivity — Specificity — PPV — NPV
Triage: Hanging presentations or protocols	— Multiclass classification	— Display layout and organization	— End user preference studies
Treatment response prediction (tumour response assessment)	— Binary and multiclass classification	— RECIST based response prediction	— AUROC
Treatment selection	— Binary and multiclass classification — Text synthesis/generation	— Radiotherapy modality selection (e.g. protons vs photons) — Suggestion of treatment that optimizes curative effect	— AUROC — Sensitivity — Specificity — PPV — NPV — Clinical utility (when the AI system is not a binary or multiclass classifier)
Risk assessment	— Time to event/survival analysis	— Assessment of risk of future cancer, disease recurrence or treatment related toxicity	— Concordance index — Brier score — Time dependent ROC curve

TABLE 2. EXAMPLES OF SPECIFIC CLINICAL TASKS, THEIR ASSOCIATED DATA SCIENCE TASKS AND HOW PERFORMANCE OF THE AI SYSTEM MAY BE MEASURED (cont.)

Specific clinical tasks (see Section 3.1)	Data science tasks	Examples of clinical use	Common metrics
Radiotherapy treatment planning	— Image synthesis/generation	— Synthetic CT generation — Dose or fluence calculation	— Difference maps between real and synthetic images — Downstream effects such as radiation dose calculation differences — SSIM — Multiscale SSIM
Image quality improvement (optimization)	— Regression — Image synthesis/generation	— Optimization of imaging acquisition protocol and tracking of relevant dose metrics	— PSNR — CNR — AUROC — SSIM
Patient specific and equipment QC	— Detection of outliers	— Identification of errors in radiotherapy treatment plan	— PPV — NPV

Note: AI: artificial intelligence; AUROC: area under the receiver operating characteristic curve; CADe: computer aided detection; CADt: computer aided triage; CADx: computer aided diagnostics; CNR: contrast to noise ratio; CT: computed tomography; FROC: free-response receiver operating characteristic; NPV: negative predictive value; PPV: positive predictive value; PSNR: peak signal to noise ratio; QC: quality control; RECIST: Response Evaluation Criteria in Solid Tumours; ROC: receiver operating characteristic; SSIM: structural similarity index metric.

The CQMP needs to be aware of the various sources of mathematical bias in order to ensure the correct, safe and equitable utilization of medical imaging based AI systems. Identifying potential biases is essential for ensuring the performance, fairness and trustworthiness of AI systems [49]. If such biases are not identified and addressed, clinical deployment may compromise the intended function of the AI system, potentially perpetuating inequities through erroneous outputs or biased performance in the intended population.

Data used for the training, testing, QC and clinical use of imaging based AI systems need to be diverse and representative of the true variability of the clinical end point, so that the AI systems perform appropriately in the intended population and can be considered ethical and trustworthy [50]. AI systems can propagate or amplify biases introduced at various stages, from AI system design, data collection and training to AI system deployment, potentially resulting in systematic differences in the treatment of different population groups. Five main areas of possible bias within medical imaging AI/machine learning have been identified: (1) data collection, (2) data preparation and annotation, (3) AI system development, (4) AI system evaluation and (5) AI system deployment, as depicted in Fig. 4 [51]. Within these categories, 29 sources of potential bias

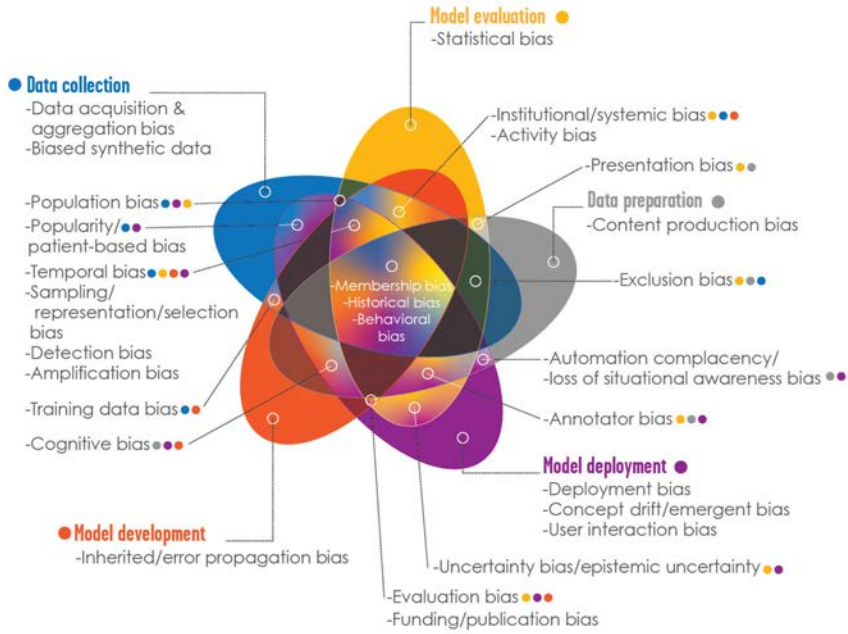


FIG. 4. Potential biases in an AI system (reproduced from Ref. [51] with permission).

have been identified (many of which can affect multiple steps), together with corresponding mitigation strategies [39, 51–54]. Many important sources of bias are related to population characteristics; the population whose data are used to develop and test an AI system needs to be clearly understood and compared with the population in which the AI system is intended to be used. For example, an AI system for assessing breast density that is trained on a dataset consisting only of dense breasts is unlikely to perform well in a population with a broad range of breast densities. Similarly, AI systems developed to diagnose melanoma using training data predominantly from subjects with lighter skin tones might not perform well in subjects with darker skin tones.

Imaging devices and protocols can also be an important source of bias. Although equipment and protocols are implicitly considered part of the 'intended population' description, they might not be explicitly documented as such. For example, the use of different equipment (e.g. an MRI scanner from a different manufacturer), different equipment capabilities (e.g. use of time of flight PET, or 80 kV vs 140 kV X ray tube voltage in CT) or different imaging protocols (e.g. CT convolutional kernels, radiotracer uptake time or MRI sequences) can lead to erroneous outputs if these conditions were not similar during the training phase.

Similarly, changes such as technical developments, shifts in demographics or lifestyle factors and changes in patient care can, over time, lead to a distributional shift (i.e. data drift). As a consequence, the performance of an AI system may degrade over time.

4. ROLES AND RESPONSIBILITIES OF CLINICALLY QUALIFIED MEDICAL PHYSICISTS THROUGHOUT THE CLINICAL IMPLEMENTATION OF ARTIFICIAL INTELLIGENCE SYSTEMS

According to IAEA Training Course Series No. 83 [4], there are six main areas of responsibility involving CQMPs and the clinical use of AI across all medical physics specialities:

(a) Development of technical specifications of the equipment;
(b) Equipment acceptance and commissioning;
(c) Optimization of the physical aspects of medical procedures;
(d) Quality management of the physical, technical and safety aspects;

(e) Education and training of other healthcare professionals;
(f) Scientific research and development.

The subsequent sections summarize the roles and responsibilities of the CQMP relevant to the clinical implementation of AI systems in medical imaging and radiotherapy. These roles and responsibilities are schematically presented in Fig. 5. It is stressed that the CQMP will often be involved as part of a multidisciplinary team consisting of professionals with a wide variety of backgrounds, including medical, IT, purchasing, data science, security and privacy experts. The fulfilment of these roles and responsibilities therefore constitutes

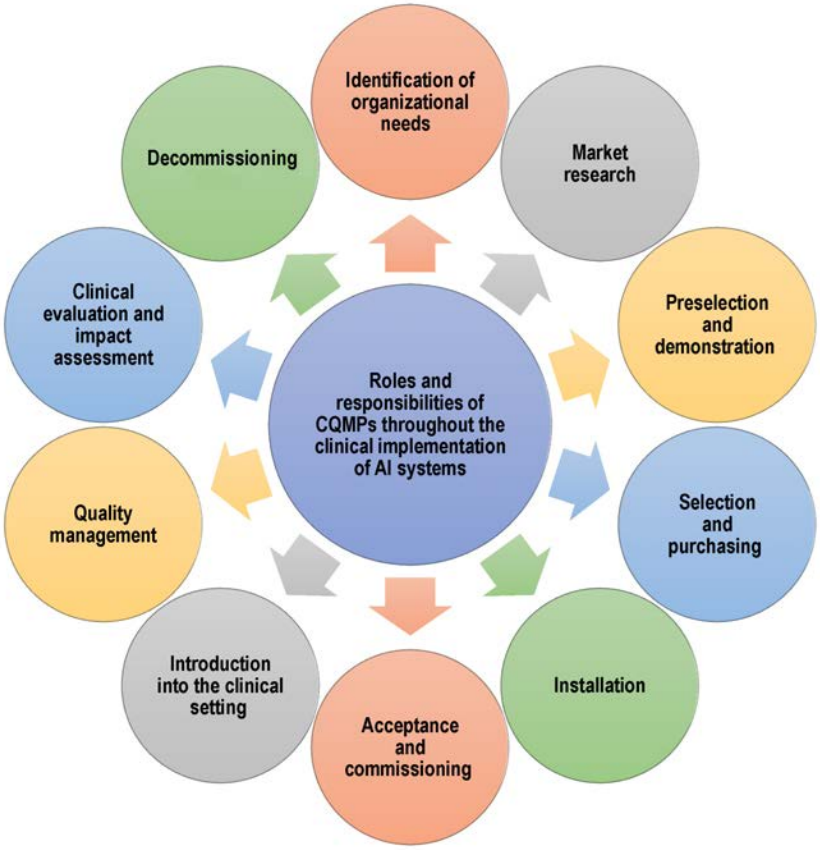


FIG. 5. Schematic representation of the roles and responsibilities of clinically qualified medical physicists (CQMPs) throughout the clinical implementation of artificial intelligence (AI) systems.

a coordinated team effort. The successful implementation of imaging based AI systems relies on project management and leadership skills from the CQMP, similar to those needed for the implementation of other medical technologies, such as digital imaging in radiology [52, 55]. Detailed considerations for the clinical implementation of AI systems from the perspective of the CQMP are presented in Section 5.

4.1. IDENTIFICATION OF ORGANIZATIONAL NEEDS

The CQMP plays an important role in ensuring that the organizational needs to be addressed by an imaging based AI system are clearly identified. The responsibility of the CQMP is to co-lead the team to ensure that the AI system meets the clinical need and to define the relevant functional requirements. The CQMP and their team are expected to consider what success would look like after the AI system is introduced, as well as to clearly establish what the outputs and intended use of the AI system are, what the intended patient population is, what the current performance of human actors is, and what technological solutions are currently used in the clinical care path. All of these considerations inform the definition of the necessary AI system performance. In addition to these functional requirements, non-functional requirements also have to be considered, including workflow considerations, data maturity (i.e. effective management and utilization of data), deployment, autonomy, risks, liabilities, transparency, change management and expectation management. Finally, the CQMP needs to be involved as a technical expert in the multidisciplinary team responsible for formulation of a business case, which includes a cost–benefit analysis of the proposed AI system. For detailed considerations regarding the identification of organizational needs from the perspective of the CQMP, see Section 5.1.

4.2. MARKET RESEARCH

The CQMP needs to provide input as part of the multidisciplinary team that performs comprehensive market research, in which existing AI system solutions are identified and evaluated. Important aspects of this phase include gathering AI system product fact sheets from manufacturers and understanding the different options available in terms of product, licensing and service models. The CQMP and their team are then expected to vet these options to produce a shortlist of relevant potential AI systems. For detailed considerations regarding market research from the perspective of the CQMP, see Section 5.2.

4.3. PRESELECTION AND DEMONSTRATION

The CQMP needs to be involved in the team that evaluates the AI systems from the shortlist generated during the market research phase, identifying those suitable for preselection and demonstration (if possible). In this phase, the CQMP works with other experts in the team, such as IT professionals and end users, to ensure that each AI system is evaluated from multiple viewpoints, including the end user, technical and documentation perspectives. This process results in an in-depth understanding of each AI system's characteristics and the differences between the systems under consideration. For detailed considerations regarding preselection and demonstration from the perspective of the CQMP, see Section 5.3.

4.4. SELECTION AND PURCHASING

The previous phases have provided insights into the organizational needs and the AI systems available in the market. In this phase, one AI system solution is selected and purchased. In the selection and purchasing phase, also known as procurement, the CQMP needs to be involved in defining specifications and requirements, tendering (requesting information and proposals), reviewing and formally approving the submitted bids, evaluating and ranking the received proposals and negotiating terms and conditions with the potential manufacturers.

The CQMP is expected to provide technical input for the formal, concrete list of requirements so that manufacturers can respond with specific offers. This list includes general requirements (e.g. compatibility, scalability, regulatory compliance, staffing, training), functional requirements (e.g. intended use, input data and intended population, output data and range, workflow integration) and technical requirements (e.g. interoperability with current and future solutions, security). Performance metrics against which the AI system will be evaluated also need to be defined at this stage. Additional considerations include QA programme and auditing requirements, as well as maintenance, support, improvements, contract termination and end of life aspects of the AI system. Any data transfer to or processing by third parties, including post-market surveillance by the manufacturer, needs to be understood in order to assess the regulatory impact on the healthcare organization. Finally, a pricing, licensing and subscription model needs to be chosen.

For detailed considerations regarding selection and purchasing from the perspective of the CQMP, see Section 5.4.

4.5. INSTALLATION

Once an AI system has been selected and purchased, the CQMP is responsible for overseeing, with the team, its proper installation. In coordination with IT, security and regulatory professionals, the CQMP needs to conduct pre-installation assessment to ensure that the organization is ready to install the AI system. The installation process, including hardware and software set-up as well as the actual installation, integration and testing of the AI system, then needs to be overseen by the CQMP. Supported by IT and security professionals, the CQMP is responsible for ensuring that all necessary (cyber)security protections are in place. The CQMP then has to lead a team effort to assess the (technical) performance of the AI system through stress testing and evaluation to determine whether the AI system meets the desired and specified technical performance requirements or whether additional IT resources are necessary to achieve the desired performance. After confirming proper technical operation, IT personnel, under CQMP guidance, are expected to establish monitoring and maintenance procedures (see Section 5.5.6). Finally, the CQMP needs to oversee the process of adding and managing the users of the AI system and the training of specific staff as necessary for acceptance and commissioning. For detailed considerations regarding installation from the perspective of the CQMP, see Section 5.5.

4.6. ACCEPTANCE AND COMMISSIONING

The CQMP is responsible for ensuring that acceptance (manufacturer specified criteria) and commissioning (user or regulatory agency specified criteria, or both) of the AI system comply with national and international guidelines and regulations. The CQMP accomplishes this by coordinating interactions between IT experts and the manufacturer during the installation of the AI system, performing the agreed acceptance tests, demonstrating compliance in the specifications as described in the contract with the manufacturer, and leading the final performance evaluation of the system using standardized validation datasets or local data (or both) [4, 56]. The acceptance and commissioning process includes evaluation of the AI system's performance at multiple levels: stand alone computer performance as well as clinical performance as assessed by the healthcare professionals who will serve as the end users of the system. In addition to AI system performance testing, the CQMP ensures routine QC of security, privacy and stability (repeatability), as well as efficient data transfer between the AI system display and its source and target systems, such as picture archiving and communication systems (PACSs) and other hospital information systems (HISs). For acceptance and commissioning, the CQMP is responsible for preparing datasets tailored to the clinical requirements

of the AI system, including manufacturer provided acceptance data and local clinical test sets for commissioning, taking into account both the clinical task and the intended clinical population. For some AI systems, digital or physical phantoms can be used for this purpose. Note that the size and characteristics of the commissioning dataset will depend on the AI system, its intended use, the data maturity and the capabilities of the healthcare organization. The CQMP needs to write a description of the expected clinical workflows before and after the introduction of the AI system, as well as to design routine performance testing (also known as QC testing) for each specific AI system. The goal of routine performance testing and participation in post-market surveillance is to ensure that the AI system continues to operate as expected with respect to both clinical and technical performance, according to the baseline established during acceptance and commissioning, including for recently collected data. For detailed considerations regarding acceptance and commissioning from the perspective of the CQMP, see Section 5.6.

4.7. INTRODUCTION INTO THE CLINICAL SETTING

After acceptance and commissioning, the CQMP, as part of a multidisciplinary team, is responsible for overseeing the introduction of the AI system into the clinical setting. This phase includes setting up a production environment, establishing the workflow of clinical use, releasing technical and clinical documentation, setting operational objectives, defining the scope of use of the AI system and training clinical end users and support staff. Additionally, the team is expected to establish feedback mechanisms, contingency plans, and post-deployment monitoring procedures. The introduction of the AI system into the clinical setting may initially be limited in scope, followed by expansion to broader clinical use. For detailed considerations regarding the introduction of an AI system into the clinical setting from the perspective of the CQMP, see Section 5.7.

4.8. QUALITY MANAGEMENT

The CQMP leads the quality management and risk assessment process for the AI system, in collaboration with the team of experts, and is responsible for preparing a protocol and guidelines for its clinical use. After the AI system has been introduced into the clinical environment, the CQMP and their team are expected to establish and maintain a comprehensive QA programme, including automated or manual QC tests for routine, case specific and ad hoc QC. Additionally, the team needs to implement processes for risk assessment,

incident management and reporting, documentation and participation in post-market surveillance. For detailed considerations regarding the development and implementation of a QA programme for AI systems from the perspective of the CQMP, see Sections 5.6 and 5.8.

4.9. CLINICAL EVALUATION AND IMPACT ASSESSMENT

Once the AI system is in routine clinical use and supported by an active QA programme, the CQMP participates in a multidisciplinary team that evaluates whether the AI system is meeting the identified organizational needs and clinical expectations, through a workflow audit. The audit assesses how the AI system is being used in the clinical environment and identifies factors that are facilitating or hindering its effective use. On the basis of these evaluations, the CQMP and their team are expected to identify possible improvements. Because patients, staff, data characteristics, organizations and AI systems change over time, both objective and subjective clinical evaluations need to be repeated at regular intervals to ensure continued clinical relevance, safety and effectiveness. For detailed considerations regarding clinical evaluation and impact assessment from the perspective of the CQMP, see Section 5.9.

4.10. DECOMMISSIONING

At some point, a healthcare organization may decide to discontinue the use of an AI system, or a manufacturer may declare an AI system to be reaching its end of life, end of support or both. In such cases, the CQMP is responsible for overseeing the proper decommissioning of the AI system. This process typically requires assembling a team with specific expertise, including clinical end users and IT professionals. As an initial step, the clinical processes that will be affected need to be identified and alternative solutions considered. This may include replacing the AI system with another AI system. Once the effect on the clinical workflow is clear, the CQMP needs to consult with the other experts on how to adjust the clinical processes that previously relied on the AI system. In parallel, the CQMP leads the revision of the QA programme to ensure that it appropriately addresses the decommissioning process. If an AI system continues to be used beyond the manufacturer specified end of life or end of support, the healthcare organization and the CQMP are responsible for ensuring that this continued operation complies with the applicable national laws and regulations. For detailed considerations regarding decommissioning from the perspective of the CQMP, see Section 5.10.

5. DETAILED CONSIDERATIONS FOR CLINICAL IMPLEMENTATION OF ARTIFICIAL INTELLIGENCE SYSTEMS FROM THE PERSPECTIVE OF THE CLINICALLY QUALIFIED MEDICAL PHYSICIST

5.1. IDENTIFICATION OF ORGANIZATIONAL NEEDS

This section provides detailed information for CQMPs during the initial stages of acquiring and implementing imaging based AI systems, noting that the CQMP is co-leading a multidisciplinary expert team when performing these tasks. These early phase considerations are broadly applicable to all clinical uses of AI systems but are discussed here specifically in the context of imaging based AI systems. They focus on addressing organizational needs and guiding the CQMP and their healthcare organization in defining the key issues to be addressed in clinical practice. These considerations can be summarized as follows:

(a) Definition of the intended clinical purpose: One of the first steps is to clearly understand why the healthcare organization is considering the adoption of an imaging based AI system. What specific objectives or clinical challenges is the organization seeking to address through the use of AI? Establishing this purpose is critical for setting clear goals for AI system implementation.
(b) Identification of the clinical problem to be addressed: It is important to define the precise clinical problem that the AI system is intended to solve. This involves a rigorous analysis of the healthcare organization's needs that may be amenable to imaging based AI solutions, whether related to diagnostics, treatment planning or other aspects of patient care.
(c) Definition of the intended role of the AI system: Once the clinical need has been identified, the CQMP needs to outline what role the AI system is expected to play in addressing it. What functions or tasks will the AI system perform? How can the AI system enhance or augment the capabilities of the clinical team? Is there an existing structured dataset pertaining to the functions or tasks? Who are the end users for the AI system? How will the AI system be integrated into the clinical workflow to augment the clinical team's work? Clear definition of these aspects helps ensure alignment between the AI system, the clinical workflow and the specific expectations of the healthcare organization.
(d) Selection of an appropriate AI system approach: AI systems differ widely in design, functionality and performance. What type of AI system might be most suitable for addressing the identified clinical challenges? What are the

performance metrics and thresholds that the AI system is expected to achieve to be clinically relevant? Will the AI system function within the technical infrastructure and clinical environment of the healthcare organization?

5.1.1. Functional requirements

Functional requirements define what an AI system has to do within the clinical setting. The CQMP is involved in establishing the functional requirements [4] in collaboration with clinicians, medical radiation technologists, IT staff and other relevant domain experts. To establish these requirements, core questions to be addressed with the end users include the following:

(a) What is the clinical problem to be solved?
(b) What are the specific clinical needs that the AI system can address?
(c) What level of performance is needed for the AI system to be effective, in terms of operating speed and performance of the clinical task?
(d) How will the success of the AI system be measured or assessed at the time of commissioning and during operation?
(e) What are the potential downsides of having an AI system and what are its limitations?

5.1.1.1. Defining the problem

When considering the integration of an AI system into clinical practice, it is crucial to ensure that its adoption aligns with clinical needs and broader healthcare objectives. Healthcare professionals may initially express interest in, or request, a specific AI system. However, such requests do not always reflect the underlying clinical needs. Prioritizing the true clinical need is therefore critical to ensuring that AI system adoption is driven by genuine and meaningful healthcare objectives.

To effectively integrate an AI system, the CQMP is expected to apply a structured approach to identifying the true clinical needs. This needs assessment involves engaging with relevant healthcare stakeholders to understand the specific challenges they face in their daily practices. The process typically includes reviewing current workflows, identifying bottlenecks and assessing where an AI system could potentially provide value. Structured questioning techniques can be used to explore the motivations behind requests for an AI system. For example, stakeholders may be asked why automation is desired, how it improves upon current manual processes, and what impact time consuming tasks have on clinical efficiency and patient care. This process also enables clarification of user expectations and the definition of success criteria, supporting evidence based

decision making that is aligned with organizational needs. By following this structured approach, the CQMP can help ensure that AI system integration efforts are driven by clinical needs and strategic objectives, thereby increasing the likelihood of successful implementation and maximizing the potential benefits of AI technologies for improving healthcare delivery and patient outcomes.

