Contents lists available at ScienceDirect

Clinical Oncology

journal homepage: www.clinicaloncologyonline.net

Improving the Quality of Care in Radiation Oncology using Artificial Intelligence

S.M.H. Luk [*][†], E.C. Ford [*], M.H. Phillips [*], A.M. Kalet [*]

[*] Department of Radiation Oncology, University of Washington, Seattle, Washington, USA
[†] Department of Radiation Oncology, University of Vermont Medical Center, Burlington, Vermont, USA

Received 23 September 2021; received in revised form 20 October 2021; accepted 12 November 2021

Abstract

Radiation therapy is a complex process involving multiple professionals and steps from simulation to treatment planning to delivery, and these procedures are prone to error. Additionally, the imaging and treatment delivery equipment in radiotherapy is highly complex and interconnected and represents another risk point in the quality of care. Numerous quality assurance tasks are carried out to ensure quality and to detect and prevent potential errors in the process of care. Recent developments in artificial intelligence provide potential tools to the radiation oncology community to improve the efficiency and performance of quality assurance efforts. Targets for artificial intelligence enhancement include the quality assurance of treatment plans, target and tissue structure delineation used in the plans, delivery of the plans and the radiotherapy delivery equipment itself. Here we review recent developments of artificial intelligence applications that aim to improve quality assurance processes in radiation therapy and discuss some of the challenges and limitations that require further development work to realise the potential of artificial intelligence for quality assurance.

Key words: Artificial intelligence; quality assurance; radiation oncology

Introduction

Radiation therapy is a complex healthcare process involving multiple professionals and steps from simulation to treatment planning to delivery [1].

Many of the processes are prone to error and impact outcomes. Thus, numerous quality assurance tasks are carried out to detect and minimise the frequency and effect of these potential errors. Quality assurance standards are discussed in various reports, including task group reports from American Association of Physicists in Medicine (AAPM) (e.g. [2–5]). Completing these tasks requires substantial human effort and the performance of quality assurance tasks can be limited by available resources. The recent developments of artificial intelligence provide potential tools to the radiation oncology community to improve both the efficiency and/or performance of quality assurance efforts.

Artificial intelligence refers to manmade machines that show intelligence similar to natural creatures such as humans. In this context, we define artificial intelligence as computer software that mimics human reasoning to some degree by learning from historical data. Machine learning algorithms, such as neural networks, and statistical models, such as Bayesian networks, are a few examples of algorithms that could be used as artificial intelligence. Recently, due to the improvements in computation power and the amount of digital data available, artificial intelligence has grown rapidly as an academic field and is now prolific among commercial industries. Artificial intelligence serves as an excellent candidate to support different activities in radiation oncology because of its ability to learn (synthesise new knowledge) from historical clinical data (a vast source of existing knowledge) and mimic human reasoning in these decision-making processes. Figure 1 shows the

Author for correspondence: S.M.H. Luk, Department of Radiation Oncology, University of Vermont Medical Center, 111 Colchester Ave, Burlington, VT, 05401, USA.

E-mail address: samuel.luk@uvmhealth.org (S.M.H. Luk).

https://doi.org/10.1016/j.clon.2021.11.011
0936-6555/© 2021 The Royal College of Radiologists. Published by Elsevier Ltd. All rights reserved.

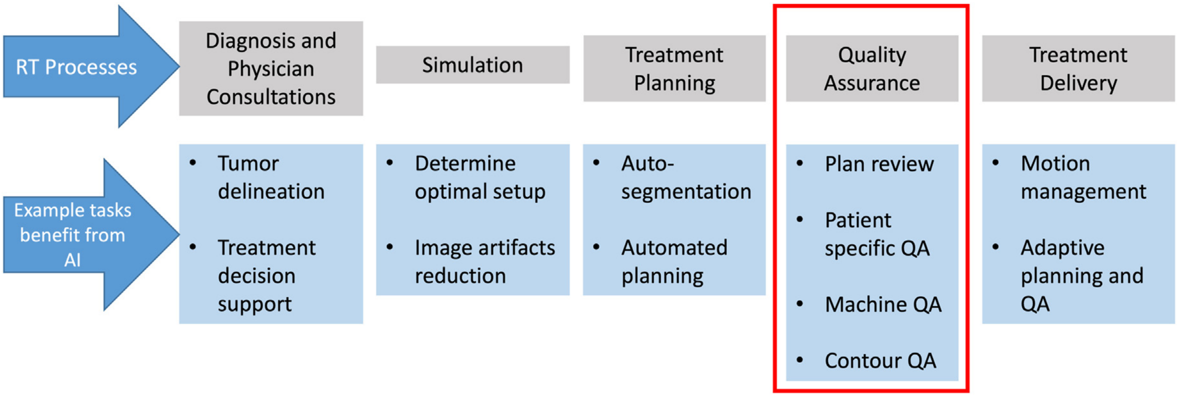

Fig 1. Processes in radiation therapy and the tasks that could benefit from the development of artificial intelligence applications. This review focuses on the quality assurance aspect in the processes.

processes in radiation therapy and some example tasks that could benefit from the development of artificial intelligence applications.

Currently, artificial intelligence is implemented in a limited manner in treatment planning and auto-segmentation of structures, with other ongoing development work on other aspects in radiation oncology. One of the main contributions of artificial intelligence tools is in the improvement of quality of care and the detection of errors, as evidenced by recently published reviews [6–10]. The importance of safety in radiation oncology, the automation of time-consuming quality assurance tasks and personalising quality assurance to specific patients are a few reasons that it is important to develop artificial intelligence tools for error detection.

Here we review recent developments of artificial intelligence applications that aim to improve quality assurance processes in radiation therapy. We focus especially on machine quality assurance, patient-specific quality assurance (PSQA), treatment plan review and the quality assurance of contours. We also discuss some of the challenges and limitations that require further development work to realise the potential of artificial intelligence for quality assurance.

Quality Assurance of Radiotherapy Machines

The quality assurance of treatment machines is one of the essential parts of a comprehensive quality assurance programme to ensure the delivered dose is within 5% of the prescribed dose [2]. The quality assurance procedures of linear accelerators (linacs) have been well described in various reports [3,11–13], and include dosimetry, mechanical, safety, imaging and respiratory gating. The frequencies of these procedures are tabulated into daily, weekly, monthly and annually with different tolerances. Artificial intelligence models are useful for a few of the machine quality assurance tasks, including building trends of machine performance, finding outliers in linac performance from periodical quality assurance results and improving the efficiency of quality assurance procedures.

