Radiotherapy and Oncology 197 (2024) 110345

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original Article

A joint ESTRO and AAPM guideline for development, clinical validation and reporting of artificial intelligence models in radiation therapy

Coen Hurkmans [a,b,*], Jean-Emmanuel Bibault [c], Kristy K. Brock [d], Wouter van Elmpt [e], Mary Feng [f], Clifton David Fuller [g], Barbara A. Jereczek-Fossa [h,i], Stine Korreman [j,k], Guillaume Landry [l,m,n], Frederic Madesta [o,p,q], Chuck Mayo [r], Alan McWilliam [s], Filipe Moura [t], Ludvig P. Muren [j,k], Issam El Naqa [u], Jan Seuntjens [v], Vincenzo Valentini [w,x], Michael Velec [y]

[a] Department of Radiation Oncology, Catharina Hospital, Eindhoven, the Netherlands
[b] Department of Electrical Engineering, Technical University Eindhoven, Eindhoven, the Netherlands
[c] Department of Radiation Oncology, Georges Pompidou European Hospital, Paris, France
[d] Departments of Imaging Physics and Radiation Physics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[e] Department of Radiation Oncology (MAASTRO), GROW – School for Oncology and Reproduction, Maastricht University Medical Centre+, Maastricht, the Netherlands
[f] University of California San Francisco, San Francisco, CA, USA
[g] Department of Radiation Oncology, The University of Texas MD Anderson Cancer, Houston, TX
[h] Dept. of Oncology and Hemato-oncology, University of Milan, Milan, Italy
[i] Dept. of Radiation Oncology, IEO European Institute of Oncology IRCCS, Milan, Italy
[j] Department of Clinical Medicine, Aarhus University, Aarhus, Denmark
[k] Danish Center for Particle Therapy, Aarhus University Hospital, Aarhus, Denmark
[l] Department of Radiation Oncology, LMU University Hospital, LMU Munich, Munich, Germany
[m] German Cancer Consortium (DKTK), Partner Site Munich, a Partnership between DKFZ and LMU University Hospital Munich, Germany
[n] Bavarian Cancer Research Center (BZKF), Partner Site Munich, Munich, Germany
[o] Department of Computational Neuroscience, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
[p] Institute for Applied Medical Informatics, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
[q] Center for Biomedical Artificial Intelligence (bAIome), University Medical Center Hamburg-Eppendorf, Hamburg, Germany
[r] Institute for Healthcare Policy and Innovation, University of Michigan, USA
[s] Division of Cancer Sciences, The University of Manchester, Manchester, UK
[t] CrossI&D Lisbon Research Center, Portuguese Red Cross Higher Health School Lisbon, Portugal
[u] Department of Machine Learning, Moffitt Cancer Center, Tampa, FL 33612, USA
[v] Princess Margaret Cancer Centre, Radiation Medicine Program, University Health Network & Departments of Radiation Oncology and Medical Biophysics, University of Toronto, Toronto, Canada
[w] Department of Diagnostic Imaging, Oncological Radiotherapy and Hematology, Fondazione Policlinico Universitario "Agostino Gemelli" IRCCS, Rome, Italy
[x] Università Cattolica del Sacro Cuore, Rome, Italy
[y] Radiation Medicine Program, Princess Margaret Cancer Centre and Department of Radiation Oncology, University of Toronto, Toronto, Canada

ARTICLE INFO

Keywords:
Artificial Intelligence
Deep learning
Machine Learning

ABSTRACT

Background and purpose: Artificial Intelligence (AI) models in radiation therapy are being developed with increasing pace. Despite this, the radiation therapy community has not widely adopted these models in clinical practice. A cohesive guideline on how to develop, report and clinically validate AI algorithms might help bridge this gap.

Abbreviations: AAPM, American Association of Physics in Medicine; AI, Artificial Intelligence; ASTRO, American Society for Therapeutic Radiation Oncology; AUC, Area Under the Curve; CLAIM, checklist for AI in medical imaging; CLAMP, checklist for AI/ML applications in Medical Physics; CNN, Convolutional Neural Network; CT, Computer Tomography; DL, Deep Learning; DVH, Dose Volume Histogram; EPID, Electronic Portal Imaging Device; ESTRO, European Society for Therapeutic Radiation Oncology; GAN, Generative Adversarial Network; IMRT, Intensity Modulated Radiotherapy; OAR, Organ At Risk; ML, Machine Learning; MLC, MultiLeaf Collimator; MR, Magnetic Resonance; MRF, Magnetic Resonance Fingerprinting; MRI, Magnetic Resonance Imaging; NTCP, Normal Tissue Complication Probability; PROBAST, Prediction model Risk Of Bias ASsessment Tool; psQA, patient-specific QA; QA, Quality Assurance; RO, Radiation Oncologist; RT, Radiation Therapy; RTT, Radiation Therapist; TCP, Tumour Control Probability; TNM, Tumour, Node, Metastases; TRIPOD, Transparent Reporting of a multivariable prediction model of Individual Prognosis Or Diagnosis; UICC, Union Internationale Contre le Cancer; VMAT, volumetric modulated arc technique.

* Corresponding author at: Department of Radiation Oncology, Catharina Hospital, Michelangelolaan 2, PO Box 1350, 5602 ZA Eindhoven, The Netherlands.
E-mail address: Coen.hurkmans@cze.nl (C. Hurkmans).

https://doi.org/10.1016/j.radonc.2024.110345
Received 23 May 2024; Accepted 23 May 2024
Available online 3 June 2024

Guideline
Radiation Therapy
Segmentation
Treatment planning
Ethics
Quality Assurance

Methods and materials: A Delphi process with all co-authors was followed to determine which topics should be addressed in this comprehensive guideline. Separate sections of the guideline, including Statements, were written by subgroups of the authors and discussed with the whole group at several meetings. Statements were formulated and scored as highly recommended or recommended.
Results: The following topics were found most relevant: Decision making, image analysis, volume segmentation, treatment planning, patient specific quality assurance of treatment delivery, adaptive treatment, outcome prediction, training, validation and testing of AI model parameters, model availability for others to verify, model quality assurance/updates and upgrades, ethics. Key references were given together with an outlook on current hurdles and possibilities to overcome these. 19 Statements were formulated.
Conclusion: A cohesive guideline has been written which addresses main topics regarding AI in radiation therapy. It will help to guide development, as well as transparent and consistent reporting and validation of new AI tools and facilitate adoption.

Due to the increase in computational power and the emergence of big data, the field of artificial intelligence (AI) is rapidly developing. AI applications are already widely used, e.g., in natural language processing, search engines and facial recognition software. AI has also entered the field of health care, and guidelines in this wider field are emerging [8,17,19,20,26,40,58,68,74,88–90]. Radiation therapy (RT), having a strongly data-driven workflow, is an active field of AI model development. AI models have for example already been developed in the field of image reconstruction, volumetric segmentation, treatment planning and delivery, outcome prediction and quality assurance (QA). Despite the many efforts in this area, the RT community has not widely adopted these AI models in clinical practice and their availability, applicability, quality, generalizability, interpretability and safety are still a matter of concern. Based on our own experience we see this is partially due to the lack of a cohesive guideline on how to develop, report and clinically validate AI algorithms within the radiation therapy domain.

This guideline provides an overview of the latest publications, focusing on already existing reviews per sub domain of AI in radiotherapy. Based on the literature, knowledge and experience of the authors, suggestions were given on how to develop and clinically validate new AI models for use in the radiation therapy domain and how to report their results scientifically and consistently. Consistent reporting enables fair comparisons of different models and benchmarking against existing approaches.

Materials & methods

This guideline aims to stimulate the development, validation and safe clinical implementation of AI in radiotherapy.

There is no global consensus on the precise definition of AI. Machine Learning (ML) models and a sub-category of ML models called Deep Learning (DL) models, are the most widely known models and are the models focussed on in this report. Other models like large language models are not yet used in the radiation therapy domain and therefore not addressed here but have great potential [41,59].

Representing the international contribution and impact of this effort, the writing committee was formed with medical physicists, radiation oncologists, radiation therapists and research members from ESTRO, ASTRO and AAPM who all have ample experience within the field of AI for radiotherapy, either in research and/or a clinical setting, including software development in collaboration with industry. To determine the topics and subtopics for inclusion and the guideline structure, a Delphi process was followed [43]. Questionnaires were anonymously collected and the results were discussed as a group. Topics were included if at least 70 % of members voted for it. For each topic, a subgroup formulated up to 3 Statements that were categorized as highly recommended or recommended, which were discussed at subsequent committee meetings. To give the readers of this guideline a good overview and limit the length of this guideline we have referenced some key reviews on specific topics without discussion the underlying original work in detail. The readers are encouraged to read those original papers if they want to know more details on the topic.