An example of identifying the true clinical need is illustrated by the following hypothetical conversation pertaining to an AI system for automatic segmentation of organs at risk in radiotherapy:

— *Request*: "I want an AI system that automatically segments organs at risk in radiotherapy."
— *Question*: "Why do you want it?"
— *Response*: "Because I saw it at a conference and it seemed to work well."
— *Question*: "Why did it seem to work well?"
— *Response*: "It automates organ segmentation with decent quality."
— *Question*: "Why is this automation needed?"
— *Response*: "Because I currently perform the segmentation manually."
— *Question*: "Why is manual segmentation a challenge?"
— *Response*: "Because it consumes a significant amount of time."
— *Question*: "Why is time consumption a concern?"
— *Response*: "Because I want to see more patients and achieve a better work–life balance."

Through such interactions, the CQMP can explore the various layers of the problem. This process not only helps in accurately defining the problem but also helps reveal user expectations and success criteria. In the above example, the AI system might be perceived as successful by the healthcare professional only if it meaningfully reduces individual workload. At the organizational level, this reduction in workload may translate into improved staff retention as a key objective for the AI system. However, in another healthcare organization facing high clinical demand, success may instead be defined by the ability to provide high quality treatment to more patients. Understanding what both end users and their healthcare organization expect from an AI system is extremely important when considering an AI system.

5.1.1.2. Defining success

The preceding sections consider a systematic process of defining the clinical need, understanding how a suitable AI system can provide actionable outputs and specifying the necessary level of performance [57]. Together, these

steps help formulate a clear understanding of what success means in the context of implementing an AI system.

In this context, the definition of success goes beyond implementation itself; it extends to how the use of an AI system could transform the healthcare provided by the organization. It involves envisioning the 'before' and 'after' scenarios that unfold with the introduction of the AI system into clinical practice. To measure success effectively, several elements can be considered:

(a) Clinical need realization: Success lies in meeting the clinical need identified earlier. Will the use of an AI system effectively address the problem and deliver actionable outputs that make a meaningful difference in efficiency or efficacy?
(b) Actionable output implementation: Success also depends on how well the expected actionable outputs provided by an AI system can be integrated into the healthcare workflow. Will an AI system streamline processes, reduce errors or enhance decision making?
(c) Performance attainment: Success demands that an AI system achieves the necessary level of performance. Whether this involves making healthcare more efficient (e.g. by automating repetitive tasks with high accuracy) or more effective (e.g. by providing robust decision support), an AI system has to meet or exceed the predefined performance criteria. Success is marked by an AI system's ability to consistently deliver reliable results at the necessary performance level.
(d) Realistic expectations: To measure success accurately, it is essential to evaluate whether the initial expectations for the introduction of an AI system are realistic. Are the organizational goals achievable, and can the system realistically meet them? How are the healthcare organization, its processes and staff currently performing? What potential negative effects might arise from having an AI system (e.g. deskilling of the workforce)? Measuring or monitoring success involves aligning expectations with achievable outcomes and acknowledging both the actual impact of the AI system on healthcare processes as well as its shortcomings [58].

5.1.1.3. AI system output and intended use

Each AI system generates output, but the true value of these outputs lies in their capacity to drive actions. Once the problem has been clearly defined, as discussed in the previous section, the crucial question emerges: What will the end user do with the output provided by an AI system — that is, what is its intended use? This question cannot be emphasized enough, as the effectiveness of an AI system hinges on the actions it prompts. Without meaningful actions, the clinical

need remains unaddressed, rendering AI system implementation futile in terms of both time and resources. In general, AI systems can be categorized as those that improve efficiency, those that improve efficacy and those that improve both.

The actions associated with the outputs are closely related to the level of performance needed from an AI system [58–60]. For example, an AI system that predicts the effectiveness of a highly toxic and expensive drug will need a higher level of specificity and reliability than an AI system used for some other applications. This relationship between the actions associated with an AI system's outputs and its necessary performance is further expanded in the subsequent section.

5.1.1.4. Intended population

Imaging based AI systems are developed, trained and tested using data derived from, or related to, actual patients. When such a system is then applied to a patient within the current clinical context of the healthcare organization, its performance is fundamentally tied to the similarity between that new patient and the patients whose data were originally used during the development process. The patient population in which an AI system is to be deployed therefore needs to be aligned with the intended population of the AI system as specified by the manufacturer.

This underscores the importance of considering specific patient population factors early in the implementation process of an AI system. A mismatch between an AI system's training population and the patient population in the healthcare organization may lead to suboptimal performance, inequities and bias, reduced clinical utility and increased risks.

The CQMP therefore needs to thoroughly assess the patient population within which an AI system will be deployed. This includes understanding the clinical context and identifying the patient groups that will benefit most from the system's capabilities. In addition to clinical data — which may encompass demographic information (including ethnicity and socioeconomic status), medical histories, disease profiles, treatment records and patient outcomes — imaging based AI systems require careful consideration of details specific to imaging. These include factors such as imaging modality (e.g. CT, MRI, radiography, mammography, PET, SPECT), equipment manufacturer, model and software version, image acquisition technique (e.g. exposure parameters) and image reconstruction settings (e.g. slice thickness in CT). Patient preparation, positioning and scanning protocol for the imaging procedure (e.g. fasting or time between injection and scan for fluorodeoxyglucose PET) may also be important. Other potentially relevant physical factors that could influence the outcome (e.g. temperature effects in quantitative MRI) also need to be considered [61].

These details can significantly bias and influence an AI system's performance and its ability to generate accurate and clinically appropriate output.

5.1.1.5. Assessing the reference performance

To be successful, an AI system needs to be better than, or at least comparable with, human operators or existing solutions in terms of accuracy, speed or reliability, depending on how success is defined. This necessitates a comparison between an AI system and current performance. Therefore, the performance of current technical solutions or human operators in the healthcare organization needs to be assessed at this stage — that is, the reference performance has to be determined. In the process of such an assessment, the following aspects are important to consider:

(a) Variability in expertise: Human operators, including radiologists, radiation oncologists and medical physicists, exhibit various levels of expertise and experience. This variability can lead to different interpretations and decisions in complex medical scenarios. When evaluating AI systems against human performance, or when evaluating the performance of humans with the aid of AI against that of humans without AI, this range of expertise needs to be taken into account. AI systems might outperform less experienced professionals but might not always match the insights of highly experienced specialists.
(b) Biases and errors: Human decision making in medicine can be influenced by cognitive biases [62], fatigue and other factors that lead to errors or inconsistencies. Although AI systems are not subject to these human limitations, they may have their own biases, derived from the data on which they are trained. Understanding and accounting for these differing bias and error profiles is crucial in assessing the true value of AI systems in medical settings [48, 51].

Defining appropriate reference performance helps in setting realistic expectations:

— Balanced evaluation of AI systems: It is important to set realistic expectations for AI systems by conducting balanced evaluations that compare AI system performance not only against average human performance but also across a spectrum of human expertise. For example, interobserver variation in segmentation among radiation oncologists may be such that segmentation by an AI system may fall within the inherent variability among observers.

This helps identify the specific contexts in which AI systems can be most beneficial.

— Acknowledging current limitations: Recognizing the limitations of humans and current solutions is also crucial. It is important to estimate whether, and to what extent, an AI system might realistically improve efficiency or efficacy. Similarly, it is important to assess the limitations of the AI system to ensure optimal complementarity between human operators and AI systems.

5.1.1.6. Defining the necessary AI system output performance

The actions prompted by AI system outputs determine the necessary quality and performance of the system. Defining the necessary level of performance of the AI system is therefore essential and requires careful consideration. Relevant evaluation criteria may include the following:

(a) Acceptable performance metric levels: For tasks such as detecting anomalies in medical images (e.g. mammograms), what level of accuracy is acceptable to avoid missing critical findings or generating excessive false alarms? Performance metrics such as sensitivity, specificity, positive predictive value and negative predictive value, as well as the desired level of performance, are task dependent and use case specific [48].

(b) Quantitative metrics: For imaging based AI systems such as those used for organ segmentation in CT scans or for classification in pathology image analysis, what specific level of misclassification is acceptable to ensure clinical relevance and utility [40]?

(c) Impactful time savings: What amount of time needs to be saved through automation for it to have a meaningful impact on clinical workflow and efficiency? Assessing the potential time savings helps justify the integration of an AI system into existing processes [63–65].

(d) Error tolerance: Understanding the degree of error tolerance is crucial. For instance, how many AI system generated errors can be accepted, what magnitude of error is permissible, what is the likelihood of detecting an error, and how frequently do healthcare professionals need to intervene to correct those errors?

(e) Consequences of AI system errors: Healthcare providers need to evaluate the potential consequences of AI system errors. For instance, in personalized treatment recommendations, what happens if an AI system output is incorrect, and how significant might the repercussions be?

(f) Current decision quality: It is essential to assess the quality of current decision making processes without AI system intervention, as described in

Section 5.1.1.5. This benchmark helps determine the extent to which an AI system can enhance decision outcomes.

(g) Predictive accuracy: When predicting future outcomes on the basis of current choices, the CQMP needs to work with healthcare professionals to assess clinicians' ability to predict outcomes and how an AI system might complement or improve that predictive accuracy.

(h) Guideline stringency: The level of adherence to clinical guidelines can influence the required accuracy of AI system output. If an AI system causes a deviation from well established guidelines, it needs to be evaluated according to comparable evidence standards, and ideally the use of such an AI system would be reflected in the guidelines.

(i) Evaluating AI system metrics: The CQMP needs to work with healthcare providers to weigh the importance and appropriateness of various metrics for a given task and context (see Section 3.2). These metrics help assess the overall reliability and performance of AI system generated outputs.

Relevant examples of performance levels pertaining to different imaging based AI systems include the following:

— Automated tumour detection: In diagnostic radiology, the required performance level for AI systems that detect tumours in medical images, such as mammograms, CT images or MRI images, is determined by the need to ensure high sensitivity and specificity. This need depends on whether the detection task is part of a screening programme (which mandates high specificity) or a diagnostic task (which mandates high sensitivity). The overall clinical process supported by an AI system is expected to minimize false negatives to avoid missing disease while maintaining a low false positive rate to reduce unnecessary follow-up procedures such as biopsies.

— Organ segmentation: For AI systems that segment organs or structures in medical images (e.g. CT or MRI images), precise performance is essential. It is particularly critical in radiotherapy planning, where accurate organ delineation ensures patient safety and effective treatment. Many metrics are available for assessing segmentations, including the Sørensen–Dice similarity coefficient and Hausdorff distance [66], which can be measured relative to interobserver and intraobserver variability and the time saved for a clinically acceptable segmentation (see Section 3.2).

— Imaging based predictions: Imaging based AI classification systems may predict the likelihood of malignancy, local control, toxicity rates or survival outcomes. Healthcare professionals need to evaluate the potential workflow and outcome consequences of an AI system, especially if detrimental effects

are possible. Acceptable performance levels have to be aligned with current standards for the specific use case.

— Radiotherapy planning: When AI systems predict risks related to radiotherapy outcomes (e.g. toxicity, tumour control), these can inform personalized radiotherapy plans. In such cases, healthcare professionals and their organizations need to consider the robustness of the AI predictions and the implications of any deviations from established treatment protocols.

In conclusion, assessing the necessary level of performance in healthcare AI systems is a multifaceted process that depends on the application type, potential consequences of errors, reproducibility or repeatability of the computer outputs and existing decision making processes. The performance expectations need to align with the specific needs and goals of the healthcare setting, while considering the impact on patient care and outcomes. Setting appropriate quantitative metrics for the expected performance of AI systems may help in subsequent phases of AI system implementation.

5.1.2. Non-functional requirements

The previous sections outline the requirements of the imaging based AI system related to the specific clinical task. This section focuses on how the AI system will perform these functions and how it will be integrated into the healthcare environment. These non-functional requirements play a critical role in ensuring the successful integration, usability and effectiveness of the AI system. In this context, the CQMP needs to look beyond the AI system itself. The maturity of the healthcare organization is a critical factor for success and also needs to be considered.

5.1.2.1. Workflow considerations

Although an AI system may be perceived as a separate entity, it rarely operates in isolation. AI systems are integrated into the clinical workflow (e.g. a PACS), rely on data input from other components in the workflow and generate output to be used in subsequent components, with computer and user interfaces.

For effective workflow integration of an AI system, the CQMP needs to consider the following aspects:

(a) Input requirements: What specific imaging and other data are acceptable as input to the AI system?
(b) Input format: In what format can the input data be provided (e.g. Digital Imaging and Communications in Medicine (DICOM) files)?

(c) Data acquisition interfaces: Is an application programming interface (API) expected for collecting input data?
(d) Data availability: How quickly can the expected input data be prepared and made available to the AI system, given the resources available for data collection and transfer?
(e) Output characteristics: What kind of output is expected?
(f) Output availability: How quickly does the output need to be available (e.g. in real time)?
(g) Output destination: To whom or to what system will the output be provided, and in which format?

Understanding and addressing these integration aspects is pivotal for the successful implementation of AI systems in healthcare. Achieving this necessitates smooth interoperability with existing workflows and systems, while minimizing disruptions and ensuring actual clinical use of the AI system.

5.1.2.2. Data maturity

Data maturity encompasses the organization's readiness and capability to leverage data effectively for the testing and QA of AI systems and to use AI driven output in patient care. The healthcare organization needs to possess the requisite data for the acceptance, commissioning and quality management of an AI system [67]. This pertains not only to the input data needed by the AI system, but also to the availability of local reference standards that can be used to initially evaluate and subsequently monitor the AI system's performance. The collection of such data will also be needed throughout the life cycle of the AI system, as monitoring, continual evaluation and testing remain necessary because patients, data and the AI system itself may change over time.

Introducing AI into healthcare processes therefore imposes demands on both initial and ongoing data collection within the healthcare organization. There might even be contractual or legal obligations to share data, especially regarding incidents and errors, with the AI system manufacturer or regulatory authorities, similar to what is requested in post-market surveillance for pharmaceuticals and medical devices. The CQMP, in collaboration with other professionals, needs to determine whether their healthcare organization has the infrastructure, resources and processes necessary to meet these data collection, transfer and monitoring demands, and whether these processes comply with the regulatory environment to which the healthcare organization is subject.

5.1.2.3. AI system deployment

There are various ways to deploy an AI system, each with its own technical, cost related and operational considerations. The choice of deployment method is intrinsically tied to where and when the AI generated outputs are needed. From a technical perspective, AI deployment options include:

(a) Cloud based deployment: Hosting AI systems outside the healthcare organization on cloud infrastructure offers scalability, flexibility and accessibility. It allows healthcare organizations to harness AI capabilities without the need for extensive on-premises technical resources. Cloud based solutions are well suited for applications that depend on remote access and data sharing. In cloud based deployment, privacy of patient data and regulatory compliance are essential considerations. Aspects that need to be considered include where the specific cloud instance is located; where the data can be sent, processed and stored; and who may have access. Cloud based deployment also requires a stable Internet connection with sufficient bandwidth. Cost estimations need to be performed to assess the sustainability of the AI system.

(b) Software as a service (SaaS): SaaS solutions are a specific type of cloud based deployment [68, 69] that offer a streamlined and user friendly approach to AI deployment. They are typically accessible via the Internet, simplifying access for healthcare professionals across various settings. In a SaaS solution, the manufacturer is responsible for the complete deployment, including support, updates and upgrades, while review and acceptance testing remain with the CQMP.

(c) On-premises deployment: On-premises deployment provides greater control over AI systems and data, ensuring that sensitive patient information remains within the healthcare organization. However, this approach requires sufficient local computing hardware and access to IT system experts who can support installation, maintenance and troubleshooting.

(d) Integration into existing software: AI systems can also be seamlessly integrated into existing healthcare image acquisition and software systems. Examples include a treatment planning system (TPS) integrated into adaptive radiotherapy delivery or an AI driven image reconstruction system that is included in software provided by the scanner manufacturer. This approach ensures that AI outputs become an integral part of the clinical workflow but can make it difficult to separate the impact of the AI system from that of the software system in which it is embedded.

The choice of the preferred deployment method depends on the technical infrastructure and capabilities of the healthcare organization, including computing resources, data storage, network bandwidth, Internet bandwidth, cybersecurity and compatibility with existing systems. Legal and regulatory requirements in the Member State where the healthcare organization is located may also influence feasible deployment options. Finally, the financial aspects of deployment, including initial set-up costs, ongoing maintenance expenses and licensing fees, need to align with the healthcare organization's budget and financial sustainability.

5.1.2.4. Autonomy, risks and liabilities

The autonomy of an AI system is another important consideration. For example, an AI system may support clinicians by generating recommendations, whereas another AI system may be permitted to operate without direct human oversight (i.e. autonomously). Striking an appropriate balance between autonomy and user control is essential to ensure safe and effective AI implementation, and this balance may change over time as AI systems improve and users' confidence in them increases. Moreover, the permitted degree of autonomy may be prescribed by law or by the regulatory requirements of the Member State.

Introducing AI systems with a certain degree of autonomy into clinical workflows alters the existing risk landscape. Such systems may produce additional risks or challenges in patient care, data security or regulatory compliance. The CQMP therefore needs to consider conducting a comprehensive risk assessment both in the phase of the selection process and during the deployment of an AI system (see Sections 5.4 and 5.6). The purpose of this risk assessment is to identify potential vulnerabilities, develop mitigation strategies, ensure patient safety and clarify liability and responsibility in the event of errors.

5.1.2.5. Transparency, explainability and interpretability

The types of imaging based AI system cover a wide spectrum, ranging from relatively simple and interpretable algorithms such as decision trees to complex, inherently non-intuitive models such as deep learning networks. The need for openness, transparency, explainability and interpretability of AI systems in healthcare depends significantly on their intended actions and objectives [59, 60, 70, 71]. For example:

(a) Automating human tasks: When an AI system automates a task historically performed by humans, the primary objective is the system's raw performance. In such cases, transparency may be less of a concern because,

by definition, a human can validate or correct the outputs generated by the AI. An illustrative example is the automatic arranging ('hanging') of images for interpretation by humans, based on a preferred protocol [72]. Other examples are the automatic segmentation of normal tissues for input to radiotherapy treatment planning, segmentation of tumours to measure maximum diameter for subsequent use in response assessment, or automated segmentation of PET images for standardized uptake value (SUV) measurement [4, 73, 74].

(b) Enhancing clinical decision making: When an AI system provides recommendations or makes decisions that influence patient management, treatment or care plans, the demands for interpretability and transparency may increase substantially. For example, an AI system that recommends specific treatment plans for patients with prostate cancer benefits from being able to explain its decision making process. Another example is the diagnosis of breast cancer from mammographic images based on output from an AI system, where the focus is on achieving high diagnostic accuracy (in terms of sensitivity and specificity). Interpretability is particularly critical when the AI system output leads to changes in treatment strategies or therapeutic interventions. In these cases, healthcare professionals are expected to have a clear understanding of the rationale behind the AI recommendations. However, if the performance of a non-interpretable 'black box' AI system far exceeds that of an interpretable alternative, the black box model might still be preferred if it has been sufficiently validated (e.g. in a multicentric clinical trial), incorporated into clinical guidelines and practice, and demonstrated to meet high performance standards. In these situations, the ultimate responsibility for clinical decisions influenced by the AI system is likely to remain with the clinician who incorporates the AI into their decision making process, but the CQMP and their team will need to carefully evaluate the regulatory environment of the Member State.

Recognizing the need for transparency, or the extent to which an AI system is understandable, is a critical consideration during the selection phase. At this stage, the CQMP and the healthcare organization need to assess the nature of the actions resulting from the intended use of the AI system and align them with the appropriate level of performance and transparency. Formal clinical evaluation studies might be necessary to support these assessments.

5.1.2.6. Change and expectation management

Imaging based AI systems, when implemented effectively, have the potential to improve the efficiency or efficacy (or both) of the healthcare process.

However, their introduction inevitably brings changes to staff roles and functions. Effective change management is therefore critically important to support a smooth transition and acceptance of AI driven changes. CQMPs, other healthcare professionals and members of the wider healthcare team may experience a range of reactions to the implementation of an AI system. Some may express concerns about job security, fearing that AI systems could potentially replace their roles. Additionally, they might be concerned about the complexities of new technologies that they might not fully understand. Others might trust AI systems too readily or have unrealistic expectations of the capabilities of AI systems. Such over-reliance can lead to inappropriate use of an AI system, for example, using an AI system approved as a second reader as a primary reader instead, which could result in adverse patient outcomes, such as underdiagnosis or overtreatment. Educating staff and fostering a clear understanding of the AI system's intended role and limitations are essential to address these concerns.

Another challenge of introducing AI systems is potential loss of skill (i.e. deskilling) and expertise, especially if processes previously performed by humans become automated [75]. The CQMP and the healthcare organization need to consider the extent to which deskilling may be permissible, bearing in mind the risks that could arise if the AI system is temporarily unavailable. A contingency plan is needed for both planned and unexpected downtimes during which AI system output would not be available to end users. Such contingency plans to avoid deskilling in the event of AI system failure are well established in other industries, such as the aviation industry, where pilots are required to maintain their skills through regular simulation training.

At an organizational level, assessing readiness for change is crucial. This involves evaluating the existing culture, the historical acceptance of new technologies and the willingness of staff to adapt. Without a supportive organizational culture and staff readiness to embrace change, AI system implementation might not achieve its intended benefits.

In summary, the CQMP and the healthcare organization have to consider what training and education may be needed to foster a clear understanding of an AI system's intended role, capabilities and limitations, as well as to mitigate potential loss of skills as the system is integrated into the healthcare environment.

5.1.3. Business case

Once the functional and non-functional requirements of the desired AI system have been identified, the business case can be defined. The adoption of an imaging based AI system ultimately needs to be underpinned by a compelling and well structured business case. The business case extends beyond the initial

cost of the AI system itself and encompasses a comprehensive evaluation of the entire implementation process within the healthcare organization.

The CQMP needs to be aware that the acquisition and use of an AI system are contingent upon such a business case and that they have a responsibility to be involved as a technical expert in the multidisciplinary team responsible for its formulation. The business case includes the following elements:

(a) AI system costs: The business case includes an estimate of the direct costs associated with the AI system, encompassing purchase or subscription price, licensing fees and ongoing maintenance expenses.
(b) Implementation costs: Equally important are the estimates related to the integration and implementation of the AI system within the healthcare organization, including requirements for routine QC. These may include data capture, data storage infrastructure, IT and computing resources, cybersecurity, clinical user education, staff training, additional staff recruitment, and software development costs such as for fine tuning the AI system to the local patient population or creating interoperability with existing IT systems, such as electronic health record (EHR) systems. Cost considerations may also include the workforce and resources needed to validate an AI system on local data and establish mechanisms for effective post-market surveillance.
(c) Operational efficiency: An effective business case needs to weigh the potential operational efficiency gains or losses associated with AI system adoption. Efficiency may improve through streamlined processes, reduced throughput times and optimized resource allocations. However, efficiencies could also decrease because of premature clinical adoption issues or incompatibilities with existing organizational workflows.
(d) Clinical efficacy: Consideration needs to be given to the potential enhancement of clinical efficacy enabled by the AI system. Improvements in diagnosis, personalization of treatment, reduction of errors and improved patient management can lead to better patient outcomes. But no AI system is perfect; its errors may simply differ from human errors. Therefore, a reduction in clinical efficacy for some patients (e.g. increased false positives) needs to be evaluated against the potential gains for others (e.g. improved sensitivity) or broader gains in efficiency (e.g. faster turnaround times).
(e) Financial impact: The business case has to estimate the potential for reduced operational costs, optimized resource allocation and possible increases in reimbursement due to the use of the AI system, as well as indirect additional reimbursement through increased patient volumes.
(f) Strategic considerations: The business case also has to address aspects related to the expected longevity of the AI system, including scalability,

future upgrades, manufacturer support and potential embedding within existing digital ecosystems.