Dosimetric quality assurance involves the measurement of linac radiation delivered to quality assurance phantoms or ion chambers with the goal of ensuring that radiation parameters have not deviated over tolerances from commissioning baselines. Trends of the measurement data are important indicators of linac performance and failure rates. Standard methods can be used to evaluate dosimetric trends. For example, Chan *et al.* [14] found cyclic trends on linac performance, such as output drift, when they visually examined the daily quality assurance data over 5 years. Artificial intelligence models could provide more predictions of trends on linac performance using these rich but rarely analysed data. Li and Chan [15] used 5 years of daily dosimetric quality assurance measurement results to develop an artificial neural network (ANN) that predicts future trends of linac dosimetric performance and detects potential anomalous events. The ANN was compared with the typical statistical autoregressive moving average time series prediction model and shown to have similar prediction accuracy and lower mean-squared error. Similarly, Zhao *et al.* [16] trained a multivariate regression model with beam data acquired during water tank measurements for commissioning/annual quality assurance of multiple Varian TrueBeam in nine institutions to predict beam-specific PDDs and profiles of different field sizes using only a 10 × 10 cm^2 field as input. The predictions were within 1% for PDDs and profiles were within 4%. The largest uncertainties were at the build-up region for PDDs and penumbra for profiles. The model could be used in annual dosimetric quality assurance of linacs of the same model and could reduce the number of measurements required.

Quality assurance of imaging systems on linacs becomes more important with the rapid development of image-guided radiotherapy (IGRT) and adaptive radiotherapy. The quality of pretreatment and mid-treatment images is essential for the success of IGRT and adaptive radiotherapy, as problems such as image artifacts could affect clinical decisions. Valdes *et al.* [17] developed an automated procedure for quality assurance of the IGRT system on a linac. Image artifacts were artificially introduced during the reconstruction process and a support vector machine (SVM) algorithm was trained to identify these artifacts. The SVM was able to identify beam hardening, rings and scatter artifacts in the automated quality assurance process.

Mechanical checks are carried out to ensure that the mechanical components of linacs and ancillary devices (e.g. lasers) are operating as expected. Although traditionally these measurements are carried out using mechanical tools, automated processes are emerging to identify deviations of mechanical components, such as multileaf collimators (MLC). Approaches use both linac log files and electronic portal imaging device (EPID) measurements [18]. El Naqa *et al.* [19] used a support vector data description clustering model to classify EPID test results of a quality assurance phantom into normal versus outliers for image-based mechanical quality assurance. These artificial intelligence models have the potential to provide reliable and reproducible automation of mechanical quality assurance for the linac.

The development of artificial intelligence models for the quality assurance of imaging and treatment delivery devices is an area that has promising potential for further research. The quality assurance of devices involves a large amount of structured quality assurance data measured daily, monthly, annually, etc., all of which provides a valuable dataset for the development of a reliable artificial intelligence model. The differences in quality assurance protocols and the quality of measurements are the challenges in making these artificial intelligence models operate reproducibly at different institutions [20,21].

Patient-specific Quality Assurance

PSQA has been a focus of artificial intelligence for quality assurance development in radiation oncology. PSQA involves the quality assurance of intensity-modulated radiation therapy (IMRT) plans prior to treatment. Although PSQA is widely used and potentially capable of preventing catastrophic mistakes, studies have suggested that PSQA is one of the least effective methods for detecting errors overall [22]. This underscores the need to enhance the effectiveness and efficiency of the methods involved. As such, PSQA is an excellent target for the implementation of artificial intelligence as a prediction and classification problem. Both supervised and unsupervised machine learning models, as well as other statistical approaches, have been used to develop artificial intelligence models for PSQA. These developments aim to improve the efficiency of the time-consuming PSQA measurement process and inform physicists on treatment plans that could potentially fail quality assurance prior to measurement.

Table 1 provides a summary of studies to date that use artificial intelligence models for PSQA. Common to many of these approaches is the use of the gamma criteria as an end point. That is, many studies seek to develop artificial intelligence models to predict whether a plan will ‘pass’ or ‘fail’ according to a cut-off in the gamma passing rate (GPR), i.e. the percentage of measurement points that exceed a threshold gamma value, typically set at 1. One such early study from Valdes *et al.* [23] developed a virtual IMRT quality assurance framework using a Poisson regression model to predict GPR using plans extracted from the Eclipse treatment planning system (TPS; Varian Inc.) coupled with measurements from a two-dimensional diode array. This model was further validated against measurements using portal dosimetry at another institution and results indicated that the model can predict GPR using different measurement techniques and across multiple institutions [24]. Other studies have aimed to predict GPR using convolution neural networks (CNN) [25] and multiple tree-based machine learning algorithms [26] using predefined features designed by domain experts. In a somewhat different approach, Interian *et al.* [27] trained a CNN using fluence maps calculated from the TPS to predict GPR and showed a similar performance with the domain expert model in the study from Valdes *et al.* [23]. Other similar studies are shown in Table 1. Of note some of these studies examined volumetric modulated arc therapy (VMAT) plans (versus IMRT) and show a similar performance [28–31].

The abovementioned models detect quality assurance failures caused by all potential sources of error. There are also artificial intelligence models that seek to detect specific causes of failure. One series of studies examined failures related to the MLC positional errors. Wootton *et al.* [32] and Nyflot *et al.* [33] used planner gamma distributions of measured dose and the TPS calculated dose of 186 IMRT fields delivered to a phantom and measured with EPID to construct a classifier model to detect MLC errors. Doses from the TPS were calculated with no error, random and systematic mispositioning of MLCs. Features were either defined via a radiomics analysis on two-dimensional gamma images [32] or extracted from a CNN [33] with triplet learning, and multiple classifier models were compared. Other studies have used log files from linacs (which record the positions of all the mechanical components during treatment) to develop machine learning models [34–36] to predict delivered leaf positions. These predicted positions were compared with a planned position to determine errors in MLC.