Results

The Delphi process took place between Q4 2021 and Q2 2022. After 3 rounds, consensus was reached on the main structure of the report as it appears now. The questionnaires can be found in Appendix A. After 2 rounds, most of the topics and subtopics were set and there were only a few details for which there were preferences, but agreement did not reach 70 %. For these details it was decided to continue with the preference topics.

The following topics were found most relevant (Fig. 1): Decision making, image analysis, Volume segmentation, treatment planning, patient specific quality assurance of treatment delivery, adaptive treatment, outcome prediction, training, validation and testing of AI parameters, model availability for others to verify, model QA/updates and upgrades, ethics.

Decision making

The decision-making process is a reasoning process based on assumptions of values, preferences and beliefs of the decision-maker. Decisions are shared between the physician and patient, and decision modelling within healthcare could incorporate expert domain knowledge, probabilistic reasoning and patient preferences. AI can guide decision-making in radiation oncology at different steps: Patient initial evaluation, clinical treatment strategy, dose prescription and toxicity prediction and management.

Patient initial evaluation involves a consultation that takes into account the patient's symptoms, medical history, examination, pathological, genomic and imaging data [18]. These will guide the radiation oncologist's clinical treatment strategy. This strategy is typically guided by national and international guidelines combined with an intuitive prediction of the benefit and potential toxicities of a treatment, but AI could have a relevant role in that setting. Dose prescription is determined by internationally and/or nationally accepted standards whether they have been defined in clinical trials or not. Variations in tumour biology and radiosensitivity are not often considered. AI might enable the personalization of radiotherapy dose prescription by predicting the radiation sensitivity and toxicity [4,80]. Many models have been developed and published to predict treatment response or toxicity, but almost none are used in the daily routine. One reason for this is that they have not been validated. An example of a model that is used in daily routine is the Dutch Normal Tissue Complication Probability (NTCP) model to select patients for proton therapy [53]. Because these models should ultimately be used to define clinical treatment plans, from which the delivered dose will depend on, they need to be rigorously developed and externally validated.

Statement 1: Decision-making should strictly rely on models created in accordance with published guidelines on development & in silico validation [40,74,88] (highly recommended).

Statement 2: Models used for decision-making should be validated through careful monitoring of patient's conditions in terms of toxicity and tumour response through the treatment [20] (highly recommended) and in prospective clinical trials [1858,65] when possible (recommended).

Image analysis

Image analysis include a variety of processes which are amenable to applications of DL. While this area of research and implementation is expanding rapidly, existing applications may be generally grouped into the acquisition/reconstruction, generation of synthetic images and image-registration.

Acquisition and reconstruction methods have been widely investigated in the diagnostic literature, resulting in accelerated acquisition times, reduced artifacts, and/or image quality which is equivalent to standard procedures. For Computer Tomography (CT), an overview of DL techniques for image denoising, metal artifact reduction and super-resolution imaging is presented by Li et al. [55]. It is imperative that QA is performed when using DL augmented reconstruction to ensure that CT Hounsfield unit based dose calculation is not altered. DL-based approaches for cone beam CT have generated higher quality images and reduce motion artefacts, allowing for accurate dose deposition assessment, with quality comparable to standard CT [78].

DL reconstruction methods for Magnetic Resonance Imaging (MRI) have substantially improved reconstruction capacity [11]. A potentially transformative application of DL is reconstructing multiple sequences from a single acquisition, using a technique referred to as "MR fingerprinting (MRF)". MRF leverages pseudorandom acquisition parameters to match signal trajectories to denote an identifiable unique pattern-match (or "fingerprint") on a voxelized basis, affording simultaneous generation of multiple MRI signals (such as T1, T2, diffusion, etc.) for each tissue voxel from a single scan. The use of DL models for pattern-matching has allowed the speed and robustness of these MRF methods to be substantially enhanced and raises the potential for reproducible DL-based multiparametric MRF for radiotherapy [16,38,81]. The general field of DL for MR-guided radiotherapy is still young. With DL models for various parts of the workflow for an MR-linac but also for x-ray or ultrasound based systems, real-time motion management might become a reality [60,69]. These techniques should be at a minimum benchmarked to standard acquisitions for QA before implementation in the clinic.

With regard to Positron Emission Tomography-CT (PET-CT) and also PET-MR, deep learning approaches have improved image quality and use spatial information to account for time dependent motion variance [95]. Consequently, these improvements have to be taken into account when delineating targets or with therapy response assessment, as most published data is still based on non-motion correction/Time-of-Flight techniques.

Comparatively extensive radiotherapy specific research has been undertaken on the generation of comparable images from alternative modalities. "Pseudo-" image generation involves training a deep learning platform with comparable acquisitions using another format or modality (e.g., matched MR and CT images) and constructing a virtualized version of one of the imaging pairs. Generative adversarial networks (GANs) have accelerated this process. It is advantageous as a method of forgoing CT-simulation by allowing acquired pre-treatment MR images or image guided RT images (e.g., MR-linac or cone-beam CT platforms) to be used directly for dose calculation. The vast majority of these efforts have used data from a single or limited number of centres, with only a few reports with external validation [77]. GAN-based methods are potentially susceptible to dependencies on training data, and therefore should necessarily undergo assessments for generalizability before implementation outside of the original use case. For example, a pseudo-CT application using a particular slice thickness reconstruction standardized at one facility, may show significant

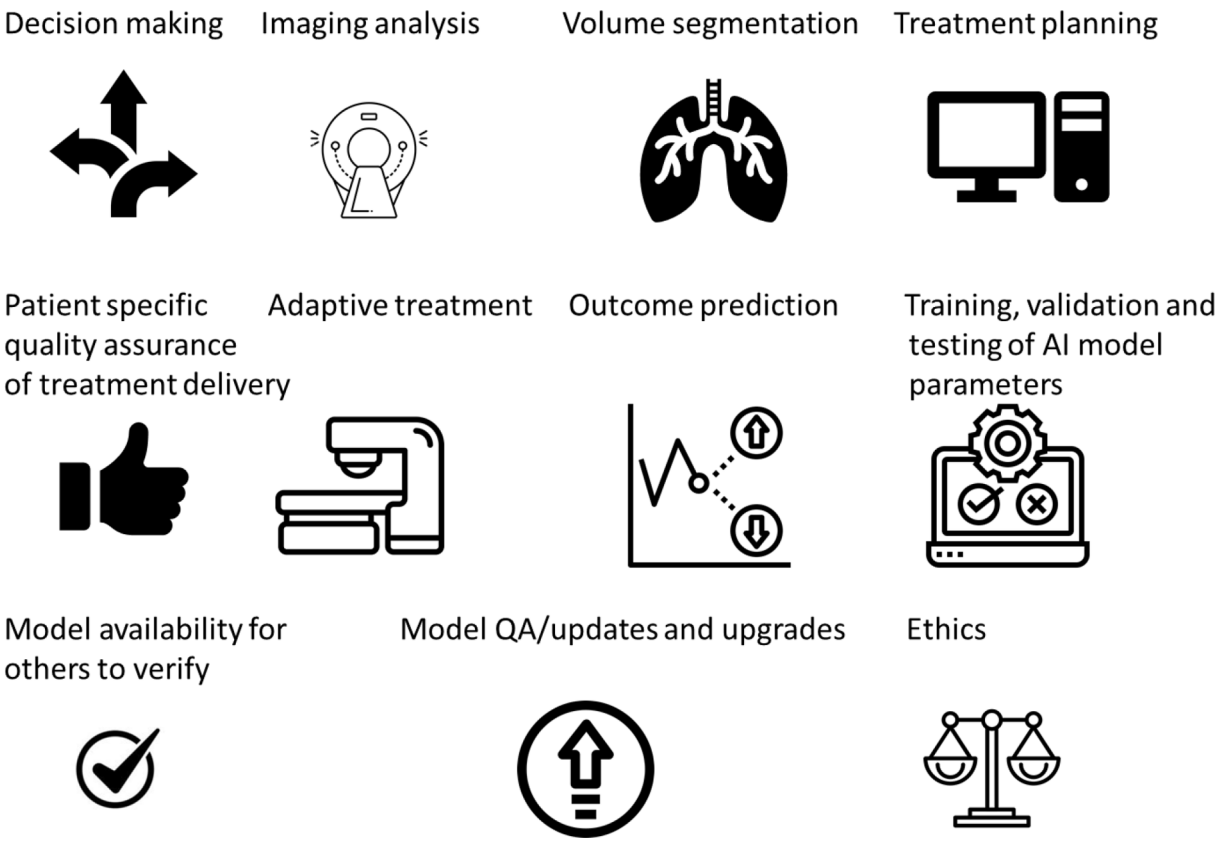

Fig. 1. Main radiation therapy topics considered in this guideline.

performance differences when applied to CTs with a different institute's acquisition protocol. Jabbarpour et al. reported a multicentric application of pseudo-CT generation from MR images [47], and represents an example of generalizability and external validation which should be encouraged for applications generally.