The central objective of the business case is to holistically weigh the costs against the tangible and intangible benefits of adopting an AI system. Although still qualitative in this early phase of the AI system selection process, the business case provides an initial indication of whether the potential value generated by an AI system justifies the investments and organizational changes it entails. Only when the business case is deemed positive, which is certainly not just a monetary consideration, is the adoption of an imaging based AI system to be pursued.

5.2. MARKET RESEARCH

Identification of potential manufacturers and products for the planned clinical application falls under the responsibility of the CQMP, supported by other members of the multidisciplinary team where relevant. After the clinical need that an AI system is intended to address has been clearly defined, the functional and non-functional requirements have been considered and the business case has been evaluated, the next step is to conduct market research to assess the available AI systems. This includes reviewing the product fact sheet; understanding the available product, licensing and service models; and conducting appropriate vetting of the manufacturer and the AI system. The result of market research is a shortlist of potential AI systems that could be further investigated. The process may begin with a review of key system documentation provided by manufacturers and of relevant scientific literature, to learn as much as possible about the systems. It may also draw on other resources such as registries of regulatory agencies and the experience of existing users [13, 76].

5.2.1. Product fact sheet

The manufacturer of an AI system is expected to provide a concise overview of facts about their product. Details from the fact sheet are important both for assessing which AI system to recommend for purchase and for determining how to deploy, monitor and potentially fine tune the system once purchased. At a minimum, the fact sheet states the claims of the AI system and the regulatory requirements it meets, as well as its intended use, intended users and intended population, together with the basic technical requirements and the version of the AI system to which the fact sheet refers. It is also expected to contain the definition and necessary details of the underlying algorithm and the training dataset characteristics. A model card template is shown in Appendix I.

As an example, the fact sheet of an AI system for diagnosing lesions on breast MRI images describes how well the AI system performs (the claims) when used as a secondary reader (the intended use) to support qualified radiologists (the intended users) for patients with suspected lesions measured using T1 weighted, dynamic contrast enhanced acquisition with a gadolinium based contrast agent on 1.5 T scanners with dedicated breast coils (the intended population). The fact sheet may also state that the system expects DICOM based input from a PACS, that the AI system is intended as a stand alone software application within a Linux based environment and that it requires a specific type of graphics processing unit (GPU) for real time reading.

The first consideration for the CQMP is whether the intended use and intended users of the AI system match the healthcare organization's requirements, and whether the manufacturer's claims are sufficient to meet the organizational needs (see Section 5.1). Figure 6 depicts the relationship between the organizational needs and the AI system's intended use.

The CQMP then needs to assess whether the healthcare organization's target population for the AI system aligns with the intended population described in the fact sheet. In later phases of the selection process, performance levels provided by the manufacturer have to be evaluated in comparison with those assessed in the organization itself.

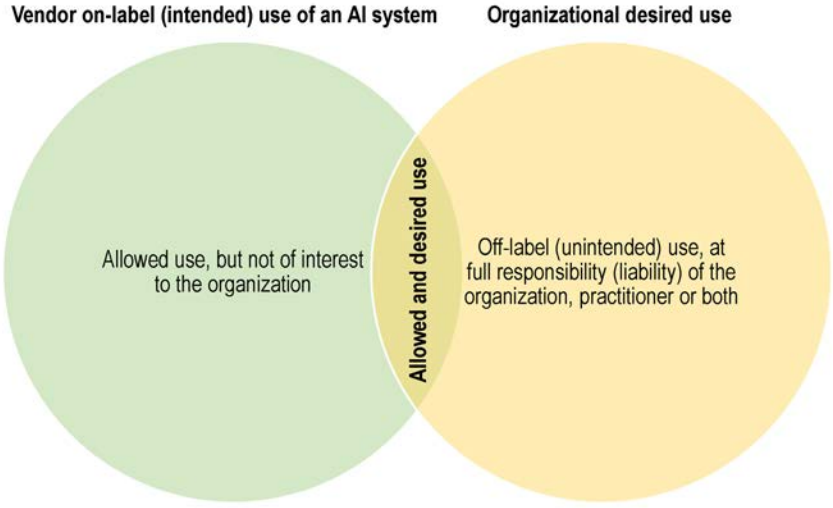


FIG. 6. The relationship between the organizational needs and the artificial intelligence (AI) system's intended use.

Likewise, the CQMP needs to consider the technical compatibility of the AI system and its requirements with the IT infrastructure of the healthcare organization. For proper use of an AI system, knowledge is needed regarding the formats of images and other input data (e.g. DICOM or other formats). The AI system may specify interfaces or means of displaying and transferring AI output (or both) to the end user and methods for sending and storing its output. The AI system may also have specific computational and network requirements, as well as requirements for imaging devices and deployment methods (e.g. cloud based deployment). By assessing technical compatibility, the CQMP can determine whether the AI system could be implemented directly or whether additional infrastructure or software components are necessary.

The CQMP also needs to be aware of any explainability or interpretability aspects and other system characteristics noted by the AI system manufacturer, and to assess whether these align with the organizational needs (see Section 5.1).

5.2.2. Product, licensing and service models

During the market research phase, the CQMP is expected to gather information on product, licensing and service requirements [4, 14], as is done for other medical imaging systems recommended, commissioned and maintained by CQMPs [15, 77].

When assessing an AI system for potential purchase, the CQMP needs to become familiar with the various sales and service models offered by different manufacturers This includes understanding AI system models, potential upgrades and fixes, pricing structures and the long term viability of the manufacturer. Various pricing models may exist, including price per patient, price per read, price per site or annual subscription. In addition, the reliability of the manufacturer's support services and the associated service level agreement costs need to be considered. The total cost of a product across its life cycle, often referred to as total cost of ownership, is also an important factor [78].

The CQMP additionally needs to understand who may need access to data used or produced by the AI system. This includes determining whether the manufacturer claims access rights to the AI system's input and/or output data. Some AI developers may request access solely for QC, whereas others may seek permission to use the data for further training or improvement of their algorithms. Such access may depend on the policies of the healthcare organization's data use committee. It is therefore the CQMP's responsibility to understand the data governance requirements within their organization and country, in order to ensure the compliance of any new installation of an AI system and to address data privacy and security concerns.

5.2.3. Vetting

As part of the market research process, the CQMP needs sufficient resources to undertake due diligence on the manufacturer and the AI system they offer. This involves collecting up-to-date information on the manufacturer's AI portfolio, including available products, market share, reliability, longevity in the medical imaging field and existing customer satisfaction ratings.

5.3. PRESELECTION AND DEMONSTRATION

In the preselection and demonstration phase, the CQMP identifies a shortlist of AI systems for demonstration by their manufacturers, based on the information gathered during the market research phase. The objective of this phase is to gain clearer insight into what each system can offer and to provide input for the subsequent phase of selection and purchasing. Demonstrations may include data provided by the manufacturer but will also need to incorporate data provided by the healthcare organization to allow a realistic assessment of system performance in the local context.

5.3.1. Demonstration preparation

To prepare for the demonstration, the CQMP and their team need to do the following:

(a) Ask the manufacturer about any preparation needed for the demonstration. This may include IT resources, connectivity, time, physical space and the staff who need to be present for a successful demonstration.
(b) Ask the manufacturer to specify the imaging and other data requirements so that the demonstration can be performed using data from the healthcare organization. Using local data allows a more appropriate assessment of the AI system.
(c) Collect the necessary imaging and other relevant data before planning the demonstration.
(d) De-identify the data as needed to ensure the privacy of patients.
(e) Ensure that a non-disclosure agreement, a confidentiality agreement or both are signed between the manufacturer and the CQMP's healthcare organization to ensure that the data used during the demonstration and the system outputs generated remain with the healthcare organization.
(f) Ensure that the demonstration is attended by IT staff, end users and CQMPs.

5.3.2. Demonstration by manufacturer: End user perspective

A demonstration from the end user perspective needs to be conducted with at least the end users and the CQMP present. It is important for the first part of this demonstration to be performed using data supplied by the manufacturer so that the intended use from the manufacturer's perspective is clear.

The CQMP needs to consider and query the manufacturer on the following topics:

(a) The intended use of the AI system according to the manufacturer:
    (i) The clinical question being addressed;
    (ii) The expected AI system output, including the type and format of the output.
(b) The intended population:
    (i) The data used by the manufacturer in the demonstration, including their origin and whether any preprocessing, imputation, cleaning or similar steps were performed by the manufacturer and would be needed during clinical operation.
    (ii) The distribution of data used in the AI system training according to various clinical, demographics and other relevant attributes. The CQMP needs to compare the manufacturer's claimed data distribution (as specified in their regulatory documents) with that of their own healthcare organization.
(c) User friendliness and user interfaces:
    (i) Clarity of the input requirements, including user friendliness, operator convenience and presence of safeguards;
    (ii) Intuitiveness of the computer and user interface (e.g. 'knobology' [79]), output readability, clarity and adjustability;
    (iii) Usability across different skill levels and user roles.
(d) Performance of the AI system:
    (i) How the manufacturer defines and determines the system's performance (metrics and statistical analyses);
    (ii) Demonstration of situations in which the system performs well and less well;
    (iii) Demonstration of situations in which the system is less reliable, if such examples are available;
    (iv) Troubleshooting strategies;
    (v) Explainability, interpretability and transparency aspects of the AI system (e.g. how the AI system reaches its decisions).
(e) Training requirements:
    (i) Availability of a training manual or similar material;

(ii) Training for technical operators (e.g. medical radiation technologists, residents) who select and input images and data;
(iii) Training for end users who incorporate the AI system output into their clinical decision making or procedures.

The second part of the end user demonstration is similar but is conducted using data from the CQMP's healthcare organization rather than data from the manufacturer. This provides additional insight into how well the system performs on previously unseen data and data specific to the healthcare organization. If the demonstration is conducted within the organization, any negative effects on the user experience due to IT infrastructure limitations will be apparent, allowing assessment of the system's data interoperability and interfacing.

Following the demonstration, the CQMP and end users need to prepare a report on the insights gained from the end user perspective.

5.3.3. Demonstration by manufacturer: Technical perspective

A demonstration from a technical perspective needs to be conducted with at least IT staff or technical staff (or both) and the CQMP present. This part of the demonstration stage focuses on how the product can fit into the technical environment of the CQMP's healthcare organization and into the workflow.

The CQMP needs to consider and query the manufacturer on the following topics:

(a) Fact sheet information, including version, release date and user, and whether, for a given patient, dataset or both, these details can be recorded, retrieved and digitally exported (e.g. to an EHR system or PACS).
(b) Manufacturer's vision for how the AI system can best fit into the CQMP's healthcare organization from an IT and technology perspective:
    (i) IT resource requirements of the AI system;
    (ii) Speed or processing time of the solution;
    (iii) Proven interoperability with the organization's HISs, such as the PACS, EHR system, record and verify system, TPS, oncology information system (OIS) and radiology information system (RIS);
    (iv) Proven compliance with the DICOM standard;
    (v) Proven interoperability with image acquisition devices, specifically with the manufacturers and AI systems currently used by the CQMP's healthcare organization.
(c) Fit of the AI system within the CQMP's healthcare organization from a workflow perspective:

(i) How fine tuning of the solution to local organizational data can be achieved, will be achieved or both;
(ii) The proposed QA programme to be implemented by the CQMP's healthcare organization.

(d) Service level aspects:
(i) Service agreements, including uptime/downtime and response time.
(ii) Administrative and/or support interfaces offered by the manufacturer.
(iii) Regularity aspects and methods of updates to the AI system by the manufacturer.
(iv) Manufacturer openness and adherence to standards, specifically:
— Whether data and functionality are separated;
— Whether data generated by the manufacturer can be accessed;
— The database schema or information model used by the manufacturer;
— APIs or other methods of access to data that the manufacturer generates;
— Health information standards to which the manufacturer adheres.

(e) Availability of different modes of the AI system, including:
(i) Clinical mode, for routine clinical procedures;
(ii) Service, maintenance and QC mode, where insights into usage, errors and quality of the solution can be monitored;
(iii) Research mode, offering more configuration options.

(f) Post-market surveillance:
(i) Whether the manufacturer's incident management system is appropriate, easy to use and resource efficient;
(ii) The manufacturer's post-market surveillance strategy and methods, including whether relevant findings and QC results are shared and how.

Following the demonstration, the CQMP and IT staff need to prepare a report on the insights gained from the technical perspective.

5.3.4. Demonstration by manufacturer: Documentation perspective

After the demonstration, the manufacturer needs to supply the CQMP and the team responsible for the AI system demonstration with all relevant documentation regarding the AI system.

At least the following documents are expected:

(a) User manual, including training material on the proper interpretation and utilization of the AI system output.

(b) Administrative and configuration manual.
(c) Technical and installation manual.
(d) Documentation on the core AI algorithm(s) used in the AI system, including a description of:
    (i) Intended use.
    (ii) Developer information.
    (iii) Intended population, including training and test data and potential biases.
    (iv) Performance metrics.
    (v) Limitations, uncertainties and risks.
    (vi) References to scientific studies on the AI system that the CQMP could evaluate using checklists such as:
        — CLAIM[1] [24];
        — FUTURE-AI[2] [27];
        — TRIPOD-AI[3] [20];
        — Other checklists described in Section 2.2.
(e) IT requirements.
(f) Interoperability and conformance statements.
(g) Certifications (e.g. ISO, FDA, CE marking) [80, 81].
(h) Preliminary service level agreements.
(i) Licensing, delivery and purchasing models that the manufacturer offers.
(j) Preliminary cost indication.
(k) Manufacturer strategy, mission and roadmap for the AI system and related solutions.
(l) Contingency plans relevant to company viability and long term support.

The CQMP and their team need to review the manufacturer's documentation for quality, completeness and accuracy.

5.4. SELECTION AND PURCHASING

The CQMP needs to be included in the team that handles the selection and purchasing phase, also referred to as procurement. In many organizations, procurement is a formal process that usually includes at least the following steps:

[1] https://pubs.rsna.org/page/ai/claim
[2] https://future-ai.eu/checklist/
[3] https://www.tripod-statement.org/wp-content/uploads/2019/12/TRIPODAI_checklist.pdf

(a) Defining specifications and requirements;
(b) Tendering (requesting information and proposals);
(c) Reviewing and formally approving the submitted bids;
(d) Evaluating and ranking the received proposals and negotiating terms and conditions with potential manufacturers.

The CQMP needs to be familiar with the procurement process within the healthcare organization. As part of the purchasing team, the CQMP focuses on defining the specifications of the desired AI system and evaluating bids against those specifications. The current section outlines several specification topics to guide the CQMP and relevant IT professionals in drafting a specification document.

5.4.1. General requirements

5.4.1.1. Compatibility

The AI system is expected to integrate seamlessly with relevant existing healthcare IT systems and equipment, including medical image acquisition systems, image processing systems, the PACS, the RIS, image processing workstations, the EHR system, the TPS, the OIS and the record and verify system.

5.4.1.2. Scalability

The AI system needs to be scalable to accommodate future growth in the volume of medical images and other data.

5.4.1.3. Regulatory compliance

The AI system, which can be considered a medical device, needs to comply with relevant healthcare regulations, medical device regulations, information security standards, information standards, data privacy laws and medical standards [82]. The specific compliance requirements depend on the regulatory environment of the Member State and the healthcare organization. Any certifications and regulatory approvals need to be supplied. An overview of regulatory considerations for AI systems in healthcare is provided in Ref. [14].

5.4.1.4. Staffing

The manufacturer of the AI system needs to propose staffing requirements, in terms of both skill sets and levels, for proper operation, local maintenance, QC and support [82].

5.4.1.5. Training

The manufacturer of the AI system is expected to provide comprehensive training materials and support for end users to ensure efficient utilization of the AI system. The manufacturer may also provide specific training for IT administrators, clinical users and technical staff, including the CQMP.

5.4.2. Functional requirements

5.4.2.1. Intended use within the healthcare organization

The AI system's intended use, as deemed by regulatory processes, needs to be aligned with the clinical needs of the healthcare organization. The intended use is to be specified by the CQMP together with the end users. The end user context, expectations of the AI system and expected interactions with the AI system need to be described.

5.4.2.2. Input data and intended population

The CQMP and their team define the input data and intended population, including:

(a) The input data needed for the AI system to fulfil its intended use;
(b) Inclusion and exclusion criteria for fulfilling the intended use, such as the range of imaging techniques, imaging protocols, manufacturers and models of imaging devices, the range of abnormalities and the range of patient characteristics (e.g. age, race, ethnicity, gender) where supported based on the training data;
(c) Input data syntax and semantics, including input file formats (e.g. DICOM or other), API specifications, data quality requirements, accepted sources of image acquisition data and the expected data workflow, including the intended population as noted in regulatory documents, if relevant;
(d) Details on the data used to train and test the AI system.

5.4.2.3. Output data and range

The CQMP and their team define the expected output data:

(a) The expected type of data to be output by the AI system;
(b) Output data syntax and semantics, including output file formats, derived image data, presentation of output data to the end user (user interfaces) and the use of unique identifiers to ensure traceability and integrity of output data.

5.4.2.4. Workflow integration

The CQMP and their team define requirements for integrating the AI system seamlessly into the existing clinical workflow and IT infrastructure to enhance efficiency and reduce manual intervention (see Section 5.1.2.1).

5.4.3. Performance metrics

5.4.3.1. Accuracy

The AI system has to achieve an acceptable level of accuracy in its intended use in its intended population. The CQMP and their team define how accuracy is assessed and monitored and what constitutes an acceptable level of accuracy. The CQMP may refer to earlier phases of the selection process to define accuracy requirements and ensure alignment with the organizational needs (see Section 5.1).

5.4.3.2. Speed

The AI system has to provide its output within a clinically meaningful timeframe. The CQMP and their team define the required speed of the AI system, including its performance when integrated into the healthcare organization's workflow. The CQMP needs to refer to previous phases of the selection process to ensure that the speed of the AI system is sufficient to meet the clinical needs.

5.4.3.3. Errors, uncertainty and risk mitigation

The AI system needs to operate with an acceptable level of error and uncertainty, and any risks need to be mitigated. The CQMP and their team define requirements for a detailed risk analysis based on the healthcare organization's functional and non-functional requirements and on the manufacturer's

documentation. The following aspects need to be considered by the CQMP and their team:

(a) The type and size of errors that are classified as acceptable or unacceptable;
(b) The technical errors documented by the manufacturer, including error numbers, descriptions and solutions;
(c) Any additional technical errors identified during operation of the AI system;
(d) Strategies or methods for mitigating risks that need to be incorporated into the AI system, as well as mitigation approaches to be applied downstream or upstream from the AI system;
(e) The requirements for a detailed risk analysis conducted by the manufacturer;
(f) Methods for reporting suspected errors to the manufacturer and for resolving them;
(g) Methods for receiving notifications of errors reported by other users that have been verified by the manufacturer and could affect the proper functioning of the AI system.

This systematic analysis is often referred to as failure mode and effects analysis (FMEA). Risk analysis in general is further described in Section 5.6.7.

5.4.4. Non-functional requirements

As described in Section 4.1, the CQMP has a role in defining non-functional requirements that are critical for ensuring the successful integration, usability and effectiveness of the AI system in the clinical environment, which have to be considered when identifying the organizational needs (see also Section 5.1.2).

5.4.4.1. Architecture

The AI system needs to fit into the IT architecture of the healthcare organization. Depending on the organization, this architecture may include elements such as on-premises versus cloud based deployment, various operating systems and versions, connectivity and related infrastructure. The CQMP has to request that members of the IT team provide these requirements. The manufacturer has to describe how their system meets, or does not meet, the architectural requirements.

5.4.4.2. Operating requirements

The CQMP and their team need to define the technical requirements for operation of the AI system, on the basis of the functional and non-functional requirements and the manufacturer's documentation. These operating requirements include:

(a) Basic infrastructure requirements, such as physical space, power supply and cooling;
(b) Installation and operation requirements, including Internet and network speeds, network settings, computing, storage, operating systems and integration software;
(c) Fine tuning requirements, including reference data, computing and storage;
(d) QA requirements, including requirements for reference data, computing, storage and operating systems;
(e) Energy efficiency and sustainability of the AI system [83, 84].

5.4.4.3. Information security

The CQMP and their team define requirements related to information security, usually on the basis of national regulations [82, 85]. At a minimum, the following information security requirements need to be met by the manufacturer:

(a) Secure handling and secure communication of data during:
    (i) Operation of the AI system (e.g. how data are handled by the manufacturer when an AI system is deployed in the cloud);
    (ii) QC activities performed by the manufacturer;
    (iii) Archival procedures;
    (iv) Services and upgrades.
(b) Management of access rights to the AI system for clinical use, QC and other purposes.
(c) Responsibility and liability of the manufacturer for security breaches and incidents arising from the use of the AI system.
(d) Information security certification.

5.4.4.4. Interoperability

The AI system has to support common industry standards for syntactic and semantic interoperability, facilitating integration within relevant healthcare systems (e.g. DICOM, HL7 FHIR). The CQMP may, together with IT personnel, add organization specific interoperability requirements.

5.4.5. Quality assurance and auditing

5.4.5.1. Quality assurance

QA of the AI system relevant to the use case, in terms of consistency and performance level, has to be possible [86]. The QC test results need to allow for effective and efficient interpretation by the personnel responsible for QA. The CQMP needs to specify what routine QC activities are planned as part of the QA programme (see Section 5.8.1).

5.4.5.2. Audit trail

The CQMP and their team need to ensure that the AI system provides an audit trail of all input data, output data and errors. An audit trail ensures that the healthcare organization can audit interactions between the AI system, users and data for accountability and traceability.

5.4.6. Maintenance, support and improvements

5.4.6.1. Updates and upgrades

The manufacturer needs to provide regular updates and upgrades to ensure that the AI system remains current and effective. As described in Section 5.6, the CQMP and their team need to verify the performance level of the AI system before and after updates or upgrades, along with its interoperability with relevant IT infrastructure, using, for example, reference data composed of both manufacturer and organization test sets. The CQMP and their team therefore need to define the requirements for updates and upgrades, including their regularity, the minimum lifespan of updates and upgrades, the need for the manufacturer to set up a testing environment prior to updates, and the availability of rollback to previous versions.