These models all achieved excellent prediction accuracy on GPR or ‘pass’/‘fail’ of IMRT and VMAT PSQA (see Table 1). However, there are multiple limitations on using gamma criteria as the observed end point of a plan that ‘passes’/‘fails’ quality assurance tests [4]. First, GPR data are usually biased [29], as most clinical data have high GPR for the gamma criteria that are used clinically (e.g. 2%/2 mm or 3%/3 mm). This can result in imbalanced datasets in which only a small percentage of cases are in the ‘fail’ class. Multiple efforts have been undertaken to improve the prediction accuracy in this situation, including the use of a balanced class sampling technique [30]. The use of stricter criteria (e.g. 1%/1 mm) has also been used to increase the number of failed cases [37], although this may not reflect clinical practice. Second, artificial intelligence models usually require large datasets for training and validation. Tomori *et al.* [37] developed a novel strategy to establish a deep learning-based GPR prediction model for VMAT using dummy plans instead of historic patient data, which could potentially be a benefit in developing artificial intelligence

Table 1
Artificial intelligence models developed for patient-specific quality assurance based on gamma criteria

Reference	Artificial intelligence model	Data	Detector	Key results
[23]	Poisson regression model	498 IMRT plans	2D diode array (MapCheck2)	3%/3 mm GPR predicted with an error smaller than 3%
[25]	Convolutional neural network	60 prostate IMRT plans	Gafchromic EBT3 film	Strong or moderate correlation between various predicted and measured GPR (Spearman rank correlation coefficients range from 0.51 to 0.62 in test set)
[26]	Tree-based machine learning algorithms (AdaBoost, Random Forest, XG- Boost)	182 IMRT plans (1497 beams)	EPID	Random Forest and Ada Boost perform slightly better; 95–98% of predictions within 3% of the measured GPR; maximum error = 4.5%
[27]	Convolution neural network	498 IMRT plans	2D diode array (MapCheck2)	3%/3 mm GPR prediction has a mean absolute error of 0.7 versus do main expert model of 0.74
[28]	Regression tree analysis, multiple regression analysis and neural networks	600 VMAT plans	Helical diode array (Arc- Check)	Prediction errors (mean +standard deviation) are mostly within 3% Neural network perform slightly better than the other 2 Models
[29]	Poisson Lasso regression model and Random Forest classification model	176 GYN and 127 H&N VMAT plans	2D ion chamber array (MatriXX)	Mean prediction error of Poisson Lasso model ranges from 1.81 to 4.18%. for different GPR criteria. Poisson Lasso has higher specificity while Random Forest has 100% sensitivity
[30]	Auto-encoder-based classification-regression model	576 plans VMAT	2D ion chamber array (MatriXX)	Absolute prediction error ranges from 1.76 to 4.66% for different GPR criteria, which is lower than Poisson Lasso model (2.10–5.29%)
[31]	Forests of extra-trees, mutual information and linear regression for feature selection; linear models, support vector machines, tree-based models and neural network to predict GPR	500 plans VMAT	2D diode array (MapCheck2)	Support vector machine model has the lowest mean absolute error (3.75%)
[32]	Logistic regression models	186 IMRT beams from 23 patients	EPID	Area under receiver operator characteristic curve is 0.761 on classifying a plan with and without error compared with 0.512 for threshold- based gamma analysis
[33]	Convolution neural network with triplet learning versus handcrafted approach for feature extraction; support vector machines, multi-layer perceptrons, decision trees and k-nearest- neighbours for classification	186 IMRT beams from 23 patients	EPID	Deep learning versus handcrafted features - 77.3% versus 66.3% accuracy for two-class (erroneous versus error-free) experiment and 64.3% versus 53.7% for three-class (error-free versus random versus systemic multileaf collimator error) experiment

EPID, electronic portal imaging device; GPR, gamma passing rate; GYN, gynaecology; H&N, head and neck; IMRT, intensity-modulated radiotherapy; VMAT, volumetric modulated arc therapy.

models for PSQA. Finally, gamma criteria itself may not be a meaningful end point. Gamma metric-based quality assurance does not seem to be sensitive to any but the largest errors. One example of this can be seen in a large multi-institutional study from IROC-H [38], which showed that IMRT quality assurance using gamma criteria does not predict plans that failed IROC-H audits. Similar conclusions were made in other studies (e.g. [39,40]), that gamma criteria in IMRT quality assurance are insensitive to clinically relevant dosimetric errors.

Table 2
Artificial intelligence models developed for patient-specific quality assurance that use dose/fluence maps or log files

Reference	Artificial intelligence model	Data	Detector/log files	Key results
[41]	Convolutional neural network	161 beams from 104 clinical prostate VMAT plans	Cylindrical 3D detector (Delta 4)	Overall accuracy on classifying 'error-free', 'systematic error' and 'random error' is 0.944
[42,43]	Artificial neural network	30 fluence maps of IMRT prostate fields	EPID	Predicted dose maps show an average GPR of 96% with 3%/3 mm criteria
[44]	Support machine vector	1620 VMAT Fields	Cylindrical 3D detector (Delta 4)	AUC of 0.88 on classifying the median dose difference between measured and planned dose to be 'hot' ($> +1\%$), 'nor mal' (within ± 1%) and 'cold' ($< -1\%$)
[35]	Artificial neural network	10 IMRT plans re-delivered without patient	Log files	Predicted and delivered MLC position has a root mean squared error of 0.0097 mm, whereas the delivered MLC and planned MLC could be deviated up to a few mm
[36]	Simple/multiple linear regression, decision tree and ensemble method (boosted tree and bagged tree model)	142 IMRT plans and 125 VMAT plans	Trajectory log file	Correlation coefficients between predicted and actual IMRT and VMAT MLC errors are 0.88 and 0.86, respectively

AUC, area under the receiver operating characteristic curve; EPID, electronic portal imaging device; GPR, gamma passing rate; IMRT, intensity-modulated radiotherapy; MLC, multileaf collimator; VMAT, volumetric modulated arc therapy.