As there is no ground truth in pseudo image generation, QA should at least take place at the use case level. It should thus be directly clear that the image is a pseudo image. At a minimum, when pseudo images are utilised, there should be a formal reporting of the application/code including annotation of the version (or code deposition, if open source or development-level), notation of the characteristics of the input/ and output of the process (e.g., "pseudo-MR T2 fat suppressed images were generated from contrast CT simulation"), description of whether there was formal QA through reference of pseudo- images to a baseline image at the individual patient/case-level (for example, initial use of a simulation CT, with subsequence pseudo-CT generation for adaptive dose calculation), and notation of whether reference images or pseudo-images alone were utilized for specific process (e.g., dose calculation or target delineation).

Statement 3: When DL reconstruction algorithms are utilized, there should be a formal designation of the method (at the level of the vendor-supplied software version – a model fact sheet), annotation of the basic data, and supplementary data regarding resultant reconstruction parameters (recommendation).

Volume segmentation

The use of AI for volume segmentation is arguably the most developed application for radiation therapy. Several commercial products are now available, as well as open-source code developed in the research setting and numerous publications reporting on independent cohorts of patients as well as publicly available curated datasets. The architectures used to train AI segmentation models have been rapidly developing, benefiting from developments in AI in computer vision outside of the medical field and advances in graphical processor unit power. These combined areas have enabled the progression from 2D models, to 3D patch based models, to fully 3D models that automatically configure themselves (e.g. no new Unet), to techniques that learn from multi-modality images.

One of the most significant challenges with the development and clinical deployment of AI for volume segmentation is its validation, specifically in determining precisely what the "gold standard" or ground truth segmentation is. Numerous studies have demonstrated the variability in contouring both normal tissues (higher level of agreement), gross tumour volumes (variable agreement), and clinical target volumes intended to encompass microscopic disease (large variability). Variability exists between and within observers as well as clinical practices. Consensus methods (e.g. Simultaneous truth and performance level estimation) and guidelines can improve the consistency of these contours [22,73].

Once a suitable ground truth is established, several methods exist to validate the segmentation accuracy including qualitative acceptability by the clinical user, workflow efficiencies, geometric accuracy versus the ground truth, dosimetric impact when using the segmentations to develop treatment plans, and finally outcomes assessment, e.g. differences in the actions taken or determined outcomes based on the differences between the ground truth and the AI segmentation. Published studies vary widely in the mechanisms used to report accuracy in AI segmentation making comparisons between studies challenging [7,42,51]. Although the chosen metrics depend on the research question to be answered, it would be good if at least qualitative (e.g. Likert scale for usability) and quantitative (e.g. Hausdorff Distance and (surface)-Dice Similarity Coefficient, time gain) metrics would always be reported. The inter- and intra-observer variability in contouring tumours and normal tissues in the clinical setting should also be considered when evaluating AI segmentations [5]. The field should encourage the development of these clinical benchmarks for comparison.

At this stage in the clinical deployment of AI for volume segmentation, simple geometric validations, are insufficient to predict usefulness in a prospective clinical setting. For example, as these tools are mainly used for more efficient segmentation, a timing study should be performed. The translation of geometric differences in tumour and normal tissue segmentation into clinical factors, e.g. meeting dosimetric criteria vary depending on their spatial location and their interaction with the treatment planning and delivery processes. For example, Chen et al. showed that dose distributions of target volumes were unaffected when auto-segmented organ contours were used in the design of treatment plans compared to manual delineations, whereas the impact of automated segmentation on the doses to OARs was complicated [12]. It is possible that the simple reliance of segmentation accuracy on geometric measures will both limit the clinical deployment of AI segmentation in areas where the current accuracy is sufficient and potentially encourage the clinical deployment of AI segmentation in areas where the accuracy is not sufficient (e.g. due to highly conformal treatments, small numbers or fraction, and aggressive treatment intent). It is recommended that research leaders and professional societies invest in the development and public dissemination of or blinded access to benchmark datasets to enable the systematic testing and validation of new algorithms on standardized data consistent across studies. Researchers must be as comprehensive as reasonably achievable in reporting the limitations that exist in their studies performed using either these benchmarking tools or non-public data that cannot be overcome at publication.

Treatment planning

The field of AI treatment planning is developing very fast. There have been some reviews in the past 3 years but it is likely new insights will continue to appear (e.g. [70]). Primarily, it is important to distinguish the various types of output from such models. Some models predict a Dose Volume Histogram (DVH), while others predict one 3D dose distribution or a pareto optimal range of dose distributions. Eriksson and Zhang presented a model for robust 3D dose prediction for proton beams [32]. There are also models which can predict fluence maps [94] or segment shapes [72] using 2D projections of the segmented organs in the beams-eye view directions. For all of these outputs, it is still necessary to translate the predicted output to a deliverable plan with all the required machine settings. This actual plan should fall into the same treatment plan class solution as the plans used as input into the AI model in order for the solution to be directly useful clinically. For example, an AI model trained on Intensity Modulated Radiotherapy (IMRT) plans will in general not be able to predict a dose distribution for a Volumetric Modulated Arc Technique (VMAT) technique with sufficient accuracy and passing QA measurements [56].

For all of the above methods, it is intrinsically assumed that the class solution is known; beam energy, number and direction of beams or arc start and stop angles need to be selected beforehand and are consistent between training, validation and clinical use. Automated non AI methods to determine optimal beam angles are available, for example the 4pi algorithm [24] while AI based orientation selection is in the early stage of development [79].

Also, it is intrinsically assumed that the training data represents the clinical scenario in which the model will be applied. It should be clear whether the treatment plans used for training were actually used in the clinic or not or if some plans were further improved or removed from the training set and based on which criteria. This should be combined with a clear description of the targets and OARs on which the plans were made, preferably with a reference to published delineation guidelines and standard nomenclature [66].

When evaluating the models, the metrics used should be based on the output within the treatment planning process for which the AI model is developed, but preferably also on the actual deliverable plan, i.e., taking possible post processing and final plan calculation into account, as these

can differ [85]. For example, some papers report on the predicted dose before translation to a deliverable plan and some on the dose after translation, i.e., an actual plan. A general review of plan quality metrics is given by Hernandez [39], who distinguish dose distribution, plan robustness and plan complexity metrics and give examples of such metrics. There is no consensus yet on generic robustness metrics. Moreover, actual clinical employment of a model may lead to plans perceived less quality than in a retrospective analysis [67]. This may depend on the input data (e.g. use of clinical data or data generated specifically for model development) or on the subjective qualitative scoring by the observers [52]. It may be helpful if an AI model could also quantify the uncertainty in the dose prediction. For example, if the prediction is used as a benchmark QA tool to compare with clinical plans. Monte Carlo bootstrapping and bootstrap aggregation are proposed methods to incorporate this [71].

Statement 4: Plan quality metrics should encompass dose and also robustness and complexity metrics combined with acceptance criteria (highly recommended).

Statement 5: Qualitative scoring before and also during clinical employment for the first patients should be performed (recommended).

Patient specific QA of treatment delivery

The number of publications on patient-specific QA (psQA) is growing significantly. Osman and Maalej published a comprehensive review of 20 relevant papers that appeared until March 2021 [75]. They summarized studies based on algorithm used, anatomical site, number of input plans or beams, number of input features, QA outcome prediction metric and key results.

As input such models usually use a number of plan features and not the whole plan, and it was argued that aperture-based complexity metrics are more direct descriptors than fluence-based complexity metrics as they directly represent the delivery parameters utilized by the treatment machine. As such they may offer better insight into the disagreement between the calculated and measured doses.

Gamma pass rates or machine error detection were used as outcome prediction in the publications. The gamma pass rates were in general based on 2D or 3D measurement of the complete plan or of each beam. The machine errors mainly consisted of Multileaf collimator (MLC) positioning errors and output variations, but some studies also included errors in MLC transmission, dosimetric leaf gap, effective source size and alignment of a measuring device [50]. Generally, studies tried to predict measurements without the patient on the table. To incorporate patient specific influences on QA measurements and predictions, Wolfs et al. for example used Electronic Portal Imaging Device (EPID) measurement of actual patient treatments as input to learn to detect and classify anatomical changes and tumour regression [91].

Patient specific QA results depended on machine and MLC type, measurement detector, and whether or not multi-institutional validations were performed. Pass rates also depend on the algorithm used to calculate the gamma values [46]. Many studies used EPID based input and recently other detectors have also been used [63,92]. Predicted gamma pass rates deviated on average less than 1 % (with 3 %=2 SD) for 3 %/3mm thresholds.

Regarding the usability of AI models for error detection, Wootton et al. could, for example, demonstrated higher accuracy compared to the conventional gamma analysis with a logistic regression model [92]. Many other models were compared and support vector machine models came out best in several studies, using metrics like area under the curve (AUC) and root mean square error. However, various metrics and also simulation and classifications of errors were used, making it very hard to draw clear conclusions over multiple studies.