5.4.6.2. Support, responsibilities and liabilities

As part of the selection and purchasing (procurement) process, the healthcare organization and manufacturer need to agree on technical and end user support, as well as on a service level agreement that specifies uptime guarantees, user training and support, recovery point and recovery time objectives, disaster recovery, data recovery, support hours and response times. The agreement has to define a clear separation between the responsibilities of the manufacturer

and the responsibilities of the healthcare organization and the end user. During specification, the CQMP and their team need to specify these details.

5.4.6.3. Incident, change and release management

As part of the selection and purchasing (procurement) process, the CQMP and their team need to specify how they expect the manufacturer to handle incident reporting and the approach to change and release management. The manufacturer may allow the healthcare organization to influence future versions of the AI system; however, if this involves the use of the organization's patient data, both the healthcare organization and the manufacturer need to establish protocols and precautions to comply with data use agreements (see Section 5.4.7).

The CQMP and their team also need to consider the following:

(a) Whether they need the manufacturer to ensure that developers and other experts are available for issues and incidents that cannot be handled by the regular support team;
(b) How they expect the manufacturer to handle the decommissioning of the AI system with the healthcare organization, whether the decommissioning is initiated by the manufacturer owing to lack of support or by the healthcare organization (see Section 5.10).

5.4.7. Data transfer and processing

5.4.7.1. Use of data from the healthcare organization by the manufacturer

Within the initial agreement, the manufacturer will likely request the use of data for QC testing of the system. However, if the manufacturer intends to use data from the healthcare organization for purposes such as processing the data in the AI system, improving the AI system or other manufacturer products, incident management, additional QC and management, or research and development, a separate agreement between the manufacturer and the healthcare organization is necessary. This agreement has to address privacy, information security, cybersecurity, intellectual property, ownership of results and reimbursements. The manufacturer has to indicate whether the healthcare organization's data will be processed, the conditions under which such processing takes place, and how they will comply with data laws and regulations, as well as the organization's data use policy, if relevant (see Section 5.2.2).

5.4.7.2. Use of data from the manufacturer by the healthcare organization

Acceptance, routine QC, and re-acceptance of new versions of the AI system require reference data from the manufacturer and from the healthcare organization, or from other sources, as described in Sections 5.6 and 5.8. The manufacturer has to supply reference data of sufficient quality and diversity to test the AI system. The CQMP needs to create in-house reference standards with the help of the manufacturer. These requirements therefore need to be specified by the CQMP and their team as part of the system specification. The post-market surveillance process facilitates information sharing between the manufacturer and the user (see Section 5.3.3).

5.4.8. Contract termination and end of life

5.4.8.1. Contract termination

The healthcare organization and the manufacturer of the AI system need to specify the conditions under which their contract can be terminated. This includes the rights of the healthcare organization to terminate the contract unilaterally if the manufacturer does not meet contractual obligations. The permitted reasons for contract termination need to include the AI system's inability to achieve its specified performance or interoperability, the manufacturer's inability to provide appropriate support, and major security concerns. Contract termination may have financial consequences, which need to be specified by both the manufacturer and the healthcare organization.

5.4.8.2. End of life and removal

The healthcare organization and the manufacturer of the AI system need to agree on guarantees regarding upgrades, updates and the supported lifespan of the current version of the AI system. The manufacturer also has to specify whether the AI system can continue to be used after end of life, end of support or contract termination. In some countries, the use of systems that have reached end of life may be prohibited because the healthcare organization cannot obtain upgrades, updates or fixes, which could affect its ability to maintain regulatory approval, if relevant.

5.4.9. Pricing, licensing and subscription models

The healthcare organization and the manufacturer need to agree on the pricing structure, including initial set-up costs, licensing fees, subscription

models, service and support fees, update and upgrade fees, hosting and processing fees and any additional charges. The CQMP and their team need to provide input on whether the pricing structure, licensing model and subscription model fit the organization's needs and budget.

5.5. INSTALLATION

Once an AI system has been acquired through the selection and purchasing process described in the previous section, it needs to be integrated into the clinical environment. This section outlines the essential aspects for the successful installation of an AI system. During the installation process, the main aims are to ensure that the AI system functions from a technical perspective, that it can operate securely and that procedures are put in place for monitoring the system after installation. Following installation, the AI system will undergo acceptance and commissioning with respect to both technical and output performance levels, which is the subject of the next section.

A step-by-step guide to the installation process, highlighting the relevant tasks, is provided below. In most situations, the installation itself will be carried out by IT personnel and the manufacturer, and the role of the CQMP is to oversee and coordinate the process. Specific instructions for the installation process need to be documented in the manufacturer's manual. These instructions need to be followed by the CQMP and their team.

5.5.1. Pre-installation assessment

The selection and purchasing process defines several requirements that need to be met by the AI system. During the pre-installation assessment, the CQMP, with the support of other team members (including IT personnel) and the manufacturer, needs to develop a concrete plan for integrating the AI system into the IT architecture and clinical environment, based on the documentation for the purchased version of the system. Although these requirements were defined during the purchasing phase, they need to be verified again. The following tasks need to be carried out:

(a) Conduct a thorough assessment of hardware and software prerequisites.
(b) Verify integration of the AI system with the existing hospital network and IT architecture.
(c) Verify compatibility with existing HISs, such as PACSs.
(d) Verify regulatory compliance and alignment with data protection standards.

5.5.2. Hardware and software set-up

The AI system may need specific additional software and hardware to operate. The CQMP, together with the IT team and the manufacturer, needs to perform the basic hardware and software set-up necessary for integration of the AI system into the hospital workflow. The following tasks need to be carried out:

(a) Install or upgrade hardware components as needed.
(b) Set up the necessary operating systems, libraries and supporting software.
(c) Configure network settings for optimal connectivity, including firewall configurations and other security settings.

5.5.3. Artificial intelligence system installation, integration and testing

After the necessary software and hardware have been set up, the AI system needs to be integrated and tested to ensure basic functionality. The aim is to prepare the system for the acceptance and commissioning process. Depending on the AI system, the CQMP and their team need to carry out the following tasks:

(a) Set up a designated software environment for testing that is not connected to the clinical environment or the organization's infrastructure. Depending on the healthcare organization, this separation of environments may be available through a DTAP (development, testing, acceptance and production) environment or a staging and production environment.
(b) Install the AI system on designated servers or computing units.
(c) Integrate the AI system with relevant HISs including any preprocessing and postprocessing of information entering or leaving the system.
(d) Integrate the AI system with medical imaging devices and monitors for real time testing and clinical use.
(e) Validate data exchange between the AI system and other hospital systems.
(f) Install, configure (including user preferences) and test end user interfaces within the AI system, in external HISs or both; for example, in image viewing devices; as a report generated by the AI system in the EHR; or as an output image in the PACS, TPS or RIS.
(g) Use standard healthcare communication protocols, such as HL7 (HL7v2, HL7v3, FHIR) and DICOM.
(h) Conduct basic testing of the integrated AI system ('Does it run?') using sample datasets (see Section 5.6).
(i) Validate the basic clinical workflow with the integrated AI system through test cases and simulations.

If applicable, verify that multiple concurrent installations and active users of the AI system within the healthcare organization (including remote locations) do not affect system performance.

5.5.4. Security implementation

The installation of the AI system in the clinical environment needs to be aligned with appropriate security measures to establish robust cybersecurity and ensure patient data protection. The tasks relevant to this aspect are:

(a) Consult with IT, information security and cybersecurity staff on local security policies.
(b) Implement data encryption, de-identification and anonymization protocols, as appropriate or necessary.
(c) Set up access controls and user authentication mechanisms (e.g. single sign-on).
(d) Conduct vulnerability assessments and address any identified security gaps.

5.5.5. Hardware and software stress testing

To ensure optimal technical performance and speed of the AI system in real world clinical scenarios, the CQMP and their team need to:

(a) Check and optimize computing resource utilization (e.g. storage, network, memory, processor) for the specific task that the AI system will perform.
(b) Check and optimize external systems and connections with which the AI system is integrated, such as a PACS or EHR system.
(c) Assess system technical performance in terms of speed, storage and connectivity during usage that resembles actual clinical conditions and loads.
(d) Identify bottlenecks in the technical resources of the integrated AI system.
(e) Establish baseline technical performance metrics for the AI system, such as the time needed to process a fixed number of images, restart time and user interface response time.

AI task performance testing is discussed in Section 5.8 and is not included in the above considerations.

5.5.6. Establishment of monitoring and maintenance procedures

When installing an AI system into the clinical environment, it is essential to establish mechanisms and procedures for monitoring and maintaining the system from the outset. This is described in more detail in Section 5.8. For example, the CQMP and their team need to determine the details of asset tracking (e.g. with unique identifiers such as barcodes or radio frequency identification tags), documentation and records (e.g. purchase date, installation date, history of maintenance and any modifications made), asset life cycle management and security measures (including access controls, authentication mechanisms and physical security protocols). With regard to these aspects, the CQMP, in consultation with healthcare organization officials and the manufacturer, needs to:

(a) Enter the AI system into the appropriate service and support management systems (e.g. inventory, asset, incident, quality, configuration and manufacturer management systems).
(b) Establish procedures for remote service (e.g. diagnostics and maintenance) by the manufacturer.
(c) Define appropriate system management roles (e.g. system owner, application manager, service manager).
(d) Implement or connect to monitoring tools for tracking AI system performance (both technical and clinical AI performance metrics), including repeatability and reproducibility using a consistent set of input data.

5.5.7. User management

Different users interact with the AI system for different purposes. For example, end users may use the system to perform a clinical task, and the CQMP may carry out QC tasks on the system. The different types of user therefore need different access levels and permissions. Effective user management is crucial for the safe and successful integration and use of AI systems in clinical settings. The CQMP and IT personnel need to determine how user management for the AI system will be structured. This section outlines the processes involved in creating and managing user accounts for AI systems and emphasizes the importance of training and onboarding healthcare professionals in the use of such systems. Depending on the characteristics and capabilities of the AI system, on the extent of user management needed and on the local organizational and security policies, user management can include the following aspects:

(a) User categories and roles:
    (i) Define the potential user categories and roles, where appropriate and depending on the AI system and its use case. Such user categories might include:
        — Administrator and lead computer system manager: This user has the highest level of access and is responsible for establishing, maintaining, upgrading and backing up all other user accounts on the AI system.
        — Technical support users: These users may conduct routine QC of the AI system, as well as maintenance testing and commissioning tasks.
        — Clinical end users: These users typically use the AI system in clinical workflows and incorporate its output into their clinical decision making.
        — Clinical support users: These users typically work with the AI system in the clinical environment and may be responsible for providing correct intended use input to the AI system (e.g. images and associated data).
        — AI system manufacturer users: These users are manufacturer personnel who may need to routinely (and perhaps remotely) log into the system for QC testing. They are provided with only strictly limited access to patient data, including images.
        — Other users: There may be other user categories not included in the above list (e.g. researchers, students). And in some situations, user categories might not be appropriate, for example if the output of an imaging system goes directly into another AI system as input, without clinical end user or support user interaction. In such cases, a specific system account would be needed instead.

(b) User registration process:
    (i) Develop a user registration process for the various end user categories needing access to the AI system.
    (ii) Capture essential information (e.g. name, role, department, contact details) during registration or integrate the registration process with other hospital systems.
    (iii) Ensure registered users have completed the training described in Section 5.5.8.

(c) Access levels and permissions:
    (i) Define appropriate access levels and permissions on the basis of user roles and responsibilities. It is important to avoid granting unnecessarily broad permissions to all users.

(ii) Implement role based access control [87] to ensure appropriate access to AI system functionalities.

(d) Authentication mechanisms:
  (i) Implement secure authentication mechanisms, such as multifactor authentication, to safeguard user accounts.
  (ii) Integrate authentication with existing hospital systems for seamless access (i.e. single sign-on).

(e) User account provisioning:
  (i) Establish protocols for provisioning user accounts.
  (ii) Ensure timely activation and deactivation of accounts based on changes in personnel or roles.

(f) Audit trails:
  (i) Maintain audit trails for user account activities.
  (ii) Track logins, access attempts and modifications to user permissions for security and compliance purposes.

(g) Password policies:
  (i) Enforce strong password policies to enhance account security [88].
  (ii) Prompt users to update passwords at appropriate intervals, and implement account lockout mechanisms.

(h) User account deactivation:
  (i) Establish procedures for prompt deactivation of user accounts upon personnel changes or role modifications.
  (ii) Ensure data associated with deactivated accounts are securely archived or transferred.

5.5.8. User training and onboarding

Various staff will interact with the AI system. Some may be end users who use the system in clinical workflows and therefore need to be trained to use the AI system safely and effectively. Other users also need training, such as maintenance and support staff, QA staff, IT staff and the CQMP. Non-clinical staff, such as researchers or students, may also need to be trained to use the AI system. It is important that the CQMP is part of a team, also including the manufacturer, that establishes a robust training programme for all current and future staff involved with the AI system [4, 15].

Designated clinical end users and supporting staff need to be trained during this phase to ensure a successful acceptance and commissioning phase. Other users may be trained after acceptance and commissioning but before clinical introduction and the start of the QA programme. A training programme may include the following components:

(a) Orientation programmes:
    (i) Incorporate AI system training into orientation programmes for new staff.
    (ii) Provide an overview of the AI system, its functionalities and its impact on clinical workflow.
(b) Customized training modules:
    (i) Develop customized training modules based on the specific user roles.
    (ii) Tailor training content to the needs and expectations of diverse user groups.
    (iii) Develop alternative training modules to ensure continuity of clinical services when the AI system is unavailable.
(c) Hands-on training sessions:
    (i) Conduct hands-on training sessions to familiarize users with the user interface (if applicable) and the functionalities of the AI system.
    (ii) To build confidence and vigilance, allow users to practise using simulated scenarios, including known errors and failures, false positive cases and false negative cases.
(d) User proficiency testing: Develop tests to verify that users have achieved the necessary level of competence to use the AI system, relative to system input, output or both.
(e) Documentation and resources:
    (i) Provide comprehensive documentation, manuals and other resources for users to reference independently.
    (ii) Include user guides, frequently asked questions and video tutorials to support continuous learning.
(f) Continuous learning opportunities:
    (i) Establish a framework for continuous learning and skills development.
    (ii) Offer periodic refresher courses and updates on new features or improvements to the AI system.
(g) User feedback mechanisms:
    (i) Implement mechanisms for users to provide feedback on the usability and effectiveness of the AI system.
    (ii) Use feedback to drive iterative improvements and address user concerns.
(h) User support services:
    (i) Offer dedicated user support services, including a help desk and user forums.
    (ii) Ensure prompt responses to user queries and issues to maintain a positive user experience.

(i) Training for downtime scenarios: Provide training for users of all types to ensure that they maintain competence to perform the clinical workflow when the AI system is non-functional or unavailable (i.e. to avoid deskilling).

As mentioned, the training programme needs to be tailored to the user roles. For example, technical staff (e.g. the CQMP, IT staff, application specialists) need to be equipped with the knowledge and skills necessary to maintain and operate the AI system effectively. Such training may include:

— Installation and maintenance of the AI system;
— Routine clinical use and workflow management;
— QA procedures and QC tests;
— Identification of errors and troubleshooting strategies;
— Frequently asked questions from users.

Clinical end users and clinical support users need a thorough understanding of how to interact with the AI system to discharge the intended clinical task, with particular emphasis on what constitutes an allowed input. The following elements need to be part of their training programme:

— How to use the AI system;
— Which data (e.g. images) satisfy the approved claims and intended population for the AI system;
— The limitations of the AI system;
— How the AI system output is to be interpreted and used in the clinical workflow according to the approved claims, for example whether the AI system is used as a secondary reader, as a concurrent reader, for triage, to rule out or as an autonomous device;
— Whom to contact for support, questions and concerns.

A registry needs to be maintained to track which users have successfully completed their training and to record whether and when they need to be retrained to maintain competence (see Section 5.5.7).

5.6. ACCEPTANCE AND COMMISSIONING

As stated in IAEA Human Health Series No. 25 [15],

> "Following the installation of new equipment, CQMPs are responsible for specifying the basic standards to be applied for its acceptance and subsequent commissioning. They ensure that all units and systems function according

to their technical specification and provide advice on any deviation of equipment performance from acceptable criteria".

In the context of AI systems, the CQMP ensures that the acceptance and commissioning phase is carried out in accordance with national and international guidelines and legislation, coordinates discussions between experts from different health professional groups and the manufacturer on the installation of the AI product, performs the agreed acceptance tests, and verifies compliance with the specifications as described in the contract with the manufacturer. The CQMP also leads the final performance evaluation of the system using standardized validation datasets, local data or both. In the event of system or model updates or upgrades, the CQMP will lead a new round of acceptance and commissioning [4].

The previous section, on installation, aims to ensure that the AI system is able to function within the clinical setting. Acceptance and commissioning ensure that the AI system meets both known and expected performance levels, while routine performance testing aims to ensure ongoing performance (see Section 5.8). Thus, the acceptance and commissioning process includes performance evaluations of the AI system with respect to stand alone computer performance (technical) as well as clinical performance as assessed by the healthcare professionals (e.g. radiologists, medical physicists, radiation oncologists) who serve as the end users of the device. In the acceptance process, the AI system is evaluated against the specifications set by the manufacturer, using both a dataset provided by the manufacturer and an organizational dataset. In the commissioning process, the AI system is configured and evaluated according to the needs of the healthcare organization. This includes evaluation of the AI system with the end users and the data of the healthcare organization. The acceptance and commissioning processes also provide baseline performance metrics for routine QC within the QA programme (see Section 5.8).

5.6.1. Evaluation metrics

Evaluation metrics depend on the task, on the intended use and population as detailed in the manufacturer's specifications and on the specific clinical intended use and population defined by the healthcare organization [40]. The information used to determine the appropriate evaluation metrics may also include the device claims as approved and cleared by the appropriate regulatory body. Some AI systems aim to enhance the efficacy of clinical decision making, and others aim to optimize the efficiency of the clinical practice workflow.

The evaluation metrics may be quantitative or qualitative, depending on the type of AI system. For example, quantitative evaluation metrics may be used for an AI system that predicts the likelihood of malignancy or overall survival, while

qualitative metrics may be used for an AI system that generates unstructured text. Depending on the output, evaluation can be performed in real time for a given use case or retrospectively on a cohort of patients with a known reference standard (ground truth) established within the healthcare organization (see Section 3.2).

Efficiency metrics may also include measures of fidelity, speed of throughput and user friendliness. An example of an efficiency enhancing AI system is one that retrieves images from the acquisition system as well as prior images from the PACS and organizes their presentation for the radiologist responsible for the clinical interpretation.

In addition to AI system performance tests, routine QC is also necessary to ensure security, privacy, stability (repeatability) and efficient transfer of data between the PACS or other HISs and the AI system display (see Section 5.8).

5.6.2. Preparation for acceptance and commissioning

Technical preparation for acceptance and commissioning takes place during installation. Like the installation process (see Section 5.5), the acceptance and commissioning process requires a designated software environment for testing that is separate from the clinical environment. This separation ensures that any issues arising during acceptance and commissioning do not affect ongoing clinical workflows. Depending on the healthcare organization, this separation of environments may be available through a DTAP environment or a staging and production environment. The CQMP and their team need to consider setting up such an environment if one is not already available.

The CQMP and their team need to prepare datasets for acceptance and commissioning that are specific to the requirements of the AI system [86, 89]. For acceptance, datasets provided by the manufacturer are used. For commissioning, representative data and users from the healthcare organization itself are also expected to be included. The CQMP needs to understand the limitations set by the manufacturer in terms of intended use and population, as well as any regulatory (e.g. FDA, CE marking) claims or certified use, if available, of the AI system. In addition, the CQMP needs to assess how the AI system handles off-label use cases, in which data and tasks differ from the intended use, population or claims, as well as how it handles erroneous and low quality data and input. The CQMP and their team may need to create a dataset (or use another representative dataset) to test how the AI system handles such data.

5.6.3. Acceptance

During acceptance testing, data provided by the manufacturer are used to test whether the AI system, as installed in the healthcare organization, meets

the specifications provided by the manufacturer. Any failure to meet these specifications needs to be resolved in collaboration with the manufacturer.

5.6.4. Customization, localization, fine tuning and continuous learning

Depending on the AI system, the manufacturer and the applicable regulatory approval, the AI system may be customized (fine tuned) to the context of the healthcare organization. If customization is necessary or desirable, the CQMP and their team need to coordinate closely with the manufacturer. The typical process involves using data from the healthcare organization to modify the AI system so that it performs better within the organization. After this customization, the acceptance process (using manufacturer supplied data) needs to be repeated. If customization is successful, the performance of the customized AI system is expected to remain broadly similar on the manufacturer supplied dataset. Major differences resulting from the customization could indicate that the data and context in which the AI system was originally developed by the manufacturer differ significantly from the organizational data and context, which may introduce risks. Major differences due to customization may also prevent the use of the customized AI system altogether, depending on the applicable regulatory framework. This needs to be verified against the regulatory documentation.

During the subsequent commissioning process, the use of data that were already used to customize the AI system has to be avoided. Re-using data will lead to overly optimistic performance during commissioning. Using completely new and previously unseen data during commissioning avoids this issue.

5.6.5. Commissioning

Creating a dataset for proper commissioning calls for careful consideration. The AI system needs to be commissioned for all clinical settings and workflows in which it is intended to be used, noting the data types and clinical settings for which the AI system is not to be used (off-label or more limited on-label use as decided by the organization). Accordingly, the test dataset for commissioning needs to cover different sources of input data (e.g. different scanners, scanning protocols, organizations, and preprocessing steps between the scanner and the AI system), different patient populations (e.g. different ages, genders, races/ethnicities and diseases), different clinical settings (e.g. emergency, planned care, outpatient, self-care), different levels of problem solving (e.g. easier cases, difficult cases), different destination of output data (e.g. different monitors, display devices and information systems) and different end users (e.g. experienced vs novice users, specialist physicians vs general practitioners). The test sets used during commissioning have to be large and diverse enough to enable a clear

understanding of how the AI system performs across a wide range of settings and to ensure that the assessment uses data representative of the diversity of data in the given clinic. It is important to note that these settings may also change over time, and the CQMP needs to assess whether such changes can be anticipated and, where possible, test their potential impact.

The CQMP has to ensure that the commissioning dataset does not contain any data that were used for customization or acceptance, nor any other data from any of the patients whose data were used in those processes.

The size of the commissioning dataset depends on the AI system, its intended use, the maturity of the available data and the capabilities of the healthcare organization. Clear criteria defining the appropriate dataset size are therefore difficult to establish.