Given the shortcomings of the gamma criteria, it may be productive to use a more meaningful end point in training artificial intelligence models [20]. Some studies have appeared using different end points in the model and Table 2 provides a summary. Kimura *et al.* [41] constructed a CNN model to detect MLC positional errors in VMAT PSQA using dose-difference maps between TPS and the measured dose distribution from a cylindrical three-dimensional detector. Similarly, Mahdavi *et al.* [42,43] developed an ANN model to predict dose maps from EPID fluence maps and used as a pretreatment dose verification by comparing the dose map generated by the ANN and the TPS. Granville *et al.* [44] included both treatment plan characteristics as well as linac performance metrics as part of the features to train a SVM classifier model to predict the difference between the measured dose distribution of VMAT plans. After using a recursive feature elimination cross-validation algorithm to eliminate the least important features, the top 10 important features include five from the plan and five from the linac performance. Models that include dose, fluence maps or log files (Table 2) could potentially serve as alternatives of gamma analysis and may be favoured over GPR prediction models in the future.

Treatment Plan Review

A treatment plan review is a complex process that requires a physicist and an oncologist to review the radiotherapy treatment plan including technical parameters, accuracy of dosimetric calculation, image guidance requests, plan quality, etc. [5]. Such plans are created by multiple professionals using often multiple computer software platforms. There are multiple varieties of plan and chart reviews carried out over the whole treatment process, including a pretreatment plan review, a weekly chart review and an end-of-treatment chart review. The pretreatment plan review has been shown to be one of the most effective quality control process to detect errors such as the wrong isocentre [22]. However, the actual performance of this quality assurance measure has been shown to be lower than expectations [45]. AAPM Task Group 275 provides recommendations for plan and chart reviews [5]. The report also recommends automation of various components of the initial plan check and other chart review processes to improve efficiency and effectiveness.

There have been a number of studies that aim to automate plan and chart reviews in order to reduce error rates and increase efficiency. Many use a rules-based approach to assist the plan checking process [46–56]. These reports enumerate a number of fixed rules and judge a plan based on the match between predefined rules and plan parameters. The rules-based approach performs well on this task, but the adaptability and ability to perform more nuanced reasoning are limited [9,57]. These limitations could potentially be overcome using artificial intelligence applications as they are designed to mimic human reasoning and adapt to changes in clinical practice by updating the models with new data. However, despite the importance of treatment plan review in the quality management programme and the advancement on artificial intelligence development in radiation oncology, very few studies to date have explored the application of artificial intelligence models to assist plan review. Azmandian *et al.* [58] developed an outlier-detection model using a k-means clustering algorithm for plan review of new prostate cases. The goal was to classify whether treatment plans belong to a cluster developed using historic techniques (i.e. ‘four-field’ box prostate plans). Such approaches are promising, but it must be recognised that, like many artificial intelligence approaches, such classifiers may be able to identify problematic plans but do not provide any information for troubleshooting the causes of problems.

An alternative approach that addresses these limitations is using a probabilistic graphical model [59]. Bayesian networks are probabilistic-directed graphical models that contain two components: a graphical topology that describes the relationship between different variables and conditional probability tables that describe the probabilities of different events. The Bayesian network fits well to assist pretreatment review as it has a network structure that represents the relationships between variables that describe human understanding on a domain, conditional probabilities that mimic human reasoning in judging the suitability of plan parameters, ability to incorporate the latest clinical data and statistical tools to help troubleshooting with an understandable network structure [57]. Bayesian networks have been used for a decision support system for different purposes in radiation oncology [60–62] and Kalet *et al.* [63] developed an error detection Bayesian network (EDBN) model to assist pretreatment plan checks. The Bayesian network topology was constructed using a dependency-layered ontology for radiation oncology [64] while conditional probability tables were learnt from historical clinical data. The EDBNs mimicked human reasoning processes to determine the appropriateness of the treatment plan by assessing the conditional probabilities of a set of parameters given some initial diagnostic information, such as staging and tumour location, and flag parameters that have a low probability as a potential error. Also, the graph structure of EDBN help to determine whether a flagged parameter is an error or a deliberate variation, and identify the potential causes of error. Luk *et al.* [57] further expanded the EDBN to 29 nodes to include diagnostic, prescription, plan/beam and set-up parameters. The expanded Bayesian network achieved an area under the receiver operating characteristic curve (AUC) of 0.89 on simulated errors and addressed a few practical questions, including the time window of training data (4 years) and retraining frequency (annually). The study also identified the plan/beam parameters as the most varying parameter group over the years that reduce the performance of the EDBN.

Similarly, Chang *et al.* [65] developed and validated Bayesian network models to detect errors in physician orders. Compared with Luk *et al.* [57], this study focused on prescriptions and divided the orders into three groups: single prescription, concurrent boost and sequential boost. A Bayesian network was developed for each group to detect errors in new orders using joint posterior probabilities of the order parameters given the disease information. The model achieved an AUC of 0.986 for single prescription physician orders.

Quality Assurance on Contours

Contours of treatment targets and organs at risk are one of the most important aspects in radiotherapy, as the quality of contours affects the therapeutic and adverse effects of the treatment. These contours are drawn by clinicians with the help of different imaging modalities, anatomic knowledge and patient examinations. The process of contour delineation is a particularly high-risk aspect of radiotherapy planning, as shown by the work in AAPM Task Group 275 [5]. As such, the quality assurance of contours is essential but is usually in the form of human inspection, such as peer-review in chart rounds.

A variety of tools, including artificial intelligence models, have been developed to assist contouring for treatment planning, also known as auto-segmentation, which is out of the scope of this review. Related to this are artificial intelligence models developed to perform quality checks, such as labelling and geometry, on contours. McIntosh *et al.* [66] built a conditional random forest classifier that learnt from 17 579 regions of interest (ROIs) labelled by experts to automatically label ROIs and detect errors relating to naming and contour quality. The model achieved a classification accuracy of 91.6% on ROI labelling and AUC of 0.75 on detecting contouring errors. In another study, Rhee *et al.* [67] developed a CNN-based auto-contouring tool that was trained and validated with clinical contours of 16 head and neck structures on computed tomography scans of 3495 patients to automatically detect errors in auto-contours from a clinically validated multi-atlas-based auto-contouring tool. The model achieved average AUCs of 0.98 and 0.85 on classifying unacceptable auto-contours. These artificial intelligence models on contouring quality assurance could be used on verifying contours that are drawn by human or other auto-contouring algorithms by comparing them with artificial intelligence-generated contours using geometric measures [68], and potentially act as an integral part of the online quality assurance process for adaptive radiotherapy.