Moreover, it is not always directly clear if the metric would be clinically useful. For example, besides the prediction accuracy one would probably also want to know the specificity and sensitivity (AUC) of a model for certain clinical action levels, as an important concern is not to predict too many plans as passing while in reality these would not pass the actual measurement [44].

An interesting approach to psQA first described by Carlson et al. [10] used machine logfiles as training input to accurately predict the actual MLC positions for VMAT plans. Using these positions and recalculating the treatment plan they were able to better predict the actual achieved dose distribution and DVHs which led to higher gamma pass rates compared to measurements. Chuang et al., Osman et al. and Huang et al. expanded on this work [14,45,75].

Gong et al. developed an AI model to directly predict actual DVHs with AI input from phantoms without the need for logfiles [37]. Predicting DVHs is a promising approach which may be more clinically relevant than gamma pass rates which have been shown not to correlate well with DVHs.

To our knowledge, AI psQA models have not been implemented clinically yet. This might be partly due to the unavailability of large, well curated multi-institutional training datasets to improve and generalise the predictions and the unavailability of commercial psQA outcome prediction solutions. It would also be useful to know the uncertainty of the prediction of a model. Yang et al are among the first to investigate this [93].

Statement 6: Before models can be safely implemented clinically, the models should be validated for the combination of treatment planning system, treatment technique, patient group and radiation and measurement equipment for which it will be applied (highly recommended), preferably on large scale multi-institutional data (optional) and in the institution that will actually use it (highly recommended).

Statement 7: Patient-specific QA model results should include clinically relevant metrics like sensitivity and specificity (Receiver operating characteristic analysis) (highly recommended).

Adaptive treatment

Most models predict a treatment plan based on the image set acquired and one or more segmentations. However, there are also models that combine prediction of anatomical changes and related adapted treatment plans [15].

The adaptive RT setting may also allow the implementation of training strategies exploiting prior knowledge, for example images and segmentations from the original treatment planning. This idea is to adjust a model in a patient-specific fashion, and is well described in Eppenhof et al. [31]. Several publications have shown that in the case of segmentation, some benefits over a population model can be obtained [13,35,49,57]. In this case, care must be taken to balance tuning the model to the planning image and applicability to subsequent fraction images. Alternatively, segmentation deformation networks can also be used to propagate planning segmentations to fraction images [49].

The use of AI for both offline and online adaptive RT is the culmination of AI for volume segmentation, image analysis, planning, QA, and outcomes prediction. In an ideal setting, adaptive RT is driven by the analysis of the images and a resulting deviation in the predicted outcomes – or the opportunity to improve the outcomes based on changes in the patient anatomical or functional state. Currently, most adaptive strategies are driven by either planned intervention time points or geometric changes in the patients. Data has demonstrated that these may result in adaptations that do not benefit the clinical outcomes for the patient and miss the opportunity to improve clinical outcomes for others. Leveraging these tools to optimize the use of resources, when beneficial, will benefit radiotherapy patients. We must work to simultaneously effectively optimize the risk–benefit of the use of AI tools for adaptive radiotherapy for an individual patient as well as assessing the risk–benefit of the tools on the patient population. Specifically, research results typically report the average, standard deviation, and interquartile range of the accuracy of the adaptive tools. These results are highly useful to determine if clinical translation is warranted, however, when outliers in the data exist, we must evaluate how to determine the

patients that will have similar poor results once the tools are clinically translated. Patient specific QA methods play a critical role in preventing these outlier errors from impacting patient care. Outlier cases might benefit most from model adaptation, while the average model performance might decrease from this model adaptation. The intensity of the clinical decision making is increased when moving from the offline adaptive setting, where hours of clinical time can be dedicated to QA, to the online adaptive setting, where QA and decisions must be made in minutes.

Alternatively, as these tools progress together and the resources required to employ these tools significantly decreases, one can envision a new standard of care where every patients get an optimized treatment of the day based on daily images that have been accurately contoured, with an optimized best-in-class treatment plan with sufficient QA that has leveraged outcomes prediction to determine the optimal fractionation that should be delivered to that days anatomical and functional presentation. This is a futuristic goal, however, early results demonstrate that this level of efficient, daily adaptation can be possible for all patients. Several innovative tools are already demonstrating great promise for clinical translation such as AI-based image augmentation to improve the image quality and contrast resolution of non-diagnostic images, which may improve the ability to perform online adaptive RT [60].

Outcome prediction

With the growing importance of personalised radiotherapy, AI-based outcome prediction using the vast amount of patient-specific data have gained substantial attention. The Outcomes Working Group of the American Society of Clinical Oncology defines the outcomes of cancer treatment as a tool to be used for technical assessment and the development of cancer treatment guidelines [1]. According to the Outcomes Working Group, two types of outcomes are defined; i.e., patient outcomes (e.g., survival rate or quality of life) and cancer outcomes (e.g., toxicity, response, cost-effectiveness). When choosing a treatment plan, a single outcome could not be demonstrative of the overall patient outcome after treatment. Therefore, it is more relevant to consider three important outcomes, typically toxicity, response, and survival rate.

Outcome models can be either predictive or prognostic and both were considered when referring to prediction models in this paper. Outcome models in radiation therapy can be categorised in analytical models and data-based models. Analytical models are mainly based on dosimetric variables (e.g., dose distribution, fractionation) and can be either mechanistic (e.g., the linear quadratic model for TCP), or phenomenological in nature (e.g., Lyman model for NTCP). However, treatment response is mediated by complex interactions among patient-specific anatomic, biological and treatment conditions not captured in analytical models. Data-based models do have the ability to include multimodality imaging, high throughput biotechnology providing omics information, and capture clinical data through electronic health records. These data elements can supplement conventional radiation dose-based treatment parameters and clinical information providing a pathway to more accurate outcome predictions for an individual patient.

Minimum reporting requirements for reliable prediction models using AI [29,68] have been formulated. Five prediction model development phases can be considered: (1) clinical problem definition; (2) data preparation; (3) prediction model development; (4) prediction model validation and testing; (5) prediction model interpretation, impact assessment and implementation into daily healthcare practice [3]. The most important aspects to be considered for each phase will be summarised here.

Clinical problem definition

Defining the clinical problem consists of specifying the patient-specific parameters and endpoints. Patient-specific parameters can be clinical data, treatment data, dosimetric data, imaging data, biological data. Endpoints include TCP-related endpoints as well as normal tissue response-related endpoints and overall survival.

Data preparation (pre-processing, profiling and curation)

The data preparation phase consists of selecting the patients to construct an unbiased dataset. Considerations in this are treatment regime, follow-up time, bias with respect to underrepresented groups. Inclusion and exclusion criteria need to be specified to understand how the cohort was assembled and which patients were excluded.

Statistical profiling is used to detect inconsistent data, suspicious outliers and to recognize trends in the data set. Imputation methods can be used to correct for missing data. Data normalisation (standardisation) of highly variable data (e.g., Z-score) should be performed. Protocols should be in place that define the link between the curated data and the correct patient. The use of minimal common data elements such as the Operational Ontology for Oncology O3 [64] will improve the quality of real-world patient data. Review (auditing) of the outcome dataset should be performed by an interdisciplinary team with sufficient clinical knowledge. This could be a radiation oncology team within the hospital where the patient is treated or a trial organisation if the patients were treated within a clinical trial.

Prediction Model Development

The appropriate model architecture should be decided considering the input data types involved. A comprehensive model includes categorical data, imaging data and clinical data. The dataset should be split into training, tuning and test subsets using a cross-validation approach. Training on a balanced dataset is an important consideration as the power of the dataset is determined by the size of the minority class. To achieve data size balance, prior to training, oversampling methods are often used to increase the minority class, or selection methods are used to reduce the majority class. During the tuning process, hyperparameters and parameters used in the model can be optimised.

Prediction model validation and testing

Typically, validation and testing is performed with retrospective single-institute datasets. Prospective clinical deployment of AI-based outcome models will provide more direct evidence about model performance and is needed to pave the way to trustworthy clinical implementation on real-world data. Preferably, a study analysis plan is registered prior to model development to increase transparency and reduce bias [83].

Prediction model impact assessment and implementation

The purpose of predictive models is to support clinical decision-making and inform treatment planning. Examples of predictive models used ubiquitously in the oncology community, recommended and updated by internationally recognised regulatory agencies are the Tumour, Node, Metastasis (TNM) and the Union Internationale Contre le Cancer (UICC) models. The TNM is a survival predictor, which uses an ontology that is widely understood and is used both in clinical decision-making for individual patients, and in the definition of guidelines dictating behaviours considered beneficial for the patients. The ability of AI-based predictive models to use vastly more data and more variable data types than those encoded by the TNM system, while potentially making prediction more personalised entails greater complexity for regulatory agencies and scientific societies in defining how to formulate recommendations and update them over time.