The CQMP may consider the following approaches:

(a) In the case of an AI system used primarily for higher efficiency: Prospectively compare the clinical workflow (including end users) with and without the AI system, and record any efficiency gains or losses.
(b) In the case of an AI system focused on image quality, better or faster reconstructions or related functions: Prospectively collect data from phantoms, patients or both, in which the AI system is compared with current practice.
(c) In the case of an AI system focused on higher efficacy: Retrospectively collect data, including the labels (e.g. outcomes, diagnoses, segmentations), over a defined period from the healthcare organization. A period of one year may be considered for the retrospective data collection; however, this may vary depending on the particular disease and the clinical outcome being predicted by the model.
(d) If it is not possible to create a sufficiently large or representative test set from the healthcare organization's own data: Consider contacting organizations with similar structures and patient populations to obtain additional test sets or results from acceptance and commissioning activities carried out in a comparable setting. Another option is to look for appropriate open access datasets that match the local clinical setting.

5.6.6. Workflow description

The CQMP needs to participate in documenting the expected clinical workflows before and after the introduction of the AI system. Multiple workflows may be affected by the AI system, including clinical workflows, the QA workflow and the IT and support workflow. Each workflow can be described in terms of processes, transitions between processes, and the actors involved.

Any differential effects on workflows need special attention. For example, a workflow that is now supported by the AI system may have a negative effect on a subsequent or parallel workflow. External workflows, outside the department or healthcare organization that implemented the AI system, may also be affected.

5.6.7. Risk analysis

An AI system affects various workflows, including clinical workflows, QA workflows and IT workflows. It is the responsibility of the CQMP to conduct a risk analysis with their team for each AI system and workflow, ideally using a formal, prospective risk analysis method such as FMEA [90].

Examples of risks that need to be considered are:

(a) The AI system may become non-functional, for example as a result of power failures, failure of dependencies or loss of access to cloud based systems.
(b) Connection failures may occur with a PACS, HIS or other integrated systems.
(c) The AI system may receive erroneous input data.
(d) The AI system may produce an incorrect output.
(e) The end user may use the AI system incorrectly.
(f) The end user may interpret the output of the AI system incorrectly.
(g) The performance of the AI system may be reduced as a result of changes in patients, workflows, data or the AI system itself.
(h) The AI system may introduce bias, for example by performing less well in certain care settings or for certain patient groups.
(i) Clinicians or other end users may experience deskilling.
(j) Cybersecurity or information security failures may occur.

The CQMP, together with the end users and staff with clinical responsibility, needs to assess what the identified risks may mean for patients, the healthcare organization, end users and other stakeholders. The CQMP needs to classify the risks in terms of their probability and impact combined in an overall risk score. The CQMP also needs to consider whether and how the identified risks will be mitigated and managed.

5.6.8. Reporting acceptance and commissioning

In accordance with the guidelines of the healthcare organization, relevant regulatory bodies and national authorities, the CQMP needs to report their findings on the installed AI system in the appropriate manner.

5.6.9. Design of routine performance testing

The CQMP needs to design routine performance testing (also known as QC testing) for each specific AI system. The goal of this testing is to ensure that the AI system continues to operate as expected with respect to both efficacy and efficiency, according to the baselines established during acceptance and commissioning, including for recently collected data. Some routine performance testing procedures may already have been developed by the manufacturer, or for acceptance and commissioning, but additional procedures may be needed. Section 5.8 discusses routine performance testing procedures and their design.

5.6.10. Reacceptance and recommissioning as needed

The CQMP needs to consult with the manufacturer when the AI system is updated and may need to conduct acceptance testing again. Similarly, recommissioning may be needed if major changes occur within the AI system, in the data or in the equipment in the healthcare organization. For example, recommissioning may be needed if the healthcare organization purchases a new imaging device that produces input data for the AI system. The manufacturer needs to make clear the relevant steps for development, testing, acceptance and production. The CQMP also has to repeat customization and commissioning if the AI system, the context in which it operates or both have been changed in a major way.

5.7. INTRODUCTION INTO THE CLINICAL SETTING

After successful installation (see Section 5.5) and acceptance and commissioning (see Section 5.6), the AI system may be introduced into the clinical setting.

5.7.1. Set-up of a clinical use and production environment

After acceptance and commissioning outside the production environment, the AI system needs to be migrated to the clinical production environment. This migration needs to be tested in a manner similar to the procedures described in Section 5.5, because configurations (such as those of systems with which the AI system needs to be integrated) may differ from those in the acceptance environment. For example, clinical end users might not have the same permissions and profiles as those who performed the acceptance and commissioning, which could result in the AI system functioning differently. Migrating new systems to

the (clinical) production environment may disturb ongoing clinical operation. It is therefore advisable to carry out the migration outside normal clinical operating hours.

The IT department typically needs to plan a rollback scenario to be executed if the migration fails. A rollback usually involves recovering the infrastructure to the state (recovery point) where it had been before the installation began. Such a rollback might result in data loss if new data were produced between the recovery point and the rollback time. If migration fails, the underlying issues have to be resolved and the acceptance and commissioning process subsequently repeated.

Many of the same considerations applicable to the initial installation and acceptance and commissioning processes in the testing/acceptance environment (see Sections 5.5 and 5.6) remain relevant when an AI system is migrated to the production or clinical environment. In addition, the following tasks need to be carried out by the CQMP when moving the AI system to the production environment:

(a) Set up a testing environment for update and upgrade testing, in collaboration with IT staff.
(b) Establish procedures for updates and upgrades, ensuring compatibility with hospital systems and third party hardware and software.

5.7.2. Technical release

A smooth technical release of the AI system needs to be ensured, accompanied by technical release notes prepared by the CQMP and their team. The aims of the technical release notes are to:

(a) Outline the technical aspects of the AI system and the process and steps of installation, including any modifications or enhancements made to the system.
(b) Outline the clinical workflow, data management and QA programme.
(c) Describe how and where to report errors and incidents, and how to receive support for the AI system.

5.7.3. Clinical release

After the technical release, the AI system needs to be formally released for clinical use, and clinical release notes need to be prepared by the CQMP and their team. Some important purposes of the clinical release notes are to:

(a) Clarify and outline the objectives of integrating the AI system into the clinical workflow.
(b) Define the scope of the application, specifying its intended use and functions and its impact on medical processes.
(c) Specify the intended populations and the limitations.
(d) Specify the necessary regulatory compliance measures aligned with patient medical and data protection regulations.
(e) Detail any legal considerations linked to the implementation of the AI system in the clinical context.
(f) Describe the authorized users and the level and method of mandatory training needed to use the AI system competently.
(g) State the responsible individuals and contacts.

5.7.4. Training of technical staff and clinical end users

At this point in the clinical implementation, both technical staff and clinical end users will already have received the training needed to begin operating and using the AI system clinically (see Section 5.5.8).

At this stage, the CQMP needs to verify that all technical staff and clinical end users have completed the necessary training and that any role specific training materials, user instructions and support documentation are available. The CQMP also needs to ensure that refresher training and training for new personnel can be provided as needed.

5.7.5. User feedback mechanism

The CQMP and their team need to gather feedback from users to identify areas for improvement and to ensure the ongoing refinement of the AI system function. This feedback is important for ensuring that users continue to use the system. The CQMP has to be provided with resources to ensure that a user friendly mechanism for gathering feedback from clinical end users exists. This mechanism may be very simple, such as communication in person or by email, or it may be based on a more advanced ticketing system or service desk, if available in the healthcare organization. Once received, the feedback needs to be carefully assessed and used to make iterative improvements and enhance user satisfaction, for example through a plan–do–check–act (PDCA) cycle [91].

5.7.6. Contingency plan development and testing

If the AI system is unavailable, whether because of routine maintenance or malfunction, the end user will need to revert to the protocols used prior to the

installation of the AI system (whether that involves a manual mode or an earlier AI system). For example, the radiologist may need to read without the aid of the AI system output, or the radiation oncologist may need to define targets without the support of computerized automated segmentation. It is therefore important that the contingency plan includes measures to mitigate the risks that deskilling may pose when the AI system is unavailable. This process needs to be established and tested as part of the overall installation of the AI system in the clinical environment. If the AI system is fully integrated for a specific clinical task, the only contingency plan may be to seek alternative service providers.

5.7.7. Post-deployment monitoring

Once the AI system has been deployed, its performance and any user feedback need to be closely monitored. The CQMP and their team also need to continually assess the impact of the AI system on clinical workflows and patient outcomes through established QC protocols (see Section 5.8).

5.8. QUALITY MANAGEMENT

As stated in IAEA Human Health Series No. 25 [15], it is the responsibility of the CQMP to "participate as a team member in designing and implementing a quality management programme". Within the context of QA of an AI system, the CQMP leads the risk assessment process, together with the clinical team. They prepare a protocol for the clinical use of the AI system, monitor changes in patient cohorts and in imaging or treatment protocols, assess the efficiency and consistency of the AI system, act to ensure its continued safe and effective use and lead the establishment and implementation of a QA programme for the AI system. In the case of AI system updates, the CQMP adapts the QA protocol and is responsible for training or retraining the clinical team. The CQMP monitors AI system performance and leads reporting and discussions with the manufacturer.

The use of AI systems is acceptable if they enhance the efficacy or efficiency, or both, of healthcare processes while operating within expected, acceptable and reasonable risk levels. As a medical device, a clinical AI system has an intended use for an intended population and a corresponding claim. The role of QA is therefore twofold:

(a) To ensure that the AI system is used as intended (see Section 5.6);
(b) To ensure that the AI system functions as claimed, both to ensure safety and to ensure effectiveness.

Quality management needs to cover the performance of the AI system, the workflow and the end user. Additionally, patients, clinical workflows, information systems, devices, data and AI systems may all change over time, so in this dynamic environment, the QA programme needs to provide assurance that the AI system continues to meet the needs of the organization.

This section discusses the management of quality within the QA programme, including routine QC and related QC activities. Routine QC takes place while the AI system is available for clinical use. The management plan for QC, including routine QC, is created during acceptance and commissioning, but it may be revised during operation. The risk analysis conducted as part of that process is an important determinant of which QC tests are needed and how they are designed.

The QC tests for an AI system might not need to be designed entirely by the CQMP and their team. The manufacturer of the AI system may provide QC test protocols, which may be implemented directly within the software. Some medical imaging examinations and procedures already have mandatory professional [77, 92] or regulatory [93] auditing mechanisms in place that can supplement the AI system specific QC tests developed by the CQMP. For example, in screening mammography, the Mammography Quality Standards Act (MQSA) requirements include assessment of both the image quality of mammographic images and the diagnostic performance of the clinical end user [94]. Even so, the CQMP is encouraged to devise additional QC tests when and where relevant.

For each QC test, the CQMP needs to describe what aspect of the AI system, clinical workflow or end users is being tested, how the test is conducted, the levels of deviation that will trigger further action and what action needs to be taken. This can range from noting and accepting the deviation to taking immediate action and removing the AI system from the clinical workflow. The action to be taken depends on the risk associated with the observed deviation and on the criticality of the AI system in the workflow. In some situations, depending on the regulatory requirements of the healthcare organization, the reporting may need to be sent to a relevant regulatory body (e.g. the FDA) or to the manufacturer (e.g. for CE marked devices). In all cases, the CQMP has to discuss the deviation with the clinical team using the AI system and the other staff responsible for the AI system.

Activities within the QA programme can be grouped into four main areas [95–97]: performance testing (routine QC, case specific QC and ad hoc QC) and incident management. The following subsections describe each of these areas.

5.8.1. Routine quality control

Routine QC is conducted on a planned and scheduled basis and supervised by the CQMP. It consists of automated and manual QC tests. The following considerations apply to both:

(a) Data sources to be considered: AI systems can be evaluated against several datasets, including any manufacturer supplied QC dataset, the commissioning dataset used to assess deviation from the initial baseline, the recommissioning dataset that established a new baseline for the current version or clinical setting (or both) of the AI system and the dataset of recently collected samples to assess changes occurring in practice. The QC test protocol needs to specify which datasets and other data sources are needed, if any. Manual QC tests may necessitate the use of phantoms.
(b) Timeframe to consider: Many potential routine QC tests involve monitoring the performance of an AI system for possible underperformance. Because more than one sample is usually needed to assess performance, the QC test protocol needs to describe how long the retrospective window for sample collection needs to be or how many samples are needed at minimum. This is a balancing act. Short timeframes, and thus higher variability in the underlying data, may cause irrelevant QC failures. Long timeframes cause fewer irrelevant QC failures, but the drawback is that any QC failure detected is likely to have been present for some time.
(c) Human intervention to be expected: The QC test protocol needs to specify what intervention is expected. For example, if an AI system is used for patients outside its prescribed intended population, this may prompt the need for awareness of the potential off-label use and additional training for end users. At the other end of the spectrum, an intervention may necessitate temporarily halting the use of an AI system for patient critical applications until the issue is resolved. The AI system manufacturer may provide specific information on the expected intervention. If a decline is observed in interaction with an AI system by users, this may point to complacency, which may necessitate an effort to sensitize the users.
(d) Reporting results: Out of tolerance results of routine QC tests need to be reported to and reviewed by the responsible CQMP.

Automated and manual QC approaches are discussed below.

5.8.1.1. Automated quality control

Automated QC tests run without routine human involvement, prompting human intervention only when necessary after detecting anomalies through observation of how the system is used. The CQMP is advised to consider the following questions when creating an automated QC system:

(a) What needs to be tested? This depends on the AI system, its purpose and the impact of its use. Automated routine QC tests can monitor the AI system itself, its users and the clinical workflow. For example, automated QC tests may regularly assess the AI system performance against a test dataset supplied by the manufacturer or compiled during commissioning with known reference standards or ground truth. A concurrent AI system for screening mammography images can be automatically tested for agreement with a human reader in the last 100 patients seen [92, 98]. An AI system for segmenting normal tissues in radiotherapy planning CT can be tested for the degree of manual corrections needed in the last month. QC tests may also involve checking that the AI system receives the expected input data (i.e. that the correct data are read and that these data represent the intended use case for the AI system). Generally, each QC activity tests something identified during acceptance and commissioning and in the risk analysis.
(b) How frequently do automated QC tests need to be conducted? Some tests call for continuous monitoring, such as checks on received input data. Others may be conducted at daily, weekly or longer intervals, such as monitoring the performance of AI systems for predicting tumour control in patients, where the ground truth might not be known for months or years. AI systems also have to be checked when there is a change in the input, such as the installation of a new image acquisition device or an upgrade of the organization's operating system. In some applications, the time at which the QC test is conducted can be relevant, for example when load on the IT infrastructure may cause differences in the response times of an AI system.
(c) When is human intervention prompted? An automated QC test is expected to prompt human intervention only when anomalies in the system itself or in its usage are detected. This necessitates that a specific heuristic be defined, such as a significant decrease in AI system performance compared with its baseline or significantly decreased user interaction with the AI system. The claim and intended use stated by the manufacturer of the AI system may help inform this heuristic.
(d) How does the automated QC process need to be tested? The automated QC process itself may fail silently (i.e. stop functioning correctly without notification). During such periods, the QC process fails to prompt human

intervention as necessary, which may lead to dangerous situations. At least two aspects need to be checked. First, the response to complete failure (e.g. a software crash) of the automated QC process has to be tested. Second, the response of the automated QC process to data designed to trigger human intervention needs to be tested on a routine basis. Ideally, an AI system needs to be prevented from being used clinically if the appropriate QC process is not active.

In some circumstances, automated QC tests are challenging to design. For example, image synthesis AI systems are used to replace observations that would otherwise be used to test the AI system. In such cases, the AI system has to be tested using manual QC, for example using phantoms.

5.8.1.2. Manual quality control

Manual QC, like automated QC, exists to ensure that the AI system is used as intended and that it continues to function as claimed and as established at the baselines during commissioning. Unlike automated QC, manual QC of an AI system involves the CQMP and other personnel, including the clinical end user and IT staff, more directly in performing QC.

Manual QC may be relevant for the following purposes:

(a) QC tests that cannot be fully automated (e.g. when observations outside the clinical routine would be needed for automation). This may be the case for image synthesis AI systems, where the reference standard is not obtained in clinical routine and (digital) phantom measurements or other approaches may be needed. Another example is a QC test for an AI system embedded within a closed system that cannot be checked automatically.
(b) Testing of automated QC processes. As noted in the previous section, automated QC processes need to be regularly tested to confirm their correct functioning.
(c) QC for monitoring individual end user interactions with the AI system.
(d) Summary inspection of automated QC routines. This involves summarizing the results generated by automated QC tests (e.g. those that monitor AI system performance).
(e) Technical performance checks of the AI system (e.g. the time taken to process input data to output data, stress testing).

The CQMP is advised to consider the following questions when designing manual QC tests. Several aspects overlap with those considered for automated QC.

— What needs to be tested? This depends on the overall aim of the test. Testing of automated QC processes involves checking whether the automated QC tests function as expected and whether anything that influences the operation of the AI system has changed. For QC of AI systems that are not automated, the test depends on the AI system and its purpose. For example, an AI system for screening mammography images can be tested for agreement with a human reader or retrospectively on cases with known outcomes. An AI system for segmenting normal tissues in radiotherapy planning CT can be tested for the degree of manual correction needed. A manual QC test may also involve checking whether the AI system receives the expected input data (i.e. that the correct data are read). For technical performance, the time needed for a particular task or end to end testing may be considered.
— How frequently do manual QC tests need to be conducted? The regularity of manual QC tests depends on the risk associated with the aspect of the AI system being checked. Critical aspects need to be tested more often than those associated with a lower risk.

5.8.2. Case specific quality control

Case specific QC involves checking that the AI system output for an individual case meets expectations. This is in contrast to other routine QC tests, which typically evaluate performance across a cohort of cases.

For example, case specific QC in radiotherapy may involve flagging possibly incorrect segmentations, adjusting segmentations generated by an AI system or recomputing AI generated dose–volume histograms using Monte Carlo approaches if the AI generated dose–volume histograms are deemed unexpected or are critical for the plan. Case specific QC in AI aided diagnostic image interpretation could involve obtaining a second opinion on the interpretation if the AI system output is of concern, or could involve comparing outputs from two or more AI systems for the same clinical task.

Case specific QC can be manual or automated and is typically part of the routine clinical workflow. If it is automated, it may produce a warning to the end user, for instance if the input data, use or output of the AI system falls outside the intended use, population or both.

Pass or failure of case specific QC needs to be noted in the patient file or in another system in which this information can be traced to a specific patient. This documentation may indicate groups of patients in which the AI system is not performing well. Such groups are a useful source of novel test data for use in routine QC, for training purposes, to inform risk analyses and to support improvement of the AI system. A failed case specific QC check can also trigger a

routine QC test, which may highlight the need for retraining or recommissioning of the AI system.

If the use of an AI system results in an undesirable effect on an individual patient, this may be considered an unintended event (e.g. a patient safety incident). Depending on the severity of the effect, the incident may need to be reported to regulatory authorities, the manufacturer or both.

The need for case specific QC has to be informed by the risk analysis conducted during acceptance and commissioning.

5.8.3. Ad hoc quality control

Ad hoc QC, unlike routine and case specific QC, is not performed on a regular basis. Ad hoc QC repeats elements of the acceptance and commissioning procedure. It is performed when changes in the routine use of an AI system occur, as a result of changes either in the AI system itself or in the infrastructure surrounding the AI system. Such changes may include the release of a new version of the AI system, a major upgrade to the IT infrastructure, or the acceptance and commissioning of a new imaging device that interacts with the AI system. An increase in unintended events may also trigger ad hoc QC. The following considerations apply to ad hoc QC:

(a) In addition to the acceptance and commissioning procedure, existing automated and manual QC processes need to be reviewed for correct functioning and revised if necessary.
(b) If ad hoc QC testing fails during commissioning of a new version of an AI system, the new version cannot be released for clinical use. Collaboration with the manufacturer is advisable and may be necessary to ensure that new versions pass ad hoc QC.
(c) Major changes in an AI system may necessitate additional training of users.

5.8.4. Incident management

Incidents may occur in association with the use of AI systems for clinical purposes. All users need to know where, when and how to report incidents [99]. Incidents can range from temporary unavailability of the AI system to direct harm being done to patients. Failures in automated and manual routine QC tests may also be considered incidents, or user feedback may lead to an incident being reported. Depending on the severity of the incident, national regulations may require reporting to internal patient safety officers, regulatory bodies or both.

The occurrence of an incident generally means that the AI system will not be available until the underlying cause is resolved and the system is deemed safe for use. Several considerations are relevant in these situations:

(a) Depending on the cause of the incident, the manufacturer of the AI system may need to be informed about the issue and be involved in resolving it.
(b) Downtime has to be documented. The contract with the manufacturer of the AI system may stipulate a maximum acceptable downtime, and penalties for extended downtime may be specified.
(c) Fall-back strategies have to be devised as part of acceptance and commissioning of the AI system (see Sections 5.6 and 5.7.6) to detail what is expected to happen if a tool is not available. For critical AI systems, the fall-back strategy may be rehearsed, similar to a fire drill or flight simulator test, so that users and other personnel know what to do if a critical AI system is not available.
(d) The CQMP needs to assess whether the incident warrants a change in the AI system itself, a change in QC activities, additional user training or additional risk analysis and management.
(e) The CQMP needs to include data pertaining to the incident in the post-market surveillance to support effective corrective actions (see Section 5.3.3).

5.9. CLINICAL EVALUATION AND IMPACT ASSESSMENT

During this phase, the AI system is being used in routine clinical practice and a QA programme is in place. It can be assumed now that the AI system is operating safely and in accordance with its specifications. However, this does not constitute proof that the AI system addresses the organizational needs or meets the expectations previously defined by the CQMP and their team (see Section 5.1).

A clinical implementation and evaluation team is assumed to be periodically evaluating the AI system in the real world clinical setting and workflow. The team may need to involve end users (both healthcare professionals and clinical support staff), the CQMP, management, technical support staff, IT staff and others. This team will need to devise a mechanism for workflow audits based on the specific functioning of the AI system.

5.9.1. Effect on workflow actors

Workflows affected by the AI system were identified during acceptance and commissioning (see Section 5.6.6). As part of clinical evaluation and impact assessment, the CQMP needs to participate in a team that evaluates how the

AI system has affected the actors in these workflows, both before and after its introduction. The effects on these actors need to be described both factually (in terms of observable, objective results) and subjectively (in terms of the affected actors' opinions). Special attention needs to be paid to any differential effects; for example, an AI system may reduce workload for clinical staff but increase it for IT staff.

5.9.2. Effect on organizational needs

The CQMP needs to participate in an objective evaluation of whether the AI system has met the organizational needs identified at the beginning of the process (see Section 5.1). The metrics for this evaluation were defined in the initial identification of organizational needs phase and may include aspects such as time saved, reduction in throughput times, higher staff satisfaction, less burnout, higher number of patients, higher quality of images or treatment plans, fewer errors, better outcomes, higher diagnostic accuracy, higher patient satisfaction, lower costs and higher profits. A holistic view is needed, and both positive and negative effects of the AI system on various aspects of the healthcare organization have to be considered; for example, an AI system may improve physician satisfaction but may be costly and increase demands on IT staff.