An artificial intelligence model was also proposed to improve efficiency and reduce variability on contour quality assurance for clinical trials. Nijhuis *et al.* [69] trained two CNNs for right parotid and submandibular glands using 735 clinically delineated computed tomography scans of head and neck patients to identify deliberate contour errors. Contours highlighted by the model as erroneous were visually inspected. Among the flagged contours, 26% and 47% were actually deemed clinically suboptimal for parotid and submandibular glands, respectively. On the other hand, 11% and 13% of samples of non-flagged contours were found to be suboptimal for parotid and submandibular glands, respectively, suggesting that the automated contour quality assurance is feasible but visual inspection is still essential. This will probably be an area of active development in the future.

Discussion and Conclusion

Artificial intelligence applications have shown great promise for improving the quality assurance processes in radiation oncology. This could improve the overall safety and efficiency of radiotherapy. Artificial intelligence models could be applied on a variety of quality assurance processes. For machine quality assurance, artificial intelligence builds trends on machine performance to make predictions on potential failure of linacs, identify outliers in machine performance from periodic quality assurance results and provide automation of quality assurance tasks to improve efficiency. For PSQA, artificial intelligence models could predict IMRT quality assurance results or dose distribution in the treatment plans. This process could act as virtual quality assurance, which improves PSQA efficiency and detects potential errors before performing PSQA measurements. For treatment plan review, artificial intelligence models can factor in different information of a treatment plan to assist physicists on judging the appropriateness of the technical aspects of treatment. For contouring, artificial intelligence can check the quality of contours delineated by clinicians and/or auto-contouring algorithms, which could help to standardise the quality of contours. Compared with traditional approaches, artificial intelligence has the advantage of learning from clinical data with computational algorithms in the era with a growing amount of available data and mimic human reasoning on judgmental tasks such as quality assurance processes that require analysis on multiple information and measurement results.

In spite of this promise, there are still improvements required to support the implementation of artificial intelligence models for quality assurance in clinics. First, most proposed models used single institutional data on model development, testing and validation. Recent studies [24,70] showed the importance of external validation and the potential of incompatibility of models in different institutions due to differences in clinical practice. A potential alternative solution is to use centralised data as a platform to test and validate the artificial intelligence models. For example, the AAPM Work Group for Prevention of Errors has developed a pool of treatment plans with simulated errors and made it accessible to registered members [71]. Although the purpose of this project is for the education and training of plan review, this kind of centralised dataset could potentially be used to test the generalisability of plan review artificial intelligence models.

Second, the lack of interpretability of artificial intelligence models and results limits its applicability in the clinic, especially in quality assurance. A ′black-box’ model may help with detecting errors, but it does not provide information to help identify and solve the problems. Moreover, high false-positive rates could be a challenge [57,69], and these are inherent because of the probabilistic nature of most artificial intelligence models. A recent study [72] showed that users could lose trust and ignore results generated by the artificial intelligence model if false-positive rates are too high. Models and/or tools that help to interpret the results would be preferred for artificial intelligence quality assurance application and may address this issue. Inconsistency of model performance could relate to many factors, but data quality is one of the major underlying reasons. Using ontology and distributed learning could help to provide high-quality data for the development of better artificial intelligence applications [9,73].

Moreover, independent quality assurance procedures of artificial intelligence products are required [9] as there is a decay of data relevance for machine-learned medical prediction models [74] and reduced performance in models over time [57,75]. The quality assurance procedures need to confirm the stability of the application over time, ensure a consistent performance and require users to update the model when it is under-performing. Unlike the currently accepted tolerances derived from task group reports and other sources, there are no such standards or guidelines yet for artificial intelligence performance metrics.

In conclusion, this review provides a summary of artificial intelligence models as related to quality assurance in radiotherapy. Targets for artificial intelligence include the routine quality assurance process, machine quality assurance, PSQA, plan reviews and contour quality assurance. Combining these with other artificial intelligence applications that are developed for different processes in radiation oncology, such as auto-segmentation and automated planning, could potentially form a complete artificial intelligence-supported environment in radiation oncology and help improve quality, efficiency, reduce errors and improve patient care in radiation treatment processes.

Statement of Search Strategies Used

Literature search was carried out using PubMed and other relevant medical physics and radiation oncology journals with key words including ‘Quality Assurance’, ‘Radiation Oncology’, ‘Radiotherapy’, ‘Artificial Intelligence’, and ‘Machine Learning’. Additional publications were identified by manually reviewing references cited in articles obtained in the initial search.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Ford EC, Gaudette R, Myers L, Vanderver B, Engineer L, Zellars R, *et al.* Evaluation of safety in a radiation oncology setting using failure mode and effects analysis. *Int J Radiat Oncol Biol Phys* 2009;74:852–858. https://doi.org/10.1016/j.ijrobp.2008.10.038.

[2] Kutcher GJ, Coia L, Gillin M, Hanson WF, Leibel S, Morton RJ, *et al.* Comprehensive QA for radiation oncology - TG 40. *Med Phys* 1994;21:581–618.

[3] Klein EE, Hanley J, Bayouth J, Yin FF, Simon W, Dresser S, *et al.* Task Group 142 report: Quality assurance of medical accelerators. *Med Phys* 2009;36(9Part1):4197–4212. https://doi.org/10.1118/1.3190392.

[4] Miften M, Olch A, Mihailidis D, Moran J, Pawlicki T, Molineu A, *et al.* Tolerance limits and methodologies for IMRT

measurement-based verification QA: recommendations of AAPM Task Group No. 218. *Med Phys* 2018;45(4):e53–e83. https://doi.org/10.1002/mp.12810.

[5] Ford E, Conroy L, Dong L, de Los Santos LF, Greener A, Gwe Ya Kim G, *et al*. Strategies for effective physics plan and chart review in radiation therapy: report of AAPM Task Group 275. *Med Phys* 2020;47(6):e236–e272. https://doi.org/10.1002/mp.14030.

[6] Chan MF, Witztum A, Valdes G. Integration of AI and machine learning in radiotherapy QA. *Front Artif Intell* 2020;3:577620. https://doi.org/10.3389/frai.2020.577620.

[7] Vandewinckele L, Claessens M, Dinkla A, Brouwer C, Crijns W, Verellen D, *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.

[8] Pillai M, Adapa K, Das SK, Mazur L, Dooley J, Marks LB, *et al*. Using artificial intelligence to improve the quality and safety of radiation therapy. *J Am Coll Radiol* 2019;16(9):1267–1272. https://doi.org/10.1016/j.jacr.2019.06.001.