Statement 8: an auditing step of the patient outcome data by the interdisciplinary team responsible for the care of the patient is an important data quality safeguarding step. (highly recommended).

Statement 9: the collection and interprofessional standardisation of

structured common minimal data sets for different indications is an important step towards the building of interpretable prediction models based on real-world patient data. (this cannot be directly influenced by individuals: optional).

Training, validation and testing of AI parameters

It is recognized that the development of AI algorithms needs to be performed rigorously in order to ensure reproducibility, reliability as well as generalizability to unseen data (i.e., learning the task)[6,26]. This process has been discussed in many textbooks on AI [87] including their application in the field of medical physics [27]. AI as data-driven approaches learn their (hyper-) parameters from the underlying data dependencies (e.g., input–output mapping in supervised learning or unlabelled data patterns in unsupervised learning). Specifically, in the case of supervised machine/deep learning, which constitutes the majority of AI applications in medical physics, radiology or radiation oncology (e.g., auto-contouring, dose predictions, treatment planning, online adaptive, etc.), the data is typically partitioned into three main nonoverlapping subsets with clear description of selection criteria and pre-processing procedures [84]: training subset of the AI algorithm, tuning (often called validation) subset for tweaking of the hyper-parameters as well as it acts as an internal validation process of the algorithm, and an independent testing subset for external validation purposes to ensure generalizability of the algorithms beyond training [21]. Though there is no consensus on the sizes of these data splits, ratio splits like 60:20:20, 70:15:15, or 80:10:10 are common. The choice of the exact ratio split may be data and problem dependent. The different validation steps are highlighted in the original Transparent Reporting of a multivariable prediction model of Individual Prognosis Or Diagnosis (TRIPOD) guideline [19] and the ensuing proposed extensions to AI applications (TRIPOD-AI) and the risk of bias tool (PROBAST-AI), [17] as part of the Equator network guidelines on enhancing the quality and the transparency of health research reporting. Though these guidelines and other journal publication checklists [29,68] may address requirements for the development of AI algorithms using large retrospective analyses, they don't address the concerns regarding the prospective deployment of these algorithms in the clinic [67] nor the intricacies associated with human–computer interactions [28]. Therefore, there is a need to ensure that an AI model is not only accurate in its performance, but also it is transparent, fair and bias-free through post hoc application of AI interpretability/explainability techniques [36,62]. Another approach that is being advocated by several medical societies is intelligence augmentation by incorporating human expertise in the development cycle of AI. This intelligence augmentation approach seems to improve performance with demonstrative examples in radiology [76] and radiation oncology [61]. Thus, the notion of fully autonomous AI may remain a futuristic goal at least for the time being, particularly in areas that involve human decision making.

Statement 10: Checklists to ensure reproducible AI development in the broader sense than radiation therapy should be used (highly recommended).

Model availability for others to verify

Using externally (commercially or from scientific research) developed models typically have the intrinsic uncertainty of the end-user not being able to know all the specific details in model architecture, software implementation and the procedure followed for training, validation and testing of the AI model. This however does not mean that these models cannot be used in other settings but some care has to be taken in using these models outside of the reference conditions. Availability of the source code could give an insight in the AI architecture and implementation but is not always available. If source code is present, this might not always be sufficiently reproducible if specific (in-house) coding libraries are used, patches or version numbers not always being compatible or available. Architectural description of the model is the minimum requirement, preferably all details should be provided.

The dataset used for training, validation and testing should follow the general guidelines provided in the section on "Training, validation and testing AI parameters" and are still applicable. At least a description of the datasets and key characteristics include pre- and post-processing should be available. Furthermore, applying the AI model in a population outside the models training population, or, more general, it's intended use, should be discouraged. Model datasheets should be provided for every trained model. Example model datasheets are given for segmentation and treatment planning in the appendices and an example of a careful implementation of an externally developed model can be found in [5]. Simple correction strategies applied in statistics such as calibration of the model are not always as straightforward for AI models as for more commonly used statistical methods (e.g. logistic or linear regression). Furthermore, tuning the model parameters using new datasets is the field that is receiving also more attention of "incremental learning". Details for this are provided in the section on "Model QA/updates and upgrades".

Statement 11: Externally developed models can also be used in other institutes after careful implementation (highly recommended).

Statement 12: The end user should verify the model is trained and validated on the intended population to be used (highly recommended).

Model QA/updates and upgrades

Most AI models are trained on data of a specific historical cohort. With the introduction of novel technology (e.g. new CT scanner, updated MR acquisition protocols) or new work procedures in clinical practice (e.g. changes in treatment dose prescription or delivery technique) it is important to have two procedures in place in clinical routine [86].

The first one is a periodic QA of the performance of the AI model, the model performance may drift over time due to the above mentioned changes in workflow, patient population or equipment. The frequency and thoroughness of such a QA model will depend on the specific use cases and clinical implementation.

The second procedure to have in place is an update (typically referred to small model changes) or an upgrade (for more substantial changes) procedure. At some point the need to update or upgrade models to reflect the current state-of-the-art again may arise. Various possibilities are currently being investigated to do this most efficiently. The simplest but most crude way of deriving a new model is to train, validate and test a new model based solely on a newly acquired retrospective cohort that resembles the current way of working. Alternatively, also the previous cohort could be included to make a larger robust dataset, however one has to determine if the previous dataset also falls within the scope of the new applications. Another currently investigated procedure is the cycle sometimes referred to as rapid learning healthcare [3]. In such a cycle periodically (or even on a case-by-case level) new data is added to the training cohort of the model. Since a (slightly) different model is trained every time, validation and test procedures of these models need preferably to be fully automated for such methods to work in clinical routine.

Ethics

Despite the promise and potential of AI/ML application in the medical domain in general and radiation oncology specifically, it has generated a myriad of new ethical and legal challenges that are tampering its progress. These challenges could be concerning the patient in terms of privacy protection, the healthcare provider in terms of legal responsibility when mistakes occur, the developers in terms of transparency and mitigating the risks of bias or error, and subsequently algorithms themselves in terms of trustworthiness and acceptance. Therefore, organizations such as the European Commission have published related guidelines and white papers. These include the white

paper on AI trust and excellence, [2] which was followed by several regulations on data governance and AI harmonized rules through the Artificial Intelligence Act A of April 2021. In parallel, the United States has issued the National AI Initiative Act, which became a law in January of 2021. The Food and Drug Administration has also been occupied in regulating many AI products as they appear in the US market by classifying them as software as a medical device [34]. In addition, the Food and Drug Administration has issued several guidelines for good AI practice in collaboration with equivalent agencies in Canada (Health Canada) and in the United Kingdom (Medicines and Healthcare Products Regulatory Agency) [30]. Also the USA government has issued a guidance on AI, which includes details on healthcare applications [82]. On December 9th of 2023, the European parliament reached a provisional agreement on the new EU AI Act [33]. Important to remark is that in this act it is stated that High-risk AI systems shall be designed and developed in such a way, including with appropriate human–machine interface tools, that they can be effectively overseen by natural persons during the period in which the AI system is in use. The current legislation does as such not change the responsibility structure in a hospital environment [25]. In response to the regulatory guidelines and to address the pressing concerns regarding transparency and reproducibility of AI results, [54] research journals have been also busy issuing their own guidelines for publication purposes such as the checklist for AI in medical imaging (CLAIM) by the Radiology journal sponsored by the Radiological Society of North America and the checklist for AI in medical physics (CLAMP) [29] by the journal Medical Physics sponsored by the AAPM. Though these guidelines and checklists may help address some of the immediate ethical and legal concerns associated with AI development, there remain many open questions regarding deployment and clinical implementations in the short and long terms. These open questions include but are not limited to how to effectively monitor AI performance in the clinic and at what frequency? How to address the known data drift issues in medical records and when to update AI training data accordingly? Also unknown, how AI deployment will impact the complex healthcare provider-patient relationship and what will be the future of this relationship in AI's presence? As we learn more about AI role in the medical field and radiation oncology answers to these questions will likely be provided but more likely newer questions will be posed too.

Statement 13: Though ethical standards may vary by society and regulatory bodies in the European Union and North America, these bodies are setting requirements for AI products that need to be considered. (Highly recommended).

General Statements and discussion

Topic specific guidance and Statements have been given in the previous sections. Hereunder some more generic guidance and Statements are given in the context of the use of AI for radiotherapy.

Proper use of commercial AI tools for radiotherapy starts with setting the right requirements for the purchase of such tools. A good overview of developing, purchasing, implementing and monitoring AI tools in radiology is given in [8] and many aspects obviously overlap with AI tools for radiotherapy. Before purchasing, one has to consider challenges of interoperability of the new tools with existing healthcare systems and the complexities of integrating AI into diverse clinical workflows.