5.9.3. Identification of enablers and barriers

During this phase, it is important to determine how often the AI system is used in the clinical workflow and whether it is used by all actors for all relevant patients. Typically, this is not the case, and the CQMP needs to understand the reasons behind this (e.g. what enables some staff to use the AI system in certain cases and what hampers its use in others). Enablers and barriers can be related to individual staff members or groups who might not trust, are not trained on, or do not want to use the AI system for various reasons. The culture of a department or healthcare organization may also play a role. Another known barrier may be that the AI system does not work well, or is perceived not to work well, for specific patients or groups of patients, or in particular data settings or clinical contexts. Allowing alternative workflows to exist in parallel with the AI supported workflows can also diminish use of the AI system. Some of these barriers may become apparent only after clinical implementation. Enablers can also be identified, such as 'champion' users who articulate the benefits of the AI system, extensive training programmes and trust in the AI system. The CQMP and their team need to identify such barriers and enablers to improve or maintain successful use of the AI system.

5.9.4. Possible improvements

Once sufficient experience has been gained with the AI system, areas for improvement may be identified. These may relate to many aspects, including the AI system itself, staff training or integration into the workflow. The CQMP needs to actively elicit such possible improvements from the actors in the various workflows and initiate new activities to test and implement these improvements.

5.9.5. Continuous clinical evaluation

As patients, staff, data, organizations and AI systems change, so will the impact of the AI system on the organization's goals. The CQMP needs to participate in repeated objective and subjective clinical evaluations at regular intervals. These clinical impact evaluations are in addition to continuous QC (see Section 5.8) and ongoing training (see Section 5.5.8).

5.10. DECOMMISSIONING

An AI system may need to be taken out of service when it is no longer supported by the manufacturer (i.e. it has reached its end of life) or when it is no longer compatible with existing systems (e.g. because of upgrades or changes in equipment). An AI system may also be taken out of service when it is no longer desired by a healthcare organization, for instance because of a lack of effectiveness or because it is replaced by another AI system that addresses the same clinical need. Before an AI system is removed from clinical use, the resulting consequences need to be carefully reviewed and considered in terms of the effect on overall operations and clinical care. Such an assessment needs to be conducted in a testing environment prior to the actual removal from the clinical production environment.

In this section, the main considerations for safely decommissioning AI systems are discussed. However, the laws and regulations of the Member State in which the healthcare organization is located may impose additional considerations, and the CQMP and their team are advised to refer to them.

5.10.1. Identification of affected clinical processes

When an AI system is taken out of service, it is critically important to identify all clinical processes and workflows that are affected by the system. This necessitates a careful analysis and review of the workflows and processes downstream of the AI system. Depending on the nature of the system, these may

be straightforward to determine. Physicians will generally be familiar with and aware of AI systems that directly support their practice, such as systems used for automated segmentation in radiation treatment planning or for vessel removal in lung CT. However, the impact of other AI systems, such as cancer detection tools, may be more subtle, and the absence of such an AI system might or might not affect diagnostic performance.

In some cases, there may be limited awareness of the presence of an AI system in a workflow. For example, a system that performs triaging or flags studies for immediate review might not be visible to the end user. In addition, clinical processes may be affected if deskilling has occurred and clinical decision making is now potentially compromised.

The review of affected clinical processes has to be performed by a group of healthcare professionals who are familiar with the purpose and output of the AI system, including the physician end user, hospital administrators, medical radiation technologists, the CQMP and IT staff.

5.10.2. Identification of alternative solutions and replacement criteria

The review team will have the knowledge to identify alternative solutions and to establish criteria for replacing the AI system, which may differ from the criteria for decommissioning the system. If a replacement for the AI system is sought, the business needs have to be reviewed and revised based on experience with the existing system, including its effectiveness, ease of use, data needs, service arrangements, usage patterns and maintenance agreements. Selection, commissioning and QC then need to be carried out for the new system. The overall process for introducing a replacement AI system follows the steps outlined in Sections 5.1–5.7.

5.10.3. Revision of clinical workflow, quality assurance procedures and security protocols after end of life and end of support

If an AI system is taken out of service, QA procedures for the system itself, or for processes that involve the system, need to be revised. It is essential to identify steps in the QA procedures that assumed availability of the AI system and adjust the related instructions and procedures to function with the successor system, if any. Network destinations will need to be revised and updated on all systems that provided input data to the AI system or received data from it. Before an AI system is taken out of service, the necessary changes have to be tested within a testing environment, before removing the AI system from the clinical production environment. To safely decommission an AI system, the CQMP also

needs to assess clinical performance levels and clinical workflows together with the clinical end user.

The decommissioned AI system needs to be archived, where relevant and permitted, and needs to be removed from all devices, to prevent accidental use of a tool that may no longer be compliant or appropriate. For example, this is relevant for an AI system that is no longer approved for use with updated imaging equipment. It is also important to remove decommissioned AI systems for security reasons. In addition, audit records and QA data for the AI system need to be archived to comply with regulations.

Furthermore, when a cloud based AI system is being decommissioned, the healthcare organization is expected to work with the manufacturer to ensure that all patient and organization specific data are deleted from their cloud systems and archived on in-house systems.

5.10.4. Continued use beyond end of life and end of support

The use of an AI system beyond its end of life or end of support necessitates careful review of security protocols to prevent data or networks from being compromised. Third party solutions may be available to allow ongoing use of software in a specialized environment, for example if the necessary operating system is no longer supported or interoperable with the organization's infrastructure and equipment.

National regulatory requirements and medicolegal considerations (e.g. liability issues) associated with the continued use of an AI system beyond its end of life or end of support need to be examined and any related obligations fulfilled accordingly.

6. CONCLUSIONS

The rapid integration of AI into healthcare systems has brought, and will continue to bring, major changes in healthcare, offering opportunities for improved diagnostics, treatment planning, treatment response monitoring, outcome prediction and patient care [100].

The IAEA assists its Member States in the application of nuclear sciences and technologies to human health to ensure quality and safety in the medical uses of radiation. The overall goal of the IAEA human health programme is to enhance the capabilities of Member States to address issues related to the prevention, diagnosis and treatment of disease through the application of nuclear techniques.

Providing support to the medical physics profession helps to ensure quality and safety in the medical use of ionizing radiation. Through its programmatic efforts and its publications, the IAEA has helped to clearly define the roles and responsibilities of CQMPs in supporting the safe and effective use of AI systems in clinical environments [4].

CQMPs play a leading role in a multidisciplinary team to ensure successful implementation of imaging based AI systems within clinical environments. Specifically, CQMPs help bridge the gap between the complex algorithms and technology underlying AI systems and the practicalities of clinical settings. Their involvement encompasses a wide range of responsibilities, from conducting market research and ensuring seamless integration with the existing healthcare environment to optimizing AI systems for specific diagnostic or prognostic tasks, as well as conducting continuous QC and educating all key operators and stakeholders on the potential positive and negative effects of AI implementation. CQMPs thus serve as key intermediaries in translating the capabilities offered by cutting edge AI systems into tangible benefits for patients and healthcare providers. A summary of key tasks to be performed and aspects to be considered by the CQMP and their team when acquiring and implementing an imaging based AI system is provided in Appendix II.

Appendix I

MODEL CARD FOR CLINICAL ARTIFICIAL INTELLIGENCE SYSTEMS BASED ON MEDICAL IMAGING

A manufacturer may provide a model card to inform the CQMP and others about their product. Model cards or model sheets help present AI systems in a standardized, clear format [101]. A template model card is described below.

(a) Summary description: A short description of the AI system and its use case.
(b) Claims: The claims of the system, as submitted to regulatory authorities. The claims describe the use case and the expected performance in that role. The CQMP may use these claims to verify the performance of the AI system.
(c) AI system details: Further information about the AI system itself, including:
    (i) Developers: The individuals or organizations who developed the AI system.
    (ii) Version: The current version of the AI system.
    (iii) Release date: The date on which the current version of the AI system was released.
    (iv) Expected input: The intended input for the AI system. For imaging based AI systems, this needs to include, for example, the imaging modality, relevant details of imaging protocols and the data format. Any additional data inputs (e.g. clinical factors) are also described.
    (v) Expected output: The output produced by the AI system.
    (vi) Model architecture: The type or types of model included in the AI system (e.g. random forest, U-Net convolutional neural network).
    (vii) Training data: The characteristics of the dataset used to train the AI system, including demographic and clinical characterization of the patients included in the training data, as well as relevant details related to scanners, image acquisition, reconstruction and segmentation.
    (viii) Validation data: As described for the training data, but for the data used to validate the AI system, which were not part of the training set.
    (ix) System performance: Observed performance of the current version of the AI system, preferably on the test set used by the developers. Performance in patient subpopulations has to be differentiated (e.g. by gender, genetic mutations, tumour staging). Other aspects of the AI system, such as calibration, are also potentially relevant.

(d) Intended use:
    (i) Intended use cases: The use cases for which the AI system was developed.
    (ii) Intended users: The expected users of the AI system and their qualifications.
    (iii) Out of scope use cases: Potential use cases for which the AI system is not to be used, including use cases similar to but outside the intended use.
    (iv) Known limitations: A list of known limitations of the AI system.
(e) Fairness, safety and trustworthiness:
    (i) Fairness: A description of how fairness (i.e. the AI system performance in diverse patients and healthcare settings) of the AI system was assessed and which biases are present.
    (ii) Safety — privacy: A description of how the AI system preserves patient privacy during use. The CQMP is advised to determine whether patient or other data are transmitted to the manufacturer.
    (iii) Safety — intended use: A description of how the AI system helps ensure that it is used correctly. For example, the AI system may include algorithms to detect out of scope use cases or plausibility checks on the input and output.
    (iv) Trustworthiness: A description of how the AI system explains its outputs or decisions.
(f) Equipment and QC:
    (i) Expected equipment: Any equipment needed for using the AI system.
    (ii) Tool interface (API): A summary of the API of the AI system. The API allows integration of the AI system into a digital infrastructure (e.g. by interfacing with a PACS and exporting the output in machine readable formats for automated QC tests). The complete API has to be documented separately.
    (iii) QC tests provided by the manufacturer: A list of any QC tests integrated with the AI system, with a short description of their purpose. These tests have to be documented separately in full.

Appendix II

SUMMARY OF KEY TASKS AND CONSIDERATIONS WHEN ACQUIRING AND IMPLEMENTING AN IMAGING BASED ARTIFICIAL INTELLIGENCE SYSTEM

This appendix provides an overview of tasks to be performed and aspects to be considered by the CQMP and their team when acquiring and implementing an imaging based AI system. The most important groups of tasks and related considerations, along with references to the corresponding sections of this publication, are summarized in Table 3.

TABLE 3. SUMMARY OF KEY TASKS AND CONSIDERATIONS WHEN ACQUIRING AND IMPLEMENTING AN IMAGING BASED AI SYSTEM

Section	Task/consideration
5.1	Identification of organizational needs
5.1.1	Functional requirements
5.1.1.1	— Defining the problem
5.1.1.2	— Defining success
5.1.1.3	— AI system output and intended use
5.1.1.4	— Intended population
5.1.1.5	— Assessing the reference performance
5.1.1.6	— Defining the necessary AI system output performance
5.1.2	Non-functional requirements
5.1.2.1	— Workflow considerations
5.1.2.2	— Data maturity
5.1.2.3	— AI system deployment
5.1.2.4	— Autonomy, risks and liabilities
5.1.2.5	— Transparency, explainability and interpretability
5.1.2.6	— Change and expectation management
5.1.3	Business case
5.2	Market research
5.2.1	Product fact sheet
5.2.2	Product, licensing and service models
5.2.3	Vetting
5.3	Preselection and demonstration
5.3.1	Demonstration preparation

TABLE 3. SUMMARY OF KEY TASKS AND CONSIDERATIONS WHEN ACQUIRING AND IMPLEMENTING AN IMAGING BASED AI SYSTEM (cont.)

Section	Task/consideration
5.3.2	Demonstration by manufacturer: End user perspective
5.3.3	Demonstration by manufacturer: Technical perspective
5.3.4	Demonstration by manufacturer: Documentation perspective
5.4	Selection and purchasing
5.4.1	General requirements
5.4.1.1	— Compatibility
5.4.1.2	— Scalability
5.4.1.3	— Regulatory compliance
5.4.1.4	— Staffing
5.4.1.5	— Training
5.4.2	Functional requirements
5.4.2.1	— Intended use within the healthcare organization
5.4.2.2	— Input data and intended population
5.4.2.3	— Output data and range
5.4.2.4	— Workflow integration
5.4.3	Performance metrics
5.4.3.1	— Accuracy
5.4.3.2	— Speed
5.4.3.3	— Errors, uncertainty and risk mitigation
5.4.4	Non-functional requirements
5.4.4.1	— Architecture
5.4.4.2	— Operating requirements
5.4.4.3	— Information security
5.4.4.4	— Interoperability
5.4.5	QA and auditing
5.4.5.1	— QA
5.4.5.2	— Audit trail
5.4.6	Maintenance, support and improvements
5.4.6.1	— Updates and upgrades
5.4.6.2	— Support, responsibilities and liabilities
5.4.6.3	— Incident, change and release management
5.4.7	Data transfer and processing
5.4.7.1	— Use of data from the healthcare organization by the manufacturer

TABLE 3. SUMMARY OF KEY TASKS AND CONSIDERATIONS WHEN ACQUIRING AND IMPLEMENTING AN IMAGING BASED AI SYSTEM (cont.)

Section	Task/consideration
5.4.7.2	— Use of data from the manufacturer by the healthcare organization
5.4.8	Contract termination and end of life
5.4.8.1	— Contract termination
5.4.8.2	— End of life and removal
5.4.9	Pricing, licensing and subscription models
5.5	Installation
5.5.1	Pre-installation assessment
5.5.2	Hardware and software set-up
5.5.3	AI system installation, integration and testing
5.5.4	Security implementation
5.5.5	Hardware and software stress testing
5.5.6	Establishment of monitoring and maintenance procedures
5.5.7	User management
5.5.8	User training and onboarding
5.6	Acceptance and commissioning
5.6.1	Evaluation metrics
5.6.2	Preparation for acceptance and commissioning
5.6.3	Acceptance
5.6.4	Customization, localization, fine tuning and continuous learning
5.6.5	Commissioning
5.6.6	Workflow description
5.6.7	Risk analysis
5.6.8	Reporting acceptance and commissioning
5.6.9	Design of routine performance testing
5.6.10	Reacceptance and recommissioning as needed
5.7	Introduction into the clinical setting
5.7.1	Set-up of a clinical use and production environment
5.7.2	Technical release
5.7.3	Clinical release
5.7.4	Training of technical staff and clinical end users
5.7.5	User feedback mechanism
5.7.6	Contingency plan development and testing
5.7.7	Post-deployment monitoring

TABLE 3. SUMMARY OF KEY TASKS AND CONSIDERATIONS WHEN ACQUIRING AND IMPLEMENTING AN IMAGING BASED AI SYSTEM (cont.)

Section	Task/consideration
5.8	Quality management
5.8.1	Routine QC
5.8.1.1	— Automated QC
5.8.1.2	— Manual QC
5.8.2	Case specific QC
5.8.3	Ad hoc QC
5.8.4	Incident management
5.9	Clinical evaluation and impact assessment
5.9.1	Effect on workflow actors
5.9.2	Effect on organizational needs
5.9.3	Identification of enablers and barriers
5.9.4	Possible improvements
5.9.5	Continuous clinical evaluation
5.10	Decommissioning
5.10.1	Identification of affected clinical processes
5.10.2	Identification of alternative solutions and replacement criteria
5.10.3	Revision of clinical workflow, QA procedures and security protocols after end of life and end of support
5.10.4	Continued use beyond end of life and end of support

Note: AI: artificial intelligence; QA: quality assurance; QC: quality control.

REFERENCES

[1] BARRAGÁN-MONTERO, A., et al., Artificial intelligence and machine learning for medical imaging: A technology review, Phys. Med. **83** (2021) 242–256, https://doi.org/10.1016/j.ejmp.2021.04.016

[2] KAUL, V., ENSLIN, S., GROSS, S.A., History of artificial intelligence in medicine, Gastrointest. Endoscopy **92** (2020) 807–812, https://doi.org/10.1016/j.gie.2020.06.040

[3] YASNITSKY, L.N., Artificial intelligence and medicine: History, current state, and forecasts for the future, Curr. Hypertens. Rev. **16** (2020) 210–215, https://doi.org/10.2174/1573402116666200714150953

[4] INTERNATIONAL ATOMIC ENERGY AGENCY, Artificial Intelligence in Medical Physics: Roles, Responsibilities, Education and Training of Clinically Qualified Medical Physicists, Training Course Series No. 83, IAEA, Vienna (2023).

[5] RANSCHAERT, E.R., MOROZOV, S., ALGRA, P.R. (Eds), Artificial Intelligence in Medical Imaging: Opportunities, Applications and Risks, Springer, Cham (2019).

[6] BI, W.L., et al., Artificial intelligence in cancer imaging: Clinical challenges and applications, CA Cancer J. Clin. **69** (2019) 127–157, https://doi.org/10.3322/caac.21552

[7] EL NAQA, I., HAIDER, M.A., GIGER, M.L., TEN HAKEN, R.K., Artificial intelligence: Reshaping the practice of radiological sciences in the 21st century, Br. J. Radiol. **93** (2020) 20190855, https://doi.org/10.1259/bjr.20190855

[8] GIGER, M.L., CHAN, H.-P., BOONE, J., Anniversary paper: History and status of CAD and quantitative image analysis: The role of Medical Physics and AAPM, Med. Phys. **35** (2008) 5799–5820, https://doi.org/10.1118/1.3013555

[9] GIGER, M.L., KARSSEMEIJER, N., SCHNABEL, J.A., Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer, Annu. Rev. Biomed. Eng. **15** (2013) 327–357, https://doi.org/10.1146/annurev-bioeng-071812-152416

[10] SABOTTKE, C.F., SPIELER, B.M., The effect of image resolution on deep learning in radiography, Radiol. Artif. Intell. **2** (2020) e190015, https://doi.org/10.1148/ryai.2019190015

[11] WORLD HEALTH ORGANIZATION, Ethics and Governance of Artificial Intelligence for Health: WHO Guidance, WHO, Geneva (2021).

[12] SALAHUDDIN, Z., WOODRUFF, H.C., CHATTERJEE, A., LAMBIN, P., Transparency of deep neural networks for medical image analysis: A review of interpretability methods, Comput. Biol. Med. **140** (2022) 105111, https://doi.org/10.1016/j.compbiomed.2021.105111

[13] WORLD HEALTH ORGANIZATION, Generating Evidence for Artificial Intelligence-Based Medical Devices: A Framework for Training, Validation and Evaluation, WHO, Geneva (2021).

[14] WORLD HEALTH ORGANIZATION, Regulatory Considerations on Artificial Intelligence for Health, WHO, Geneva (2023).

[15] INTERNATIONAL ATOMIC ENERGY AGENCY, Roles and Responsibilities, and Education and Training Requirements for Clinically Qualified Medical Physicists, IAEA Human Health Series No. 25, IAEA, Vienna (2013).
[16] INTERNATIONAL ATOMIC ENERGY AGENCY, Artificial Intelligence for Accelerating Nuclear Applications, Science and Technology, IAEA, Vienna (2022).
[17] KIM, D.Y., OH, H.W., SUH, C.H., Reporting quality of research studies on AI applications in medical images according to the CLAIM guidelines in a radiology journal with a strong prominence in Asia, Korean J. Radiol. **24** (2023) 1179–1189, https://doi.org/10.3348/kjr.2023.1027
[18] SOUNDERAJAH, V., et al., Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: The STARD-AI protocol, BMJ Open **11** (2021) e047709, https://doi.org/10.1136/bmjopen-2020-047709
[19] COLLINS, G.S., et al., Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence, BMJ Open **11** (2021) e048008, https://doi.org/10.1136/bmjopen-2020-048008
[20] COLLINS, G.S., et al., TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods, BMJ **385** (2024) e078378, https://doi.org/10.1136/bmj-2023-078378
[21] MES, S.W., et al., Outcome prediction of head and neck squamous cell carcinoma by MRI radiomic signatures, Eur. Radiol. **30** (2020) 6311–6321, https://doi.org/10.1007/s00330-020-06962-y
[22] DIAZ, O., et al., Data preparation for artificial intelligence in medical imaging: A comprehensive guide to open-access platforms and tools, Phys. Med. **83** (2021) 25–37, https://doi.org/10.1016/j.ejmp.2021.02.007
[23] CROSSNOHERE, N.L., ELSAID, M., PASKETT, J., BOSE-BRILL, S., BRIDGES, J.F.P., Guidelines for artificial intelligence in medicine: Literature review and content analysis of frameworks, J. Med. Internet Res. **24** (2022) e36823, https://doi.org/10.2196/36823
[24] TEJANI, A.S., et al., Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update, Radiol. Artif. Intell. **6** (2024) e240300, https://doi.org/10.1148/ryai.240300
[25] LIU, X., et al., Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension, Nat. Med. **26** (2020) 1364–1374, https://doi.org/10.1038/s41591-020-1034-x
[26] VASEY, B., et al., Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI, Nat. Med. **28** (2022) 924–933, https://doi.org/10.1038/s41591-022-01772-9
[27] LEKADIR, K., et al., FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare, BMJ **388** (2025) e081554, https://doi.org/10.1136/bmj-2024-081554

[28] MOONS, K.G.M., et al., PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods, BMJ **388** (2025) e082505, https://doi.org/10.1136/bmj-2024-082505

[29] MOONS, K.G.M., et al., PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration, Ann. Intern. Med. **170** (2019) W1–W33, https://doi.org/10.7326/M18-1377

[30] CRUZ RIVERA, S., et al., Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension, Nat. Med. **26** (2020) 1351–1363, https://doi.org/10.1038/s41591-020-1037-7

[31] BARRAGÁN-MONTERO, A., et al., Towards a safe and efficient clinical implementation of machine learning in radiation oncology by exploring model interpretability, explainability and data-model dependency, Phys. Med. Biol. **67** (2022) 11TR01, https://doi.org/10.1088/1361-6560/ac678a

[32] NG, M.Y., et al., Perceptions of data set experts on important characteristics of health data sets ready for machine learning, JAMA Netw. Open **6** (2023) e2345892, https://doi.org/10.1001/jamanetworkopen.2023.45892

[33] JHA, A.K., et al., Radiomics: A quantitative imaging biomarker in precision oncology, Nucl. Med. Commun. **43** (2022) 483–493, https://doi.org/10.1097/MNM.0000000000001543

[34] NIOCHE, C., et al., LIFEx: A freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity, Cancer Res. **78** (2018) 4786–4789, https://doi.org/10.1158/0008-5472.CAN-18-0125

[35] VAN GRIETHUYSEN, J.J.M., et al., Computational radiomics system to decode the radiographic phenotype, Cancer Res. **77** (2017) e104–e107, https://doi.org/10.1158/0008-5472.CAN-17-0339