[9] Kalet AM, Luk SMH, Phillips MH. Radiation therapy quality assurance tasks and tools: the many roles of machine learning. *Med Phys* 2020;47(5):e168–e177. https://doi.org/10.1002/mp.13445.

[10] Simon L, Robert C, Meyer P. Artificial intelligence for quality assurance in radiotherapy. *Cancer/Radiotherapie* 2021;25: 623–626. https://doi.org/10.1016/j.canrad.2021.06.012.

[11] Hanley J, Dresser S, Simon W, Flynn R, Klein EE, Letourneau D, *et al*. AAPM Task Group 198 Report: An implementation guide for TG 142 quality assurance of medical accelerators. *Med Phys* 2021;48:e830–e885. https://doi.org/10.1002/mp.14992.

[12] McCullough SP, Alkhatib H, Antes KJ, Castillo S, Fontenot JD, Jensen AR, *et al*. AAPM Medical Physics Practice Guideline 2. b.: Commissioning and quality assurance of X-ray-based image-guided radiotherapy systems. *J Appl Clin Med Phys* 2021;22(9):1–9. https://doi.org/10.1002/acm2.13346.

[13] Smith K, Balter P, Duhon J, White GA, Vassy DL, Miller RA, *et al*. AAPM Medical Physics Practice Guideline 8.a.: Linear accelerator performance tests. *J Appl Clin Med Phys* 2017;18: 23–39. https://doi.org/10.1002/acm2.12080.

[14] Chan MF, Li Q, Tang X, Li X, Li J, Tang G, *et al*. Visual analysis of the daily QA results of photon and electron beams of a trilogy linac over a five-year period. *Int J Med Phys Clin Eng Radiat Oncol* 2015;4(4):290.

[15] Li Q, Chan MF. Predictive time-series modeling using artificial neural networks for Linac beam symmetry: an empirical study. *Ann N Y Acad Sci* 2017;1387(1):84–94. https://doi.org/10.1111/nyas.13215.

[16] Zhao W, Patil I, Han B, Yang Y, Xing L, Schuler E. Beam data modeling of linear accelerators (linacs) through machine learning and its potential applications in fast and robust linac commissioning and quality assurance. *Radiother Oncol* 2020; 153:122–129. https://doi.org/10.1016/j.radonc.2020.09.057.

[17] Valdes G, Morin O, Valenciaga Y, Kirby N, Pouliot J, Chuang C. Use of truebeam developer mode for imaging QA. *J Appl Clin Med Phys* 2015;16(4):322–333. https://doi.org/10.1120/jacmp.v16i4.5363.

[18] Eckhause T, Al-Hallaq H, Ritter T, Demarco J, Farrey K, Pawlicki T, *et al*. Automating linear accelerator quality assurance. *Med Phys* 2015;42:6074–6083. https://doi.org/10.1118/1.4931415.

[19] El Naqa I, Irrer J, Ritter TA, DeMarco J, Al-Hallaq H, Booth J, *et al*. Machine learning for automated quality assurance in radiotherapy: a proof of principle using EPID data description. *Med Phys* 2019;46(4):1914–1921. https://doi.org/10.1002/mp.13433.

[20] Kry SF, Glenn MC, Peterson CB, Branco D, Mehrens H, Steinmann A, *et al*. Independent recalculation outperforms traditional measurement-based IMRT QA methods in detecting unacceptable plans. *Med Phys* 2019;46:3700–3708. https://doi.org/10.1002/mp.13638.

[21] Glenn MC, Peterson CB, Followill DS, Howell RM, Pollard-Larkin JM, Kry SF. Reference dataset of users' photon beam modeling parameters for the Eclipse, Pinnacle, and RayStation treatment planning systems. *Med Phys* 2020;47(1):282–288. https://doi.org/10.1002/mp.13892.

[22] Ford EC, Terezakis S, Souranis A, Harris K, Gay H, Mutic S. Quality control quantification (QCQ): a tool to measure the value of quality control checks in radiation oncology. *Int J Radiat Oncol Biol Phys* 2012;84:E263–E269. https://doi.org/10.1016/j.ijrobp.2012.04.036.

[23] Valdes G, Scheuermann R, Hung C, Olszanski A, Bellerive M, Solberg T. A mathematical framework for virtual IMRT QA using machine learning. *Med Phys* 2016;43(7):4323–4334. https://doi.org/10.1118/1.4953835.

[24] Valdes G, Chan MF, Lim SB, Scheuermann R, Deasy JO, Solberg TD. IMRT QA using machine learning: a multi-institutional validation. *J Appl Clin Med Phys* 2017;18(5): 279–284. https://doi.org/10.1002/acm2.12161.

[25] Tomori S, Kadoya N, Takayama Y, Kajikawa T, Shima K, Narazaki K, *et al*. A deep learning-based prediction model for gamma evaluation in patient-specific quality assurance. *Med Phys* 2018;45(9):4055–4065. https://doi.org/10.1002/mp.13112.

[26] Lam D, Zhang X, Li H, Deshan Y, Schott B, Zhao T, *et al*. Predicting gamma passing rates for portal dosimetry-based IMRT QA using machine learning. *Med Phys* 2019;46(10): 4666–4675. https://doi.org/10.1002/mp.13752.

[27] Interian Y, Rideout V, Kearney VP, Gennatas E, Morin O, Cheung J, *et al*. Deep nets vs expert designed features in medical physics: an IMRT QA case study. *Med Phys* 2018; 45(6):2672–2680. https://doi.org/10.1002/mp.12890.

[28] Ono T, Hirashima H, Iramina H, Mukumoto N, Miyabe Y, Nakamura M, *et al*. Prediction of dosimetric accuracy for VMAT plans using plan complexity parameters via machine learning. *Med Phys* 2019;46:3823–3832. https://doi.org/10.1002/mp.13669.

[29] Li J, Wang L, Zhang X, Liu L, Li J, Chan MF, *et al*. Machine learning for patient-specific quality assurance of VMAT: prediction and classification accuracy. *Int J Radiat Oncol Biol Phys* 2019;105:893–902. https://doi.org/10.1016/j.ijrobp.2019.07.049.