Introduction of AI tools in the clinic also holds challenges for the education of the employers. Concerns are for example raised that radiation oncologists or therapists might loose the skills to delineate structures or generate a treatment plan [9]. Awareness that AI will have a large influence on clinical practice in the future and core curricula need to be adapted [48].

For AI algorithms in a medical context, the trustworthiness of AI models is of importance. "explainability" and "interpretability" might be ways to make such models more trustworthy and be of added value to make clinical decisions or to be confident enough to implement a model in clinical practice. Explainability of the model requires comprehensive insight into the underlying mechanism of the model which is for most DL models difficult to achieve. Interpretability only requires that clinicians can rationalise the prediction based on scientifically and clinically sound reasoning, e.g., by demonstrating that the prediction is consistent with clinical knowledge. Several methods exist to establish interpretability both for conventional ML as well as DL models [23]. Model explainability loses importance when the outcome of a model is easily interpreted. There is no contradiction here. For example, when a DL model predicts a tumour or organ at risk segmentation, it is useful even if it is not explainable how it is predicted, as the segmentation can be seen superimposed over the images and will undergo deep scrutiny by a radiation oncologist before acceptance. It is however still important to be aware of systematic bias and risks associated with extrapolation of model predictions outside of its initial data set.

Statement 14: An overview of the input data, detailed model prescription and model performance together with intended use and examples where the model does not perform well should be made available (see Appendix B and C for examples) (highly recommended).

Statement 15: In order to achieve clinical applicability, close and effective collaboration among the academia, industrial partners, radiation oncologist, medical physicists, radiation therapists, mathematicians, computer scientists, and data scientists is required (recommended).

Statement 16: Have procedures in place for periodic QA of your AI tool before clinical introduction (highly recommended).

Statement 17: Have procedures in place for upgrades and updates before clinical introduction (highly recommended).

Statement 18: Rigorous testing is key of AI development and serves as the gateway for its deployment (highly recommended).

Statement 19: Trustworthy AI is key for safe and beneficial implementation in radiotherapy (recommended).

The field of AI for use in radiotherapy is very rapidly evolving. Although this guideline gives a broad overview of the topic and presents recommendations that are sufficiently generic to be applicable for some time, an update of the guideline and the underlying literature within a few years seems warranted.

CRediT authorship contribution statement

Coen Hurkmans: Writing – original draft. **Jean-Emmanuel Bibault:** Writing – original draft. **Kristy K. Brock:** Writing – original draft. **Wouter van Elmpt:** Writing – original draft. **Mary Feng:** Writing – original draft. **Clifton David Fuller:** Writing – original draft. **Barbara A. Jereczek-Fossa:** Writing – original draft. **Stine Korreman:** Writing – original draft. **Guillaume Landry:** Writing – original draft. **Frederic Madesta:** Writing – original draft. **Chuck Mayo:** . **Alan McWilliam:** Writing – original draft. **Filipe Moura:** Writing – original draft. **Ludvig P. Muren:** Writing – original draft. **Issam El Naqa:** Writing – original draft. **Jan Seuntjens:** Writing – original draft. **Vincenzo Valentini:** Writing – original draft. **Michael Velec:** Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Marjan Sharabiani for her invaluable help in the conceptualization of this work, guidance through the Delphi process and fruitful scientific discussions. She was supported by a grant from the EORTC Cancer Research Fund. We acknowledge the comprehensive review of this guideline by the ESTRO and AAPM reviewers. We sincerely appreciate their valuable comments and suggestions, which helped improving the quality of the manuscript.

Disclaimer

ESTRO cannot endorse all statements or opinions made on the guidelines. Regardless of the vast professional knowledge and scientific expertise in the field of radiation oncology that ESTRO possesses, the Society cannot inspect all information to determine the truthfulness, accuracy, reliability, completeness or relevancy thereof. Under no circumstances will ESTRO be held liable for any decision taken or acted upon as a result of reliance on the content of the guidelines.

The component information of the guidelines is not intended or implied to be a substitute for professional medical advice or medical care. The advice of a medical professional should always be sought prior to commencing any form of medical treatment. To this end, all component information contained within the guidelines is done so for solely educational and scientific purposes. ESTRO and all of its staff, agents and members disclaim any and all warranties and representations with regards to the information contained on the guidelines. This includes any implied warranties and conditions that may be derived from the aforementioned guidelines.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2024.110345.

References

[1] American Society of Clinical Oncology. Outcomes of cancer treatment for technology assessment and cancer treatment guidelines. J Clin Oncol 1996;14:671–9. https://doi.org/10.1200/JCO.1996.14.2.671.

[2] White Paper on AI - A European approach to excellence and trust. European Union. 2021. commission.europa.eu.

[3] Abernethy AP, Etheredge LM, Ganz PA, et al. Rapid-learning system for cancer care. J Clin Oncol 2010;20:4268–74. https://doi.org/10.1200/JCO.2010.28.5478.

[4] Ahmed KA, Caudell JJ, El-Haddad G, et al. Radiosensitivity Differences Between Liver Metastases Based on Primary Histology Suggest Implications for Clinical Outcomes After Stereotactic Body Radiation Therapy. Int J Radiat Oncol Biol Phys 2016;95:1399–404. https://doi.org/10.1016/j.ijrobp.2016.03.050.

[5] Bakx N, van der SM, Theuws J, Bluemink H, Hurkmans C. Comparison of the output of a deep learning segmentation model for locoregional breast cancer radiotherapy trained on 2 different datasets. Tech Innov Patient Support Radiat Oncol 2023;26:100209. doi: 10.1016/j.tipsro.2023.100209.

[6] Balagurunathan Y, Mitchell R, El Naqa I. Requirements and reliability of AI in the medical context. Phys Med 2021;83:72–8. https://doi.org/10.1016/j.ejmp.2021.02.024.:72-78.

[7] Baroudi H, Brock KK, Cao W, et al. Automated Contouring and Planning in Radiation Therapy: What Is 'Clinically Acceptable'? Diagnostics (Basel) 2023;13:667. https://doi.org/10.3390/diagnostics13040667.

[8] Brady AP, Allen B, Chong J *et al.* Developing, purchasing, implementing and monitoring AI tools in radiology: Practical considerations. A multi-society statement from the ACR, CAR, ESR, RANZCR & RSNA. J Med Imaging Radiat Oncol 2024;68:7-26. doi: 10.1177/08465371231222229.

[9] Brouwer CL, Boukerroui D, Oliveira J, et al. Assessment of manual adjustment performed in clinical practice following deep learning contouring for head and neck organs at risk in radiotherapy. Phys Imaging Radiat Oncol 2020;16:54–60. https://doi.org/10.1016/j.phro.2020.10.001.

[10] Carlson JN, Park JM, Park SY, et al. A machine learning approach to the accurate prediction of multi-leaf collimator positional errors. Phys Med Biol 2016;61:2514–31. https://doi.org/10.1088/0031-9155/61/6/2514.

[11] Chandra SS, Bran LM, Liu X, et al. Deep learning in magnetic resonance image reconstruction. J Med Imaging Radiat Oncol 2021;65:564–77. https://doi.org/10.1111/1754-9485.13276.

[12] Chen A, Chen F, Li X, et al. A Feasibility Study of Deep Learning-Based Auto-Segmentation Directly Used in VMAT Planning Design and Optimization for Cervical Cancer. Front Oncol 2022;12:908903. https://doi.org/10.3389/fonc.2022.908903.

[13] Chen X, Ma X, Yan X, et al. Personalized auto-segmentation for magnetic resonance imaging-guided adaptive radiotherapy of prostate cancer. Med Phys 2022;49:4971–9. https://doi.org/10.1002/mp.15793.

[14] Chuang KC, Giles W, Adamson J. A tool for patient-specific prediction of delivery discrepancies in machine parameters using trajectory log files. Med Phys 2021;48:978–90. https://doi.org/10.1002/mp.14670.

[15] Chun J, Park JC, Olberg S, et al. Intentional deep overfit learning (IDOL): A novel deep learning strategy for adaptive radiation therapy. Med Phys 2022;49:488–96. https://doi.org/10.1002/mp.15352.

[16] Cohen O, Zhu B, Rosen MS. MR fingerprinting Deep RecOnstruction NEtwork (DRONE). Magn Reson Med 2018;80:885–94. https://doi.org/10.1002/mrm.27198.

[17] Collins GS, Dhiman P, ndaur Navarro CL, 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 2021;11:e048008–. https://doi.org/10.1136/bmjopen-2020-048008.

[18] Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet 2019;20:1577–9. https://doi.org/10.1016/S0140-6736(19)30037-6.

[19] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). Ann Intern Med 2015;19:735–6. https://doi.org/10.7326/M14-0698.

[20] Cruz RS, Liu X, Chan AW, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med 2020;26:1351–63. https://doi.org/10.1136/bmj.m3210.