[36] ZWANENBURG, A., et al., The Image Biomarker Standardization Initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping, Radiology **295** (2020) 328–338, https://doi.org/10.1148/radiol.2020191145

[37] PAI, S., et al., Foundation model for cancer imaging biomarkers, Nat. Mach. Intell. **6** (2024) 354–367, https://doi.org/10.1038/s42256-024-00807-9

[38] SAHINER, B., et al., Deep learning in medical imaging and radiation therapy, Med. Phys. **46** (2019) e1–e36, https://doi.org/10.1002/mp.13264

[39] BAUGHAN, N., et al., Sequestration of imaging studies in MIDRC: Stratified sampling to balance demographic characteristics of patients in a multi-institutional data commons, J. Med. Imaging (Bellingham) **10** (2023) 064501, https://doi.org/10.1117/1.JMI.10.6.064501

[40] MAIER-HEIN, L., et al., Metrics reloaded: Recommendations for image analysis validation, Nat. Methods **21** (2024) 195–212, https://doi.org/10.1038/s41592-023-02151-z

[41] CHEN, R.J., et al., Towards a general-purpose foundation model for computational pathology, Nat. Med. **30** (2024) 850–862, https://doi.org/10.1038/s41591-024-02857-3

[42] LU, M.Y., et al., A visual-language foundation model for computational pathology, Nat. Med. **30** (2024) 863–874, https://doi.org/10.1038/s41591-024-02856-4

[43] KIM, C., et al., Fostering transparent medical image AI via an image-text foundation model grounded in medical literature, medRxiv [preprint] (2023), https://doi.org/10.1101/2023.06.07.23291119

[44] OSUALA, R., et al., medigan: A Python library of pretrained generative models for medical image synthesis, J. Med. Imaging (Bellingham) **10** (2023) 061403, https://doi.org/10.1117/1.JMI.10.6.061403

[45] GUPTA, N., et al., Exploring prospects, hurdles, and road ahead for generative artificial intelligence in orthopedic education and training, BMC Med. Educ. **24** (2024) 1544, https://doi.org/10.1186/s12909-024-06592-8

[46] KUMAR, A., BURR, P., YOUNG, T.M., Using AI text-to-image generation to create novel illustrations for medical education: Current limitations as illustrated by hypothyroidism and Horner syndrome, JMIR Med. Educ. **10** (2024) e52155, https://doi.org/10.2196/52155

[47] AKINCI D'ANTONOLI, T., et al., Large language models in radiology: Fundamentals, applications, ethical considerations, risks, and future directions, Diagn. Interv. Radiol. **30** (2024) 80–90, https://doi.org/10.4274/dir.2023.232417

[48] DRUKKER, K., et al., MIDRC-MetricTree: A decision tree-based tool for recommending performance metrics in artificial intelligence-assisted medical image analysis, J. Med. Imaging (Bellingham) **11** (2024) 024504, https://doi.org/10.1117/1.JMI.11.2.024504

[49] CHALLEN, R., et al., Artificial intelligence, bias and clinical safety, BMJ Qual. Saf. **28** (2019) 231–237, https://doi.org/10.1136/bmjqs-2018-008370

[50] HURKMANS, C., et al., A joint ESTRO and AAPM guideline for development, clinical validation and reporting of artificial intelligence models in radiation therapy, Radiother. Oncol. **197** (2024) 110345, https://doi.org/10.1016/j.radonc.2024.110345

[51] DRUKKER, K., et al., Toward fairness in artificial intelligence for medical image analysis: Identification and mitigation of potential biases in the roadmap from data collection to model deployment, J. Med. Imaging (Bellingham) **10** (2023) 061104, https://doi.org/10.1117/1.JMI.10.6.061104

[52] GLOCKER, B., JONES, C., BERNHARDT, M., WINZECK, S., Algorithmic encoding of protected characteristics in chest X-ray disease detection models, EBioMedicine **89** (2023) 104467, https://doi.org/10.1016/j.ebiom.2023.104467

[53] REZK, E., ELTORKI, M., EL-DAKHAKHNI, W., Leveraging artificial intelligence to improve the diversity of dermatological skin color pathology: Protocol for an algorithm development and validation study, JMIR Res. Protoc. **11** (2022) e34896, https://doi.org/10.2196/34896

[54] WHITNEY, H.M., et al., Longitudinal assessment of demographic representativeness in the Medical Imaging and Data Resource Center open data commons, J. Med. Imaging (Bellingham) **10** (2023) 061105, https://doi.org/10.1117/1.JMI.10.6.061105

[55] INTERNATIONAL ATOMIC ENERGY AGENCY, Worldwide Implementation of Digital Imaging in Radiology, IAEA Human Health Series No. 28, IAEA, Vienna (2015).

[56] MAHADEVAIAH, G., et al., Artificial intelligence-based clinical decision support in modern medical physics: Selection, acceptance, commissioning, and quality assurance, Med. Phys. **47** (2020) e228–e235, https://doi.org/10.1002/mp.13562

[57] DAYE, D., et al., Implementation of clinical artificial intelligence in radiology: Who decides and how? Radiology **305** (2022) 555–563, https://doi.org/10.1148/radiol.212151

[58] CAMPBELL, R., The five "rights" of clinical decision support, J. AHIMA **84** (2013) 42–47.

[59] HAIBE-KAINS, B., et al., Transparency and reproducibility in artificial intelligence, Nature **586** (2020) E14–E16, https://doi.org/10.1038/s41586-020-2766-y

[60] HICKS, S.A., et al., On evaluation metrics for medical applications of artificial intelligence, Sci. Rep. **12** (2022) 5979, https://doi.org/10.1038/s41598-022-09954-8

[61] TARASEK, M., AKIN, O., ROBERTS, J., FOO, T., YEO, D., Heat modulation of intrinsic MR contrasts for tumor characterization, Cancers (Basel) **14** (2022) 405, https://doi.org/10.3390/cancers14020405

[62] DRATSCH, T., et al., Automation bias in mammography: The impact of artificial intelligence BI-RADS suggestions on reader performance, Radiology **307** (2023) e222176, https://doi.org/10.1148/radiol.222176

[63] LUSTBERG, T., et al., Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer, Radiother. Oncol. **126** (2018) 312–317, https://doi.org/10.1016/j.radonc.2017.11.012

[64] McINTOSH, C., et al., Clinical integration of machine learning for curative-intent radiation treatment of patients with prostate cancer, Nat. Med. **27** (2021) 999–1005, https://doi.org/10.1038/s41591-021-01359-w

[65] MONTAGUE, E., et al., How long does contouring really take? Results of the Royal College of Radiologists contouring surveys, Clin. Oncol. **36** (2024) 335–342, https://doi.org/10.1016/j.clon.2024.03.005

[66] TAHA, A.A., HANBURY, A., Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool, BMC Med. Imaging **15** (2015) 29, https://doi.org/10.1186/s12880-015-0068-x

[67] LUCA, A.R., et al., Impact of quality, type and volume of data used by deep learning models in the analysis of medical images, Inform. Med. Unlocked **29** (2022) 100911, https://doi.org/10.1016/j.imu.2022.100911

[68] LI, B., KUMAR, S., Managing software-as-a-service: Pricing and operations, Prod. Oper. Manag. **31** (2022) 2588–2608, https://doi.org/10.1111/poms.13729

[69] SHAPOURI, F., WARD, K., SETOR, T., Determinants of software as a service (SaaS) adoption, J. Comput. Inf. Syst. **64** (2024) 301–313, https://doi.org/10.1080/08874417.2023.2199270

[70] FUHRMAN, J.D., et al., A review of explainable and interpretable AI with applications in COVID-19 imaging, Med. Phys. **49** (2022) 1–14, https://doi.org/10.1002/mp.15359

[71] ALLEN, B., et al., Evaluation and real-world performance monitoring of artificial intelligence models in clinical practice: Try it, buy it, check it, J. Am. Coll. Radiol. **18** (2021) 1489–1496, https://doi.org/10.1016/j.jacr.2021.08.022

[72] WILSON, D.U., BAILEY, M.Q., CRAIG, J., The role of artificial intelligence in clinical imaging and workflows, Vet. Radiol. Ultrasound **63** Suppl. 1 (2022) 897–902, https://doi.org/10.1111/vru.13157

[73] FANG, H., SHI, K., WANG, X., ZUO, C., LAN, X., Editorial: Artificial intelligence in positron emission tomography, Front. Med. (Lausanne) **9** (2022) 848336, https://doi.org/10.3389/fmed.2022.848336

[74] JHA, A.K., MITHUN, S., RANGARAJAN, V., WEE, L., DEKKER, A., Emerging role of artificial intelligence in nuclear medicine, Nucl. Med. Commun. **42** (2021) 592–601, https://doi.org/10.1097/MNM.0000000000001381

[75] CHOUDHURY, A., CHAUDHRY, Z., Large language models and user trust: Consequence of self-referential learning loop and the deskilling of health care professionals, J. Med. Internet Res. **26** (2024) e56764, https://doi.org/10.2196/56764

[76] UNITED STATES FOOD AND DRUG ADMINISTRATION, Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan, FDA, Silver Spring, MD (2021).

[77] INTERNATIONAL ATOMIC ENERGY AGENCY, Quality Assurance Programme for Computed Tomography: Diagnostic and Therapy Applications, IAEA Human Health Series No. 19, IAEA, Vienna (2012).

[78] McCONALOGUE, E., DAVIS, P., CONNOLLY, R., Health technology assessment: The role of total cost of ownership, Bus. Syst. Res. J. **10** (2019) 180–187, https://doi.org/10.2478/bsrj-2019-0013

[79] LÖWE, A., et al., “Knobology” in Doppler ultrasound, Med. Ultrason. **23** (2021) 480–486, https://doi.org/10.11152/mu-3216

[80] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, INTERNATIONAL ELECTROTECHNICAL COMMISSION, Information Security, Cybersecurity and Privacy Protection, ISO/IEC 27001:2022, ISO/IEC, Geneva (2022).

[81] BENJAMENS, S., DHUNNOO, P., MESKÓ, B., The state of artificial intelligence-based FDA-approved medical devices and algorithms: An online database, NPJ Digit. Med. **3** (2020) 118, https://doi.org/10.1038/s41746-020-00324-0

[82] EUROPEAN PARLIAMENT AND COUNCIL OF THE EUROPEAN UNION, Regulation (EU) 2024/1689 of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act), Official Journal of the European Union, Publications Office of the European Union, Luxembourg (2024).

[83] PRAJWAL, R., et al., A study on energy consumption in AI-driven medical image segmentation, J. Imaging **11** (2025) 174, https://doi.org/10.3390/jimaging11060174

[84] TRUHN, D., MÜLLER-FRANZES, G., KATHER, J.N., The ecological footprint of medical AI, Eur. Radiol. **34** (2024) 1176–1178, https://doi.org/10.1007/s00330-023-10123-2

[85] EUROPEAN PARLIAMENT AND COUNCIL OF THE EUROPEAN UNION, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), Official Journal of the European Union, Publications Office of the European Union, Luxembourg (2016).

[86] MAHMOOD, U., et al., Artificial intelligence in medicine: Mitigating risks and maximizing benefits via quality assurance, quality control, and acceptance testing, BJR Artif. Intell. **1** (2024) ubae003, https://doi.org/10.1093/bjrai/ubae003

[87] CRUZ, J.P., KAJI, Y., YANAI, N., RBAC-SC: Role-based access control using smart contract, IEEE Access **6** (2018) 12240–12251, https://doi.org/10.1109/Access.2018.2812844

[88] UMEJIAKU, A.P., DHAKAL, P., SHENG, V.S., Balancing password security and user convenience: Exploring the potential of prompt models for password generation, Electronics (Basel) **12** (2023) 2159, https://doi.org/10.3390/electronics12102159

[89] WILLEMINK, M.J., et al., Preparing medical imaging data for machine learning, Radiology **295** (2020) 4–15, https://doi.org/10.1148/radiol.2020192224

[90] LUTTERMANN, M., et al., Automated computation of therapies using failure mode and effects analysis in the medical domain, Künstl. Intell. **38** (2024) 189–201, https://doi.org/10.1007/s13218-023-00810-z

[91] TAYLOR, M.J., et al., Systematic review of the application of the plan-do-study-act method to improve quality in healthcare, BMJ Qual. Saf. **23** (2014) 290–298, https://doi.org/10.1136/bmjqs-2013-001862

[92] AMURAO, M., et al., Quality management, quality assurance, and quality control in medical physics, J. Appl. Clin. Med. Phys. **24** (2023) e13885, https://doi.org/10.1002/acm2.13885

[93] INTERNATIONAL ATOMIC ENERGY AGENCY, Radiation Protection and Safety in Medical Uses of Ionizing Radiation, IAEA Safety Standards Series No. SSG-46, IAEA, Vienna (2018).

[94] UNITED STATES FOOD AND DRUG ADMINISTRATION, Mammography Quality Standards Act; Regulatory Amendments, FDA-2013-N-0134, FDA, Silver Spring, MD (2023).

[95] CLAESSENS, M., et al., Quality assurance for AI-based applications in radiation therapy, Semin. Radiat. Oncol. **32** (2022) 421–431, https://doi.org/10.1016/j.semradonc.2022.06.011

[96] KETOLA, J.H.J., et al., Testing process for artificial intelligence applications in radiology practice, Phys. Med. **128** (2024) 104842, https://doi.org/10.1016/j.ejmp.2024.104842

[97] VANDEWINCKELE, L., et al., Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance, Radiother. Oncol. **153** (2020) 55–66, https://doi.org/10.1016/j.radonc.2020.09.008

[98] HUQ, M.S., et al., The report of Task Group 100 of the AAPM: Application of risk analysis methods to radiation therapy quality management, Med. Phys. **43** (2016) 4209–4262, https://doi.org/10.1118/1.4947547

[99] ZAREI, M., GERSHAN, V., HOLMBERG, O., Safety in radiation oncology (SAFRON): Learning about incident causes and safety barriers in external beam radiotherapy, Phys. Med. **111** (2023) 102618, https://doi.org/10.1016/j.ejmp.2023.102618

[100] ALOWAIS, S.A., et al., Revolutionizing healthcare: The role of artificial intelligence in clinical practice, BMC Med. Educ. **23** (2023) 689, https://doi.org/10.1186/s12909-023-04698-z

[101] MITCHELL, M., et al., "Model cards for model reporting", Proc. Conf. on Fairness, Accountability, and Transparency (FAT* '19), Atlanta, GA, 2019, Association for Computing Machinery, New York (2019) 220–229, https://doi.org/10.1145/3287560.3287596

GLOSSARY

The majority of the definitions given below are derived from existing IAEA and WHO terminology, or from open source resources pertaining to the application of AI systems in healthcare. The context specific definitions presented here might not necessarily conform to definitions adopted elsewhere for international use.

acceptance testing. Verification of equipment specifications and features by representatives of the installer and the facility medical physicist. This includes rigorous testing to ensure that the equipment performs as expected and complies with all safety and regulatory standards. Acceptance testing is a critical step before equipment can be commissioned and used for patient care. It includes checking the performance specifications, confirming the safety features and ensuring that the equipment integrates properly with existing systems.

accuracy. In the context of data analysis and machine learning, a measure of how often a model correctly predicts or classifies cases. It is calculated as the ratio of correct predictions (both true positives and true negatives) to the total number of cases examined. High accuracy indicates that the model performs well across both positive and negative cases, but it might not be an adequate metric if the dataset is imbalanced between classes.

algorithm. A defined set of instructions or rules that specifies a sequence of operations for solving a particular computational problem across all instances within a given problem set.

area under the precision-recall curve (AUPR) (*or* precision-recall area under the curve, PRAUC). A metric used for datasets with a significant imbalance between classes. Measuring the area under the precision-recall curve provides insight into the trade-off between precision and recall for different thresholds.

area under the receiver operating characteristic curve (AUROC, AUC-ROC *or* AUC). A metric that measures the ability of a classification model to differentiate between classes. The receiver operating characteristic (ROC) curve plots the true positive rate against the false positive rate at various thresholds/operating points, and the AUROC quantifies this performance

with a single value, where 1.0 indicates perfect classification and 0.5 indicates performance equivalent to random guessing.

artificial intelligence (AI). An array of technologies that enable a machine or computer agent to perform tasks that usually use human intelligence, such as sensing, comprehending, learning, decision making and acting. Broadly speaking, AI encompasses computer aided diagnosis, radiomics, machine learning (including deep learning and foundation models), computer vision, expert systems and natural language processing. In recent years, new breakthroughs in deep learning have greatly accelerated AI applications and enabled the leverage of large datasets and existing experience to render workflows more efficient and to automate tasks such as prediction, detection, classification, semantic transcription, image reconstruction and processing, and sensorimotor control.

artificial intelligence (AI) system. In the context of this publication, an AI application, AI subsystem or AI software as a medical device.

bias. A measure of the systematic error between an actual or true value and a prediction by a model or a measured mean value. The bias of a model represents the tendency of a model to overpredict or to underpredict. Diverse data collection and curation strategies, together with appropriate mitigation measures during data analysis, are critically important to yield ethical, generalizable AI algorithms that produce trustworthy results across different imaging equipment, healthcare settings and patients. Bias can arise at many various stages of data input, algorithm development and output usage.

clinically qualified medical physicist (CQMP). A health professional who has completed appropriate undergraduate education in the physical or engineering sciences, followed by professional competency training that includes one to three years of postgraduate academic education in medical physics and at least two further years of structured practical training in a clinical environment in one or more medical physics specialties.

cloud computing. A model of IT infrastructure management (see also *on-premises computing*). Cloud computing delivers services such as servers, storage and applications over the Internet, offering scalability, remote accessibility and a pay-as-you-go cost model, with the service provider managing the infrastructure.

commissioning. A process carried out by the facility representative, usually a medical physicist, to ensure that equipment is ready for clinical use and to establish baseline values against which the results of subsequent routine performance tests can be compared.

computed tomography (CT). A medical imaging technique that uses X ray technology to create detailed cross-sectional images (slices) of the body. CT images are particularly useful for diagnosing various medical conditions, including bone fractures, tumours and vascular diseases. A CT scanner acquires these images by rotating an X ray tube and detector around the patient and reconstructing the resulting individual projections into two dimensional or three dimensional images. CT imaging provides superior contrast and volumetric information compared with conventional radiography, although at a higher radiation dose.

computer aided detection (CADe). A computer based approach focused on automatically identifying and highlighting potential abnormalities or lesions within medical images. CADe systems are commonly used in radiology to assist radiologists in detecting tumours, fractures or other anomalies in X rays, mammograms and other imaging modalities, especially in screening programmes where most patients are normal. In the context of radiation oncology, CADe can be used to assist in identifying target structures and critical organs in treatment planning CT scans, ensuring accurate and precise radiotherapy delivery.

computer aided diagnostics (CADx). A subfield of AI in medicine involving software and algorithms to assist healthcare practitioners in diagnosing diseases or medical conditions. CADx systems analyse medical data, such as imaging studies (e.g. X rays, CT scans, MRI), to diagnose suspect abnormalities, highlight suspicious areas and provide quantitative assessments. In medical physics, CADx can be applied to help clinical medical physicists analyse medical images, assess treatment plans and identify potential issues or deviations from the norm.

computer aided triage (CADt). The use of computer based systems, often powered by AI and machine learning, to assist healthcare professionals in prioritizing patients based on the severity of their medical conditions. CADt helps in allocating resources efficiently and ensuring that the most critical cases receive immediate attention.

concordance index (CI *or* C-index). A measure of the predictive accuracy of a model, used in both classification and survival analysis. In classification, the CI quantifies how well a model discriminates between different outcome classes, essentially reflecting the probability that, for a randomly chosen pair of samples, the model correctly predicts which one has the higher risk or likelihood of the event. In survival analysis, the CI extends this concept to time-to-event data by assessing the model's ability to correctly rank patients according to their predicted survival times, with higher values indicating better predictive accuracy. The CI accounts for censored data, which is crucial in survival analysis.

concurrent read artificial intelligence (AI) system. An AI system that serves as a concurrent reader, allowing clinicians to view both the images and the AI system output concurrently and to use both in making their initial interpretation or decision.

configuration. The adaptation of a computer system to a customer's preferences and operating environment. Configuration is performed using the flexibility built into the software and is normally a standard part of the software installation process.

convolutional neural network. A deep learning architecture that applies convolutional layers to structured data, such as images, to detect and extract spatial features.

Cox proportional hazard ratio. The hazard ratio estimated from the Cox proportional hazards model, used in survival analysis as a measure of the effect of an explanatory variable (e.g. a treatment or risk factor) on the hazard or risk of an event occurring, such as death or disease progression. This ratio compares the hazard in two groups: those with the explanatory variable (e.g. a treatment group) and those without it (e.g. a control group). A hazard ratio greater than one indicates an increased risk of the event in the group with the variable, while a ratio less than one indicates a reduced risk. The Cox model assumes that this ratio remains constant over time, a property known as 'proportional hazards'. It is a semiparametric model, meaning it does not require specifying the underlying hazard function, allowing for flexibility in analysing survival data.

customization. The adaptation of a computer system to a customer's preferences using custom designed modifications. Such modifications can be costly to

procure and maintain, so customization is typically undertaken only when configuration cannot provide the functionality that the customer needs.

data maturity. A measure of an organization's ability to effectively manage, use and leverage its data. It encompasses the quality and accuracy of data, efficient data management practices, the integration and accessibility of data across the organization and the use of analytical tools for data driven decision making. It also involves cultivating a data driven culture in which data are valued as a strategic asset. High data maturity signifies an organization's capability to harness data for innovation, efficiency and competitive advantage.

deep learning. A subset of machine learning that uses neural networks as its set of functions from inputs to outputs, involving artificial neural networks with multiple layers (deep neural networks). Deep learning aims to automatically learn and extract hierarchical representations of data, enabling the model to recognize complex patterns and features, without explicit programming. Deep learning has been highly successful in various AI applications, including image and speech recognition.

deskilling. A loss of proficiency that occurs when end users become so reliant on an AI system's output that they forget how to perform the clinical task independently.

detection, diagnosis, prognosis, prediction. A set of related clinical concepts referring to different stages of interpreting medical information:

— **Detection:** Identifying a disease, such as a tumour, or another object of interest in medical images;

— **Diagnosis:** Not only detecting an object of interest (e.g. disease) but also characterizing it, for example determining the type of cancer;

— **Prognosis:** Predicting the likely course or outcome of a disease once it has been diagnosed;

— **Prediction:** Forecasting the response of a tumour to radiation or predicting potential side effects in patients based on their individual characteristics and treatment parameters.

development, testing, acceptance and production (DTAP). A software development and IT framework that structures the life cycle of software and systems to ensure systematic progression, quality, reliability and user satisfaction in software development and deployment. It comprises four stages:

- **Development:** The initial phase, in which software is developed according to project requirements.
- **Testing:** Software is evaluated through various tests to identify and fix bugs, ensuring functionality and performance.
- **Acceptance:** Also called staging or pre-production; software is validated in an environment that simulates real world use, often including user acceptance testing.
- **Production:** The final stage, in which the software is deployed for actual use in a live environment.