[30] Wang L, Li J, Zhang S, Zhang X, Zhang Q, Chan MF, *et al*. Multitask autoencoder based classification-regression model for patient-specific VMAT QA. *Phys Med Biol* 2020;65:235023. https://doi.org/10.1088/1361-6560/abb31c.

[31] Wall PD, Fontenot JD. Application and comparison of machine learning models for predicting quality assurance outcomes in radiation therapy treatment planning. *Inform Med Unlocked* 2020;18:100292. https://doi.org/10.1016/j.imu.2020.100292.

[32] Wootton LS, Nyflot M, Chaovalitwongse WA, Ford E. Error detection in IMRT quality assurance using radiomic analysis of gamma distributions. *Int J Radiat Oncol Biol Phys* 2018; 102(1):219–228. https://doi.org/10.1016/j.ijrobp.2018.05.033.

[33] Nyflot MJ, Thammasorn P, Wootton LS, Ford EC, Chaovalitwongse WA. Deep learning for patient-specific quality assurance: identifying errors in radiotherapy delivery by radiomic analysis of gamma images with

convolutional neural networks. *Med Phys* 2019;46:456–464. https://doi.org/10.1002/mp.13338.

[34] Carlson JN, Park JM, Park SY, Park JI, Choi Y, Ye SJ. A machine learning approach to the accurate prediction of multi-leaf collimator positional errors. *Phys Med Biol* 2016;61(6): 2514–2531. https://doi.org/10.1088/0031-9155/61/6/2514.

[35] Osman AF, Maalej NM, Jayesh K. Prediction of the individual multileaf collimator positional deviations during dynamic IMRT delivery priori with artificial neural network. *Med Phys* 2020;47(4):1421–1430. https://doi.org/10.1002/mp.14014.

[36] 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(3):978–990. https://doi.org/10.1002/mp.14670.

[37] Tomori S, Kadoya N, Kajikawa T, Kimura Y, Narazaki K, Ochi T, *et al*. Systematic method for a deep learning-based prediction model for gamma evaluation in patient-specific quality assurance of volumetric modulated arc therapy. *Med Phys* 2021;48(3):1003–1018.

[38] Kry SF, Molineu A, Kerns JR, Faught AM, Huang JY, Pulliam KB, *et al*. Institutional patient-specific IMRT QA does not predict unacceptable plan delivery. *Int J Radiat Oncol Biol Phys* 2014; 90:1195–1201. https://doi.org/10.1016/j.ijrobp.2014.08.334.

[39] Kruse JJ. On the insensitivity of single field planar dosimetry to IMRT inaccuracies. *Med Phys* 2010;37:2516–2524. https://doi.org/10.1118/1.3425781.

[40] Nelms BE, Zhen H, Tomé WA. Per-beam, planar IMRT QA passing rates do not predict clinically relevant patient dose errors. *Med Phys* 2011;38:1037–1044. https://doi.org/10.1118/1.3544657.

[41] Kimura Y, Kadoya N, Tomori S, Oku Y, Jingu K. Error detection using a convolutional neural network with dose difference maps in patient-specific quality assurance for volumetric modulated arc therapy. *Phys Med* 2020;73(March):57–64. https://doi.org/10.1016/j.ejmp.2020.03.022.

[42] Mahdavi SR, Bakhshandeh M, Rostami A, Arfaee AJ. 2D dose reconstruction by artificial neural network for pretreatment verification of IMRT fields. *J Med Imaging Radiat Sc* 2018; 49(3):286–292. https://doi.org/10.1016/j.jmir.2018.05.004.

[43] Mahdavi SR, Tavakol A, Sanei M, Molana SH, Arbabi F, Rostami A, *et al*. Use of artificial neural network for pretreatment verification of intensity modulation radiation therapy fields. *Br J Radiol* 2019;92:20190355. https://doi.org/10.1259/bjr.20190355.

[44] Granville DA, Sutherland JG, Belec JG, Russa DJL. Predicting VMAT patient-specific QA results using a support vector classifier trained on treatment plan characteristics and linac QC metrics. *Phys Med Biol* 2019;64:095017. https://doi.org/10.1088/1361-6560/ab142e.

[45] Gopan O, Zeng J, Novak A, Nyflot M, Ford E. The effectiveness of pretreatment physics plan review for detecting errors in radiation therapy. *Med Phys* 2016;43:5181–5187. https://doi.org/10.1118/1.4961010.

[46] Furhang EE, Dolan J, Sillanpaa JK, Harrison LB. Automating the initial physics chart checking process. *J Appl Clin Med Phys* 2009; 10(1):129–135. https://doi.org/10.1120/jacmp.v10i1.2855.

[47] Siochi RA, Pennington EC, Waldron TJ, Bayouth JE. Radiation therapy plan checks in a paperless clinic. *J Appl Clin Med Phys* 2009;10(1):43–62. https://doi.org/10.1120/jacmp.v10i1.2905.

[48] Yang D, Moore KL. Automated radiotherapy treatment plan integrity verification. *Med Phys* 2012;39(3):1542–1551. https://doi.org/10.1118/1.3683646.

[49] Sun B, Rangaraj D, Palaniswaamy G, Yaddanapudi S, Wooten O, Yang D, *et al*. Initial experience with truebeam trajectory log files for radiation therapy delivery verification. *Pract Radiat Oncol* 2013;3(4):e199–e208. https://doi.org/10.1016/j.prro.2012.11.013.

[50] Moore KL, Kagadis GC, McNutt TR, Moiseenko V, Mutic S. Vision 20/20: Automation and advanced computing in clinical radiation oncology. *Med Phys* 2014;41(1):010901. https://doi.org/10.1118/1.4842515.

[51] Xia J, Mart C, Bayouth J. A computer aided treatment event recognition system in radiation therapy. *Med Phys* 2014;41(1): 011713. https://doi.org/10.1118/1.4852895.

[52] Dewhurst JM, Lowe M, Hardy MJ, Boylan CJ, Whitehurst P, Rowbottom CG. Autolock: a semiautomated system for radiotherapy treatment plan quality control. *J Appl Clin Med Phys* 2015;16(3):339–350. https://doi.org/10.1120/jacmp.v16i3.5396.