[21] Cui S, Tseng HH, Pakela J, Ten Haken RK, El N, I.. Introduction to machine and deep learning for medical physicists. Med Phys 2020;47:e127–47. https://doi.org/10.1002/mp.14140.

[22] Dal PA, Dirix P, Khoo V, et al. ESTRO ACROP guideline on prostate bed delineation for postoperative radiotherapy in prostate cancer. Clin Transl Radiat Oncol 2023;41:100638. https://doi.org/10.1016/j.ctro.2023.100638.

[23] Diamant A, Chatterjee A, Vallières M, Shenouda G, Seuntjens J. Deep learning in head & neck cancer outcome prediction. Sci Rep 2019;9:2764–39206. https://doi.org/10.1038/s41598-019-39206-1.

[24] Dong P, Lee P, Ruan D, et al. 4π non-coplanar liver SBRT: a novel delivery technique. Int J Radiat Oncol Biol Phys 2013;85:1360–6. https://doi.org/10.1016/j.ijrobp.2012.09.028.

[25] Drabiak K, Kyzer S, Nemov V, El Naqa I. AI and machine learning ethics, law, diversity, and global impact. Br J Radiol 2023;96:20220934. https://doi.org/10.1259/bjr.20220934.

[26] Dutch Ministry of Health. Guideline for high-quality diagnostic and prognostic applications of AI in healthcare. webpage visited 15-2-2024.

[27] El Naqa, I., Murphy, M. J. Machine and Deep Learning in Oncology, Medical Physics and Radiology. 2022. Springer.

[28] El Naqa I. Prospective clinical deployment of machine learning in radiation oncology. Nat Rev Clin Oncol 2021;18:605–6. https://doi.org/10.1038/s41571-021-00541-w.

[29] El Naqa I, Boone JM, Benedict SH, et al. AI in medical physics: guidelines for publication. Med Phys 2021;48:4711–4. https://doi.org/10.1002/mp.15170.

[30] El Naqa I, Li H, Fuhrman J, et al. Lessons learned in transitioning to AI in the medical imaging of COVID-19. J Med Imaging (Bellingham) 2021;8:010902. https://doi.org/10.1117/1.JMI.8.S1.010902.

[31] Eppenhof KAJ, Maspero M, Savenije MHF, et al. Fast contour propagation for MR-guided prostate radiotherapy using convolutional neural networks. Med Phys 2020;47:1238–48. https://doi.org/10.1002/mp.13994.

[32] Eriksson O, Zhang T. Robust automated radiation therapy treatment planning using scenario-specific dose prediction and robust dose mimicking. Med Phys 2022;49:3564–73. https://doi.org/10.1002/mp.15622.

[33] European Parliament. EU AI Act. webpage last visited 15-2-2024.

[34] Fda. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)- based software as a medical device (SaMD). Food and Drug. Administration 2019.

[35] Fransson S, Tilly D, Strand R. Patient specific deep learning based segmentation for magnetic resonance guided prostate radiotherapy. Phys Imaging Radiat Oncol 2022;23:38–42. https://doi.org/10.1016/j.phro.2022.06.001.

[36] Fuhrman JD, Gorre N, Hu Q, et al. A review of explainable and interpretable AI with applications in COVID-19 imaging. Med Phys 2022;49:1–14. https://doi.org/10.1002/mp.15359.

[37] Gong C, Zhu K, Lin C, et al. Efficient dose-volume histogram-based pretreatment patient-specific quality assurance methodology with combined deep learning and machine learning models for volumetric modulated arc radiotherapy. Med Phys 2022;49:7779–90. https://doi.org/10.1002/mp.16010.

[38] Gugliandolo SG, Pepa M, Isaksson LJ, et al. MRI-based radiomics signature for localized prostate cancer: a new clinical tool for cancer aggressiveness prediction? Sub-study of prospective phase II trial on ultra-hypofractionated radiotherapy (AIRC IG-13218). Eur Radiol 2021;31:716–28. https://doi.org/10.1007/s00330-020-07105-z.

[39] Hernandez V, Hansen CR, Widesott L, et al. What is plan quality in radiotherapy? The importance of evaluating dose metrics, complexity, and robustness of treatment plans. Radiother Oncol 2020;153:26–33. https://doi.org/10.1016/j.radonc.2020.09.038.

[40] Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. J Am Med Inform Assoc 2020;27:2011–5. https://doi.org/10.1093/jamia/ocaa088.

[41] Holmes J, Liu Z, Zhang L, Ding Y, Sio TT, McGee LA, Ashman JB, Li X, Liu T, Shen J, Liu W. Evaluating Large Language Models on a Highly-specialized Topic, Radiation Oncology Physics. Front. Oncol 2023 Jul;17:1219326. https://doi.org/10.3389/fonc.2023.1219326.

[42] Hosny A, Bitterman DS, Guthier CV, et al. Clinical validation of deep learning algorithms for radiotherapy targeting of non-small-cell lung cancer: an observational study. Lancet Digit Health 2022;4:e657–66. https://doi.org/10.1016/S2589-7500(22)00129-7.

[43] Hsu c-c., Sandford BA. The Delphi technique: making sense of consensus. 2019.

[44] Huang Y, Pi Y, Ma K, et al. Virtual Patient-Specific Quality Assurance of IMRT Using UNet++: Classification, Gamma Passing Rates Prediction, and Dose Difference Prediction. Front. Oncol 2021;%20;11:700343. https://doi.org/10.3389/fonc.2021.700343.

[45] Huang Y, Pi Y, Ma K, et al. Deep Learning for Patient-Specific Quality Assurance: Predicting Gamma Passing Rates for IMRT Based on Delivery Fluence Informed by log Files. Technol Cancer Res Treat 2022;21. https://doi.org/10.1177/15330338221104881. 15330338221104881.

[46] Hussein M, Clementel E, Eaton DJ, et al. A virtual dosimetry audit - Towards transferability of gamma index analysis between clinical trial QA groups. Radiother Oncol 2017;125:398–404. https://doi.org/10.1016/j.radonc.2017.10.012.

[47] Jabbarpour A, Mahdavi SR, Vafaei SA, et al. Unsupervised pseudo CT generation using heterogenous multicentric CT/MR images and CycleGAN: Dosimetric assessment for 3D conformal radiotherapy. Comput Biol Med 2022;143:105277. https://doi.org/10.1016/j.compbiomed.2022.105277.

[48] Kang J, Thompson RF, Aneja S, et al. National Cancer Institute Workshop on Artificial Intelligence in Radiation Oncology: Training the Next Generation. Pract Radiat Oncol 2021;11:74–83. https://doi.org/10.1016/j.prro.2020.06.001.

[49] Kawula M, Hadi I, Nierer L, et al. Patient-specific transfer learning for auto-segmentation in adaptive 0.35 T MRgRT of prostate cancer: a bi-centric evaluation. Med Phys 2023;50:1573–85. https://doi.org/10.1002/mp.16056.

[50] Kimura Y, Kadoya N, Oku Y, et al. Error detection model developed using a multi-task convolutional neural network in patient-specific quality assurance for volumetric-modulated arc therapy. Med Phys 2021;48:4769–83. https://doi.org/10.1002/mp.15031.

[51] Kiser KJ, Barman A, Stieb S, Fuller CD, Giancardo L. Novel Autosegmentation Spatial Similarity Metrics Capture the Time Required to Correct Segmentations Better Than Traditional Metrics in a Thoracic Cavity Segmentation Workflow. J Digit Imaging 2021;34:541–53. https://doi.org/10.1007/s10278-021-00460-3.

[52] Kneepkens E, Bakx N, van der Sangen M, et al. Clinical evaluation of two AI models for automated breast cancer plan generation. Radiat Oncol 2022;17:25–01993. https://doi.org/10.1186/s13014-022-01993-9.

[53] Langendijk JA, Lambin P, De Ruysscher D, et al. Selection of patients for radiotherapy with protons aiming at reduction of side effects: the model-based approach. Radiother Oncol 2013;107:267–73. https://doi.org/10.1016/j.radonc.2013.05.007.

[54] Lee D, Hu YC, Kuo L, et al. Deep learning driven predictive treatment planning for adaptive radiotherapy of lung cancer. Radiother Oncol 2022;169:57–63. https://doi.org/10.1016/j.radonc.2022.02.013.

[55] Li D, Ma L, Li J, et al. A comprehensive survey on deep learning techniques in CT image quality improvement. Med Biol Eng Compu 2022;60:2757–70. https://doi.org/10.1007/s11517-022-02631-y.

[56] Li X, Zhang J, Sheng Y, et al. Automatic IMRT planning via static field fluence prediction (AIP-SFFP): a deep learning algorithm for real-time prostate treatment planning. Phys Med Biol 2020;65:175014. https://doi.org/10.1088/1361-6560/aba5eb.