Digital Imaging and Communications in Medicine (DICOM) standard. A standard for the management of information (including images) in medical imaging. The DICOM standard is based on industry standards such as the TCP/IP network protocol and has been developed for a wide range of imaging systems (https://www.dicomstandard.org/).

dimensionality reduction. A method for transforming high dimensional data into lower dimensional data. Such methods can be used to reduce the number of computer extracted image features prior to input into a machine learning algorithm for training (i.e. feature selection).

effectiveness. The ability of a medical device to achieve clinically meaningful outcomes in its intended use as claimed by the manufacturer.

efficacy (of healthcare). The ability of a healthcare process to achieve a desired clinically meaningful outcome.

efficiency (of healthcare). The ability to complete a healthcare process given a certain amount of time and resources (human resources and others).

electronic health record (EHR). A digital collection of a person's medical information stored on a computer. An EHR includes information about a

patient's health history, such as diagnoses, medications, tests, allergies, immunizations and treatment plans. EHRs can be accessed by all healthcare providers involved in a patient's care and can be used to support clinical decision making. Also called an electronic medical record.

explainability. In the context of AI and machine learning, the ability of a model or system to provide understandable explanations for its predictions or decisions at various stages of the AI system's computations and decision making processes.

F1 score. A performance metric defined as the harmonic mean of precision and recall, providing a single metric that balances both. The F1 score is particularly useful when a balance between precision and recall is desired and in situations where an imbalance in the dataset might result in misleading accuracy metrics.

false negative. An outcome in which a model incorrectly predicts the negative class. In a medical context, a false negative occurs when a test fails to detect a disease or condition that is actually present.

false positive. An outcome in which a model incorrectly predicts the positive class. In a medical context, a false positive occurs when a test indicates the presence of a disease or condition in a patient who does not have it.

fine tuning. A machine learning technique in which an already trained model in an AI system is further trained on a specific task or dataset to adapt it for a new, related task (e.g. tailoring a model to a specific healthcare organization). Fine tuning involves updating the model parameters using a smaller dataset, often with a lower learning rate, allowing the model to retain the knowledge learned during the initial pretraining while adjusting its representations to better suit the target task.

foundation model. A large-scale AI model trained on extensive and diverse datasets, enabling it to develop a broad understanding of various topics, concepts and skills. Foundation models serve as a foundational framework upon which specialized (narrow) AI applications can be built.

free-response receiver operating characteristic (FROC). A method for assessing the accuracy of diagnostic systems that need to identify and localize multiple instances of targets within images. In medical imaging,

FROC analysis assesses how well a system correctly identifies and localizes abnormalities, such as tumours, while minimizing false positives.

ground truth. See *reference standard*.

Hausdorff distance. A performance metric that quantifies how close two subsets of a given space are to each other. Specifically, it calculates the greatest distance from any point in one set to the closest point in the other set. The Hausdorff distance is widely used in image analysis to assess the similarity or dissimilarity between two shapes or datasets. Its versatility allows it to be applied in contexts where assessing the degree of resemblance or alignment between sets is crucial.

HL7 Fast Healthcare Interoperability Resources (FHIR). A standard developed within the HL7 standards framework to enable seamless electronic communication between different healthcare systems, such as hospitals, laboratories, insurance companies and mobile applications. HL7 FHIR defines structured data formats and elements, referred to as called 'resources', and provides an interface for exchanging them (https://www.fhir.org).

HL7 standards. A family of international standards widely used for the exchange, integration, sharing and retrieval of electronic health information. HL7 standards are continually developed and maintained by Health Level Seven International (United States of America), with HL7 Version 3 now available (https://www.hl7.org).

hospital information system (HIS). A comprehensive, integrated software system designed to manage the medical, administrative, financial and legal information of a hospital. An HIS typically serves as a central repository for patient demographic and administrative data and often provides patient data as input for other clinical systems, such as the radiology information system (RIS) and picture archiving and communication systems (PACSs).

image reconstruction. The process of generating a two or three dimensional image from a set of observed data, typically acquired from imaging modalities such as computed tomography (CT), magnetic resonance imaging (MRI) or positron emission tomography (PET). Mathematical algorithms transform raw data (such as individual detector readings) into a coherent image, which is then used for diagnosis or treatment planning.

In radiotherapy, accurate image reconstruction is critical for delineating tumour volumes and planning treatment paths.

input data. The data that need to be provided to an AI system so it can perform its intended function.

intended population. The population in which an AI system is developed to fulfil its intended use as specified by the manufacturer. For example, if the intended use is automated segmentation of organs at risk in the head and neck region, the intended population could be adult patients with a tumour in the head and neck region who are receiving external beam radiotherapy.

intended use. The use case or cases explicitly described by the manufacturer of the AI system and assessed in their claim (i.e. what the AI system does); for example, automated segmentation of organs at risk in computed tomography scans of the head and neck region.

interpretability. The degree to which a model's predictions or behaviour can be understood and interpreted by humans, typically data domain experts or end users; related to explainability. An interpretable model is one whose output is understandable by the end user.

log rank test for Kaplan–Meier survival curves. A statistical method used in survival analysis to compare Kaplan–Meier survival curves across two or more groups. It tests the null hypothesis that there is no difference in survival between the groups over the study period. The log rank test focuses on the number of events (e.g. deaths or relapses) and the expected number of events in each group at various time points, taking into account censored data (i.e. subjects who have not yet had the event or are lost to follow-up). A significant result suggests a difference in survival experience between the groups.

machine learning. A subfield of AI that centres on using computer algorithms to draw inferences and find patterns in data. The algorithm learns by extracting essential features from data and then makes decisions based on inference. Machine learning focuses on developing algorithms and models that enable computers to learn from data and make predictions or decisions. It encompasses a wide range of techniques, including supervised learning (e.g. predictive modelling), unsupervised learning (e.g. clustering and dimensionality reduction), self-supervised learning (used in many

foundation models) and reinforcement learning (decision making through trial and error).

magnetic resonance imaging (MRI). A medical imaging technique that uses strong magnetic fields to generate detailed images of the internal structures of the body.

medical physicist. A health professional with specialized education and training in the concepts and techniques of applying physics in medicine, who is competent to practise independently in one or more of the subfields (specialties) of medical physics.

negative predictive value (NPV). The proportion of negative identifications made by a model that are actually correct, calculated as the number of true negatives divided by the sum of true and false negatives.

on-premises computing. A model of IT infrastructure management (see also *cloud computing*). On-premises computing involves setting up and maintaining IT infrastructure within a healthcare organization's physical location, providing full control and security over the data and systems but requiring significant upfront capital investment and ongoing maintenance.

oncology information system (OIS). A comprehensive information system designed specifically to support the management and treatment of patients with cancer. It integrates patient data across radiology, pathology and pharmacy to provide a comprehensive record of the patient's oncology treatment, including data on chemotherapy, radiotherapy and other treatment modalities. An OIS helps in treatment planning, managing therapy cycles, monitoring patient progress and documenting outcomes. It is crucial for ensuring coordinated care and maintaining detailed treatment records for oncology patients.

output data. The information produced by an AI system after processing the input data. Output data can be presented in various forms, such as a classification (including diagnosis, disease severity or stage, or a recommendation such as referability), a probability, a class activation map or an image.

picture archiving and communication system (PACS). A computer system for storing, displaying and transmitting medical images. A PACS is often combined with a radiology information system (RIS), enabling the display of images alongside clinical information and final diagnoses.

positive predictive value (PPV) (*or* precision). The proportion of positive identifications made by a model that are actually correct, calculated as the number of true positives divided by the sum of true and false positives. PPV is particularly important when the cost of false positives is high.

positron emission tomography (PET). A medical imaging technique that uses a small amount of radioactive material (a radiotracer) to visualize and measure metabolic and biochemical processes within the body. PET scans are commonly used for diagnosing and monitoring conditions such as cancer, brain disorders and heart diseases.

post-market surveillance. A systematic process for collecting and analysing information on the performance of medical devices after they have been placed on the market.

precision. See *positive predictive value*.

quality assurance (QA). The function of a management system that provides confidence that specified requirements will be fulfilled. A QA programme establishes the procedures and protocols needed to ensure that the equipment and processes used in patient care meet established standards. Within the context of AI systems, QA ensures that the system is used as intended and that appropriate quality control activities are implemented to verify that it functions as claimed, ensuring both safety and effectiveness.

quality control (QC). Part of quality management intended to verify that structures, systems and components correspond to predetermined requirements. In the context of AI systems, QC refers to the tests and monitoring activities implemented to ensure that AI systems used in clinical care perform accurately, safely, consistently and as intended throughout their entire life cycle.

radiological medical practitioner. A health professional with specialized education and training in the medical uses of radiation, who is competent to perform independently or to oversee radiological procedures within a given specialty.

radiology information system (RIS). A networked software system designed for managing medical imagery and associated data. An RIS is particularly useful for tracking radiology imaging orders and billing information and is often used in conjunction with a picture archiving and communication

system (PACS) and vendor neutral archives to manage image archives, record keeping and billing. An RIS facilitates management of patient schedules, resource management, examination performance tracking, reporting, results distribution and procedure billing.

radiomics. A method that involves automated processing of medical images and the computerized extraction of predetermined quantitative features (e.g. texture, shape and intensity) or ad hoc features (using deep learning) from medical images, such as those obtained from mammography, positron emission tomography (PET), computed tomography (CT) and magnetic resonance imaging (MRI). These features can be used to characterize tumours, tissues or organs, and can provide valuable information for diagnosis, treatment planning and outcome prediction. Their analysis is commonly facilitated by machine learning and deep learning methods.

recall. A performance metric, also known as sensitivity, that measures how well a model identifies all relevant positive cases. It is calculated as the ratio of true positive predictions to the total number of actual positive cases (true positives plus false negatives). High recall indicates that the model is effective at capturing the majority of positive cases.

reference standard (*or* ground truth). The 'true' outcome, diagnosis or measurement as determined using well established means. It is used to compare AI predicted measures with the actually observed measures and forms the basis for validating the performance of an AI application. The concept extends beyond physics based reference standards provided by standards laboratories and includes clinical outcomes, expert opinions, images and other authoritative sources.

region of interest. A specific area or subset of an image or dataset that is selected for closer examination or analysis. Regions of interest are often defined to focus on particular regions or structures (e.g. lesions or nodules) within an image, enabling quantitative measurements or detailed evaluation.

remote service. The provision of troubleshooting and technical support through computer networks, allowing the service provider to operate from elsewhere in the hospital or from the other side of the world. This approach is commonly used for IT based imaging systems such as computed tomography (CT) and magnetic resonance imaging (MRI), as well as for picture archiving and communication systems (PACSs).

repeatability. The closeness of agreement between independent test results obtained under the same conditions. In the context of quantitative imaging biomarkers, the repeatability is the closeness of agreement between measurements of the same biomarker, made under the same conditions, over a short period of time.[4]

reproducibility. The degree to which an experiment or measurement yields consistent results when conducted under different conditions (e.g. different operators, equipment, locations or times). In the context of quantitative imaging biomarkers, the reproducibility is the closeness of agreement between measurements of the same biomarker obtained under different conditions (e.g. different imaging systems, sites, operators or time points) when no true biological change has occurred.

rule-out AI system. An AI system that automatically removes normal or negative cases from the case list, so that only cases requiring end user decision making remain.

second read AI system. An AI system that provides its output only after the clinicians have made an initial interpretation or decision (i.e. prior to seeing the output from the AI system), thereby acting as a second reader to support or refine the clinicians' assessment.

segmentation (*or* delineation). In medical imaging, the process of identifying and outlining specific regions of interest within an image. For example, in radiotherapy, segmentation is used to delineate tumour boundaries and critical organs in planning images. This is a crucial step in treatment planning, ensuring that the radiation dose is accurately targeted to the tumour while minimizing exposure to healthy tissues.

software as a service (SaaS). A model in which the consumer uses a service provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through a thin client interface, such as a web browser (e.g. web based email), or through a program interface. The consumer does not manage or control the underlying cloud infrastructure (including network, servers, operating systems, storage or

[4] RAUNIG, D.L., et al., Quantitative imaging biomarkers: A review of statistical methods for technical performance assessment, Stat. Methods Med. Res. **24** (2015) 27–67, https://doi.org/10.1177/0962280214537344

even individual application capabilities), with the possible exception of limited user specific application configuration settings.

software environment. In the context of software development and deployment across different settings, the specific configuration and conditions in which a software application operates or is tested. Software environments are typically tailored to meet the requirements and constraints of different stages in the software development and deployment life cycle, such as testing, acceptance and clinical production. Each of these software environments serves a distinct purpose, with specific configurations and considerations tailored to their respective roles.

— **Testing environment:** A controlled and isolated environment in which the AI system can be thoroughly tested. It replicates aspects of the production environment but allows for experimentation, testing and debugging without affecting the clinical routine.

— **Acceptance environment:** A pre-production environment used during acceptance and commissioning to validate that the AI system meets requirements and expectations. It closely mimics the production environment to ensure proper acceptance and commissioning of the AI system.

— **Clinical production environment:** The operational environment in which the AI system is deployed and used by end users. It represents the real world setting in which the AI system performs reliably and efficiently.

Sørensen–Dice similarity coefficient (*or* Dice coefficient). A statistical measure used to quantify the similarity between two sets. The coefficient is calculated by taking the size of the intersection of the sets, multiplying it by 2 and then dividing that number by the sum of the sizes of the two sets. Typically ranging from 0 (indicating no similarity) to 1 (indicating complete similarity), this coefficient is particularly useful in data analysis, ecology and computational biology for comparing the composition of different datasets, as well as in natural language processing and information retrieval for assessing text similarity.

structured report. A standardized, structured method, defined within the Digital Imaging and Communications in Medicine (DICOM) standard, for exchanging data produced in the course of image acquisition,

post-processing and reporting. Structured reports use DICOM data elements and DICOM network services, such as storage and query/retrieve.

time dependent receiver operating characteristic (ROC) curve. A method in survival analysis used to evaluate the predictive accuracy of models over time, particularly for censored survival data. Unlike traditional ROC curves, which assess diagnostic tests at a single point in time, the time dependent ROC curve considers the probability of an event (e.g. death or disease recurrence) at multiple time points. It plots the true positive rate (sensitivity) against the false positive rate (1 − specificity) for predictions made at various times.

transfer learning. An approach in machine learning in which a model trained on one task or dataset is reused or adapted to perform a different but related task or work with a different dataset (e.g. a model for tissue segmentation is adapted to tissue characterization). The knowledge and representations learned in the source task or dataset are leveraged to improve performance in the target task, often shortening training time and reducing the amount of labelled data needed for the target task.

transparency. The degree to which the inner workings and decision making processes of a model are visible and comprehensible. In the context of AI and machine learning, transparency often refers to the availability and explainability of the computer code and documentation, and it is crucial for building trust and ensuring accountability.

treatment planning system (TPS). A device, usually a programmable electronic system, used to simulate and plan the delivery of radiation to a patient for a proposed radiotherapy treatment. In this context, the TPS usually consists of hardware, a computer operating system and specialized TPS software. Also referred to as a radiation treatment planning system (RTPS).

validation. Objective evidence that the requirements for a specific intended use have been fulfilled.

verification. Objective evidence that the specific requirements have been fulfilled.

ABBREVIATIONS

AI	artificial intelligence
API	application programming interface
CE	European conformity mark
CLAIM	Checklist for Artificial Intelligence in Medical Imaging
CQMP	clinically qualified medical physicist
CT	computed tomography
DICOM	Digital Imaging and Communications in Medicine
DTAP	development, testing, acceptance and production
EHR	electronic health record
FDA	Food and Drug Administration (United States of America)
FHIR	Fast Healthcare Interoperability Resources
FMEA	failure mode and effects analysis
HIS	hospital information system
HL7	Health Level Seven
ISO	International Organization for Standardization
IT	information technology
MRI	magnetic resonance imaging
OIS	oncology information system
PACS	picture archiving and communication system
PET	positron emission tomography
QA	quality assurance
QC	quality control
RIS	radiology information system
SaaS	software as a service
SPECT	single photon emission computed tomography
STARD-AI	Standards for Reporting of Diagnostic Accuracy Studies — AI extension
TPS	treatment planning system
TRIPOD-AI	Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis — AI extension
WHO	World Health Organization

CONTRIBUTORS TO DRAFTING AND REVIEW

Avanzo, M.	European Federation of Organisations for Medical Physics
Azangwe, G.	International Atomic Energy Agency
Brouwer, C.L.	European Society for Radiotherapy and Oncology
Carrara, M.	International Atomic Energy Agency
Ciraj-Bjelac, O.	International Atomic Energy Agency
Crijns, W.	European Federation of Organisations for Medical Physics; European Society for Radiotherapy and Oncology
Dekker, A.	Maastro Clinic; Maastricht University; Maastricht UMC+, Kingdom of the Netherlands
Diaz, O.	European Federation of Organisations for Medical Physics
Ewert, K.	Australasian College of Physical Scientists and Engineers in Medicine
Fidarova, E.	International Atomic Energy Agency
Gichoya, J.	Emory University, United States of America
Giger, M.L.	University of Chicago, United States of America
González, J.	Latin American Association of Medical Physics
Guidi, G.	European Federation of Organisations for Medical Physics
Haibe-Kains, B.	Princess Margaret Cancer Centre, University of Toronto; Structural Genomics Consortium; Vector Institute for Artificial Intelligence, Canada
Holloway, L.	Liverpool Cancer Therapy Centre and Macarthur Cancer Therapy Centre, Australia; Australasian College of Physical Scientists and Engineers in Medicine

Hu, W.	Fudan University Shanghai Cance Center, China
Jha, A.K.	Tata Memorial Hospital, India
Kendrick, J.	Australasian College of Physical Scientists and Engineers in Medicine
Kortesniemi, M.	European Federation of Organisations for Medical Physics
Liu, F.F.	University of Toronto, Canada
Mahesh, M.	International Organization for Medical Physics
Mairal, M.L.	Latin American Association of Medical Physics
Namías, M.	Latin American Association of Medical Physics
Peszyńska-Piorun, M.	European Society for Radiotherapy and Oncology
Pinto dos Santos, D.	University Medical Center Mainz, Germany
Pirchio, R.	Latin American Association of Medical Physics
Reiser, I.	University of Chicago, United States of America
Rizk, C.	International Atomic Energy Agency
Swamidas, J.	International Atomic Energy Agency
Tadic, T.	University of Toronto, Canada
Titovich, E.	International Atomic Energy Agency
Trauernicht, C.	Federation of African Medical Physics Organizations
Vallières, M.	Université de Sherbrooke, Canada
van der Merwe, D.	University of the Witwatersrand, South Africa
van Timmeren, J.E.	European Society for Radiotherapy and Oncology
Verellen, D.	Medical Physics Department, Iridium Netwerk – ZAS, University of Antwerp, Belgium
Zwanenburg, A.	National Center for Tumor Diseases (NCT), NCT/UCC Dresden, Germany

Consultancy Meetings

Vienna, Austria: 5–9 December 2022, 4–8 December 2023, 29 April–3 May 2024

IAEA HUMAN HEALTH SERIES PUBLICATIONS

The mandate of the IAEA human health programme originates from Article II of its Statute, which states that the "Agency shall seek to accelerate and enlarge the contribution of atomic energy to peace, health and prosperity throughout the world". The main objective of the human health programme is to enhance the capabilities of IAEA Member States in addressing issues related to the prevention, diagnosis and treatment of health problems through the development and application of nuclear techniques, within a framework of quality assurance.

Publications in the IAEA Human Health Series provide information in the areas of: radiation medicine, including diagnostic radiology, diagnostic and therapeutic nuclear medicine, and radiation therapy; dosimetry and medical radiation physics; and stable isotope techniques and other nuclear applications in nutrition. The publications have a broad readership and are aimed at medical practitioners, researchers and other professionals. International experts assist the IAEA Secretariat in drafting and reviewing these publications. Some of the publications in this series may also be endorsed or co-sponsored by international organizations and professional societies active in the relevant fields.

There are two categories of publications in this series:

IAEA HUMAN HEALTH SERIES

Publications in this category present analyses or provide information of an advisory nature, for example guidelines, codes and standards of practice, and quality assurance manuals. Monographs and high level educational material, such as graduate texts, are also published in this series.

IAEA HUMAN HEALTH REPORTS

Human Health Reports complement information published in the IAEA Human Health Series in areas of radiation medicine, dosimetry and medical radiation physics, and nutrition. These publications include reports of technical meetings, the results of IAEA coordinated research projects, interim reports on IAEA projects, and educational material compiled for IAEA training courses dealing with human health related subjects. In some cases, these reports may provide supporting material relating to publications issued in the IAEA Human Health Series.

All of these publications can be downloaded cost free from the IAEA web site:

http://www.iaea.org/Publications/index.html

Further information is available from:

Marketing and Sales Unit
International Atomic Energy Agency
Vienna International Centre
PO Box 100
1400 Vienna, Austria

Readers are invited to provide their impressions on these publications. Information may be provided via the IAEA web site, by mail at the address given above, or by email to:

Official.Mail@iaea.org.

Feedback on IAEA publications may be given via the on-line form available at:
www.iaea.org/publications/feedback
The form may also be used to report safety issues or submit environmental queries concerning IAEA publications.

Alternatively, contact IAEA Publishing directly:

Publishing Section, Dissemination Unit
International Atomic Energy Agency
Vienna International Centre, PO Box 100, 1400 Vienna, Austria
Telephone: +43 1 2600 22529 or 22530
Email: sales.publications@iaea.org
www.iaea.org/publications

ORDERING LOCALLY

To purchase priced IAEA publications, please contact either your preferred local supplier or the IAEA's lead distributor:

Mare Nostrum Group
39 East Parade
Harrogate
North Yorkshire
HG1 5LQ
United Kingdom
Email: enquiries@mare-nostrum.co.uk
www.mngbookshop.co.uk

Trade Orders and Enquiries:
Telephone: +44 1243 843 291
Email: trade@wiley.com

Individual Orders and Enquiries:
Email: mng.csd@wiley.com

Priced and unpriced IAEA publications may also be ordered directly from the IAEA by contacting IAEA Publishing. The recipient is responsible for shipping costs and any customs and duties.

Artificial intelligence (AI) has significant potential to impact processes in science and technology, including in the area of human health. While bringing potential benefits to healthcare, the application of AI systems also introduces new challenges and potential risks. This publication is aimed at clinically qualified medical physicists, who are health professionals uniquely positioned to bridge the gap between complex AI systems and practical clinical applications. It provides comprehensive guidance for the clinical implementation of imaging based AI systems in medical imaging and radiotherapy, addressing the entire process, from the initial assessment of needs through selection, commissioning, ongoing (quality) management and eventual decommissioning. Although the primary focus is on imaging based AI systems, the guidance provided in this publication is broadly applicable to non-imaging based AI systems as well.

International Atomic Energy Agency
Vienna, 2026
www.iaea.org

ISBN 978-92-0-130026-3