[53] Hadley SW, Kessler ML, Litzenberg DW, Lee C, Irrer J, Chen X, *et al*. Safetynet: Streamlining and automating QA in radiotherapy. *J Appl Clin Med Phys* 2016;17(1):387–395. https://doi.org/10.1120/jacmp.v17i1.5920.

[54] Holdsworth C, Kukluk J, Molodowitch C, Czerminska M, Hancox C, Cormack RA, *et al*. Computerized system for safety verification of external beam radiation therapy planning. *Int J Radiat Oncol Biol Phys* 2017;98(3):691–698. https://doi.org/10.1016/j.ijrobp.2017.03.001.

[55] Munbodh R, Bowles JK, Zaveri HP. Graph-based risk assessment and error detection in radiation therapy. *Med Phys* 2020;48:965–977. https://doi.org/10.1002/mp.14666.

[56] Covington EL, Chen X, Younge KC, Lee C, Matuszak MM, Kessler ML, *et al*. Improving treatment plan evaluation with automation. *J Appl Clin Med Phys* 2016;17(6):16–31. https://doi.org/10.1120/jacmp.v17i6.6322.

[57] Luk SMH, Meyer J, Young LA, Cao N, Ford EC, Phillips MH, *et al*. Characterization of a Bayesian network-based radiotherapy plan verification model. *Med Phys* 2019;46(5):2006–2014. https://doi.org/10.1002/mp.13515.

[58] Azmandian F, Kaeli D, Dy JG, Hutchinson E, Ancukiewicz M, Niemierko A, *et al*. Towards the development of an error checker for radiotherapy treatment plans: a preliminary study. *Phys Med Biol* 2007;52(21):6511. https://doi.org/10.1088/0031-9155/52/21/012.

[59] Jensen FV, Nielsen TD. . In: *Bayesian networks and decision graphs*, vol. 2. Springer; 2007.

[60] Smith WP, Doctor J, Meyer J, Kalet IJ, Phillips MH. A decision aid for intensity-modulated radiation-therapy plan selection in prostate cancer based on a prognostic Bayesian network and a Markov model. *Artif Intell Med* 2009;46(2):119–130. https://doi.org/10.1016/j.artmed.2008.12.002.

[61] Meyer J, Phillips MH, Cho PS, Kalet I, Doctor JN. Application of influence diagrams to prostate intensity-modulated radiation therapy plan selection. *Phys Med Biol* 2004;49(9):1637–1653. https://doi.org/10.1088/0031-9155/49/9/004.

[62] Hargrave C, Deegan T, Bednarz T, Poulsen M, Harden F, Mengersen K. An image-guided radiotherapy decision support framework incorporating a Bayesian network and visualization tool. *Med Phys* 2018;45(7):2884–2897. https://doi.org/10.1002/mp.12979.

[63] Kalet AM, Gennari JH, Ford EC, Phillips MH. Bayesian network models for error detection in radiotherapy plans. *Phys Med Biol* 2015;60(7):2735. https://doi.org/10.1088/0031-9155/60/7/2735.

[64] Kalet AM, Doctor JN, Gennari JH, Phillips MH. Developing Bayesian networks from a dependency-layered ontology: a proof-of-concept in radiation oncology. *Med Phys* 2017;44(8): 4350–4359.

[65] Chang X, Li HH, Kalet AM, Yang D. Development and validation of a Bayesian network method to detect external beam radiation therapy physician order errors. *Int J Radiat Oncol Biol Phys* 2019;105(2):423–431. https://doi.org/10.1016/j.ijrobp.2019.05.034.

[66] McIntosh C, Svistoun I, Purdie TG. Groupwise conditional random forests for automatic shape classification and contour quality assessment in radiotherapy planning. *IEEE Trans Med Imaging* 2013;32(6):1043–1057. https://doi.org/10.1109/TMI.2013.2251421.

[67] Rhee DJ, Cardenas CE, Elhalawani H, McCarroll R, Zhang L, Yang J, *et al*. Automatic detection of contouring errors using convolutional neural networks. *Med Phys* 2019;46(11):5086–5097. https://doi.org/10.1002/mp.13814.

[68] Sherer MV, Lin D, Elguindi S, Duke S, Tan LT, Cacicedo J, *et al*. Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: a critical review. *Radiother Oncol* 2021;160:185–191. https://doi.org/10.1016/j.radonc.2021.05.003.

[69] Nijhuis H, van Rooij W, Gregoire V, Overgaard J, Slotman BJ, Verbakel WF, *et al*. Investigating the potential of deep learning for patient-specific quality assurance of salivary gland contours using EORTC-1219-DAHANCA-29 clinical trial data. *Acta Oncol* 2021;60:575–581. https://doi.org/10.1080/0284186X.2020.1863463.

[70] Kalendralis P, Eyssen D, Canters R, Luk SMH, Kalet AM, van Elmpt W, *et al*. External validation of a Bayesian network for error detection in radiotherapy plans. *IEEE Trans Radiat Plasma Med Sci* 2021. https://doi.org/10.1109/trpms.2021.3070656.

[71] Schubert L, Johnson P, Kim G, Faught J. *Hands-on workshop: simulated error training for the physics plan review*. . In: *AAPM Annual Meeting TH-AB-Track*, vol. 7.

[72] Luk SMH, Ford E, Kim M, Phillips M, Hendrickson K, Kalet A. Challenges on implementing an hybrid AI-and-rules based plan check tool in clinical practice - a pilot study. In: *AAPM annual meeting PO-GePV-T-155* 2021.

[73] Phillips MH, Serra LM, Dekker A, Ghosh P, Luk SMH, Kalet A, *et al*. Ontologies in radiation oncology. *Phys Med* 2020;72:103–113. https://doi.org/10.1016/j.ejmp.2020.03.017.

[74] Chen JH, Alagappan M, Goldstein MK, Asch SM, Altman RB. Decaying relevance of clinical data towards future decisions in data-driven inpatient clinical order sets. *Int J Med Inform* 2017;102:71–79. https://doi.org/10.1016/j.ijmedinf.2017.03.006.

[75] Nakatsugawa M, Cheng Z, Kiess A, Choflet A, Bowers M, Utsunomiya K, *et al*. The needs and benefits of continuous model updates on the accuracy of RT-induced toxicity prediction models within a learning health system. *Int J Radiat Oncol Biol Phys* 2019;103(2):460–467. https://doi.org/10.1016/j.ijrobp.2018.09.038.