[57] Li Z, Zhang W, Li B, et al. Patient-specific daily updated deep learning auto-segmentation for MRI-guided adaptive radiotherapy. Radiother Oncol 2022;177:222–30. https://doi.org/10.1016/j.radonc.2022.11.004.

[58] Liu X, Cruz RS, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet Digit Health 2020;2:e537–48. https://doi.org/10.1016/S2589-7500(20)30218-1.

[59] Liu, Z, Wang, P, Li, Y, Holmes, J, Shu, P, Zhang, L, Liu, C, Liu, N, Zhu, D, Li, X, Li, Q, Patel, S. H., Sio, T. T, Liu, T, Liu, W. RadOnc-GPT: A Large Language Model for Radiation Oncology. arXiv:2309.10160 webpage visited 6-11-2023.

[60] Lombardo E, Dhont J, Page D, et al. Real-time motion management in MRI-guided radiotherapy: Current status and AI-enabled prospects. Radiother Oncol 2023:109970. https://doi.org/10.1016/j.radonc.2023.109970.

[61] Luo Y, Jolly S, Palma D, et al. A situational awareness Bayesian network approach for accurate and credible personalized adaptive radiotherapy outcomes prediction in lung cancer patients. Phys Med 2021;87:11–23. https://doi.org/10.1016/j.ejmp.2021.05.032.

[62] Luo Y, Tseng HH, Cui S, et al. Balancing accuracy and interpretability of machine learning approaches for radiation treatment outcomes modeling. BJR Open 2019;1:20190021. https://doi.org/10.1259/bjro.20190021.

[63] Matsuura T, Kawahara D, Saito A, et al. Predictive gamma passing rate of 3D detector array-based volumetric modulated arc therapy quality assurance for prostate cancer via deep learning. Phys Eng Sci Med 2022;45:1073–81. https://doi.org/10.1007/s13246-022-01172-w.

[64] Mayo CS, Feng MU, Brock KK, et al. Operational Ontology for Oncology (O3): A Professional Society-Based, Multistakeholder, Consensus-Driven Informatics Standard Supporting Clinical and Research Use of Real-World Data From Patients Treated for Cancer. Int J Radiat Oncol Biol Phys 2023;117:533–50. https://doi.org/10.1016/j.ijrobp.2023.05.033.

[65] Mayo CS, Mierzwa M, Yalamanchi P, et al. Machine Learning Model of Emergency Department Use for Patients Undergoing Treatment for Head and Neck Cancer Using Comprehensive Multifactor Electronic Health Records. JCO Clin Cancer Inform 2023;7:e2200037.

[66] Mayo CS, Moran JM, Bosch W, et al. American Association of Physicists in Medicine Task Group 263: Standardizing Nomenclatures in Radiation Oncology. Int J Radiat Oncol Biol Phys 2018;100:1057–66. https://doi.org/10.1016/j.ijrobp.2017.12.013.

[67] McIntosh C, Conroy L, Tjong MC, et al. Clinical integration of machine learning for curative-intent radiation treatment of patients with prostate cancer. Nat Med 2021;27:999–1005. https://doi.org/10.1038/s41591-021-01359-w.

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

[69] Mylonas A, Booth J, Nguyen DT. A review of artificial intelligence applications for motion tracking in radiotherapy. J Med Imaging Radiat Oncol 2021;65:596–611. https://doi.org/10.1111/1754-9485.13285.

[70] Nguyen D, Lin MH, Sher D, et al. Advances in Automated Treatment Planning. Semin Radiat Oncol 2022;32:343–50. https://doi.org/10.1016/j.semradonc.2022.06.004.

[71] Nguyen D, Sadeghnejad BA, Bohara G, et al. A comparison of Monte Carlo dropout and bootstrap aggregation on the performance and uncertainty estimation in radiation therapy dose prediction with deep learning neural networks. Phys Med Biol 2021;66:054002. https://doi.org/10.1088/1361-6560/abe04f.

[72] Ni Y, Chen S, Hibbard L, Voet P. Fast VMAT planning for prostate radiotherapy: dosimetric validation of a deep learning-based initial segment generation method. Phys Med Biol 2022;67:10–6560/ac80e5.. https://doi.org/10.1088/1361-6560/ac80e5.

[73] Niyazi M, Andratschke N, Bendszus M, et al. ESTRO-EANO guideline on target delineation and radiotherapy details for glioblastoma. Radiother Oncol 2023;184:109663. https://doi.org/10.1016/j.radonc.2023.109663.

[74] Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med 2020;26:1320–4. https://doi.org/10.1038/s41591-020-1041-y.

[75] Osman AFI, Maalej NM. Applications of machine and deep learning to patient-specific IMRT/VMAT quality assurance. J Appl Clin Med Phys 2021;22:20–36. https://doi.org/10.1002/acm2.13375.

[76] Patel BN, Rosenberg L, Willcox G, et al. Human-machine partnership with artificial intelligence for chest radiograph diagnosis. NPJ Digit Med 2019;2:111. https://doi.org/10.1038/s41746-019-0189-7.

[77] Prunaretty J, Güngör G, Gevaert T, et al. A multi-centric evaluation of self-learning GAN based pseudo-CT generation software for low field pelvic magnetic resonance imaging. Front. Oncol 2023;13:1245054. https://doi.org/10.3389/fonc.2023.1245054.

[78] Rusanov B, Hassan GM, Reynolds M, et al. Deep learning methods for enhancing cone-beam CT image quality toward adaptive radiation therapy: A systematic review. Med Phys 2022;49:6019–54. https://doi.org/10.1002/mp.15840.

[79] Sadeghnejad BA, Ogunmolu O, Jiang S, Nguyen D. A fast deep learning approach for beam orientation optimization for prostate cancer treated with intensity-modulated radiation therapy. Med Phys 2020;47:880–97. https://doi.org/10.1002/mp.13986.

[80] Scott JG, Sedor G, Scarborough JA, et al. Personalizing Radiotherapy Prescription Dose Using Genomic Markers of Radiosensitivity and Normal Tissue Toxicity in NSCLC. J Thorac Oncol 2021;16:428–38. https://doi.org/10.1016/j.jtho.2020.11.008.

[81] Song P, Eldar YC, Mazor G, Rodrigues MRD. HYDRA: Hybrid deep magnetic resonance fingerprinting. Med Phys 2019;46:4951–69. https://doi.org/10.1002/mp.13727.

[82] The White House. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. webpage last visited 30-11-2023.

[83] Thor M, Oh JH, Apte AP, Deasy JO. Registering Study Analysis Plans (SAPs) Before Dissecting Your Data-Updating and Standardizing Outcome Modeling. Front Oncol 2020;10:978. https://doi.org/10.3389/fonc.2020.00978.

[84] Uk, Department of Health and Social Care. A guide to good practice for digital and data-driven health technologies 2021.

[85] van de Sande D, Sharabiani M, Bluemink H, et al. Artificial intelligence based treatment planning of radiotherapy for locally advanced breast cancer. Phys Imaging. Radiat Oncol 2021;20:111–6. https://doi.org/10.1016/j.phro.2021.11.007.

[86] Vandewinckele L, Claessens M, Dinkla A, et al. Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. Radiother Oncol 2020;153:55–66. https://doi.org/10.1016/j.radonc.2020.09.008.

[87] Vapnik VN. The nature of statistical learning theory. Springer; 2000.

[88] Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med 2022;28:924–33. https://doi.org/10.1038/s41591-022-01772-9.

[89] Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ 2020;20:l6927. https://doi.org/10.1136/bmj.l6927.

[90] Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med 2019;170:51–8. https://doi.org/10.7326/M18-1376.

[91] Wolfs CJA, Varfalvy N, Canters RAM, et al. External validation of a hidden Markov model for gamma-based classification of anatomical changes in lung cancer patients using EPID dosimetry. Med Phys 2020;47:4675–82. https://doi.org/10.1002/mp.14385.

[92] Wootton LS, Nyflot MJ, Chaovalitwongse WA, Ford E. Error Detection in Intensity-Modulated Radiation Therapy Quality Assurance Using Radiomic Analysis of Gamma Distributions. Int J Radiat Oncol Biol Phys 2018;102:219–28. https://doi.org/10.1016/j.ijrobp.2018.05.033.

[93] Yang X, Li S, Shao Q, et al. Uncertainty-guided man-machine integrated patient-specific quality assurance. Radiother Oncol 2022;173:1–9. https://doi.org/10.1016/j.radonc.2022.05.016.
[94] Yuan Z, Wang Y, Hu P, et al. Accelerate treatment planning process using deep learning generated fluence maps for cervical cancer radiation therapy. Med Phys 2022;49:2631–41. https://doi.org/10.1002/mp.15530.
[95] Zaharchuk G, Davidzon G. Artificial Intelligence for Optimization and Interpretation of PET/CT and PET/MR Images. Semin Nucl Med 2021;51:134–42. https://doi.org/10.1053/j.semnuclmed.2020.10.001.