Physics and Imaging in Radiation Oncology 29 (2024) 100525

Contents lists available at ScienceDirect

Physics and Imaging in Radiation Oncology

journal homepage: www.sciencedirect.com/journal/physics-and-imaging-in-radiation-oncology

Original Research Article

Multi-institutional generalizability of a plan complexity machine learning model for predicting pre-treatment quality assurance results in radiotherapy

Michaël Claessens [a,b,*], Geert De Kerf [a], Verdi Vanreusel [a,b,c], Isabelle Mollaert [a], Victor Hernandez [d], Jordi Saez [e], Núria Jornet [f], Dirk Verellen [a,b]

[a] Department of Radiation Oncology, Iridium Netwerk, Wilrijk (Antwerp), Belgium
[b] Centre for Oncological Research (CORE), Integrated Personalized and Precision Oncology Network (IPPON), University of Antwerp, Antwerp, Belgium
[c] Research in Dosimetric Applications (RDA), SCK CEN, Mol (Antwerp), Belgium
[d] Department of Medical Physics, Hospital Sant Joan de Reus, IISPV, 43204 Tarragona, Spain
[e] Department of Radiation Oncology, Hospital Clínic de Barcelona, 08036 Barcelona, Spain
[f] Servei de Radiofísica i Radioprotecció, Hospital de la Santa Creu i Sant Pau, Barcelona, Spain

ARTICLE INFO

Keywords:
Radiation therapy
Machine learning
VMAT
Quality assurance
Plan complexity
Multi-institutional validation

ABSTRACT

Background and purpose: Treatment plans in radiotherapy are subject to measurement-based pre-treatment verifications. In this study, plan complexity metrics (PCMs) were calculated per beam and used as input features to develop a predictive model. The aim of this study was to determine the robustness against differences in machine type and institutional-specific quality assurance (QA).

Material and methods: A number of 567 beams were collected, where 477 passed and 90 failed the pre-treatment QA. Treatment plans of different anatomical regions were included. One type of linear accelerator was represented. For all beams, 16 PCMs were calculated. A random forest classifier was trained to distinct between acceptable and non-acceptable beams. The model was validated on other datasets to investigate its robustness. Firstly, plans for another machine type from the same institution were evaluated. Secondly, an inter-institutional validation was conducted on three datasets from different centres with their associated QA.

Results: Intra-institutionally, the PCMs beam modulation, mean MLC gap, Q1 gap, and Modulation Complexity Score were the most informative to detect failing beams. Eighty-tree percent of the failed beams (15/18) were detected correctly. The model could not detect over-modulated beams of another machine type. Inter-institutionally, the model performance reached higher accuracy for centres with comparable equipment both for treatment and QA as the local institute.

Conclusions: The study demonstrates that the robustness decreases when major differences appear in the QA platform or in planning strategies, but that it is feasible to extrapolate institutional-specific trained models between centres with similar clinical practice. Predictive models should be developed for each machine type.

1. Introduction

Volumetric modulated arc therapy (VMAT) is currently a state-of-the-art technique for the treatment of different tumour sites by providing a highly conformal dose distribution to the target volume while minimizing the dose deposition in the surrounding organs at risk (OARs) [1]. Despite its advantages compared to previous techniques, the creation of deliverable VMAT plans involves repeatedly solving a large-scale modulation problem with iteratively updated treatment parameters, which in parallel may lead to an (unnecessary) increase in treatment plan complexity [2]. Although a certain degree of complexity may be required to achieve an acceptable dose distribution, it has been reported that increasing plan complexity may lead to higher uncertainties both in dose calculation and in treatment delivery due to limitations in the calculation algorithm or in the defined beam model [2]. In addition, the dose accuracy can be further influenced by mechanical uncertainties as well as patient's position. Consequently, the general suitability of a treatment plan should not only be evaluated

* Corresponding author at: Department of Radiation Oncology, Iridium Netwerk, Oosterveldlaan 24, Antwerp, Belgium.
E-mail addresses: michael.claessens@uantwerpen.be, michael.claessens@zas.be (M. Claessens).

https://doi.org/10.1016/j.phro.2023.100525
Received 27 July 2023; Received in revised form 5 December 2023; Accepted 12 December 2023
Available online 19 December 2023

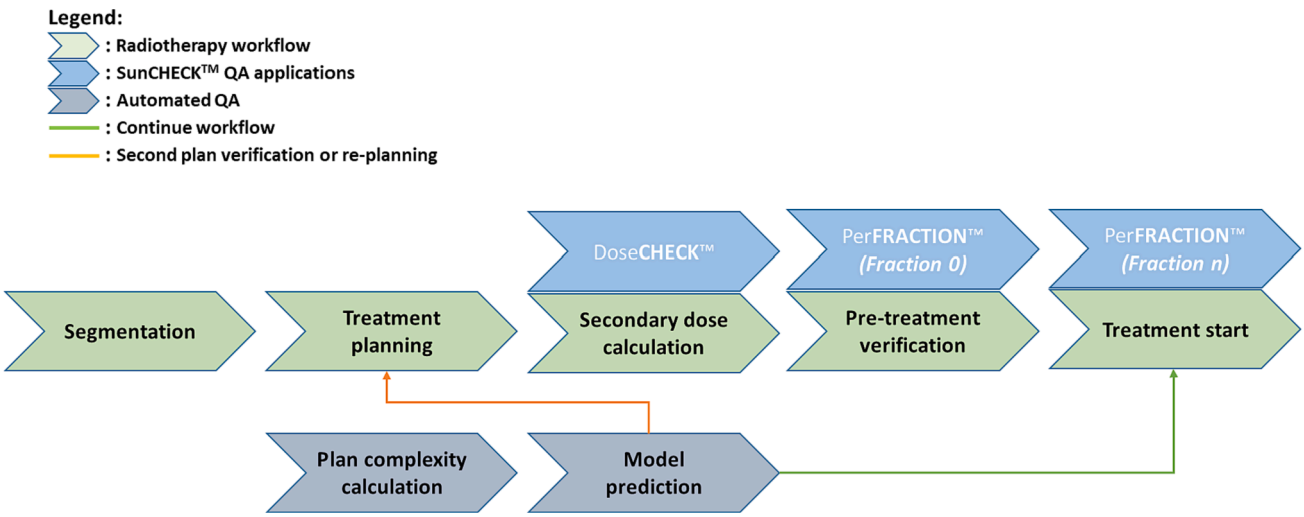

Fig. 1. Representation of the synergetic interaction of the clinical QA workflow with the AI model.

based on calculated dose distributions, but also on accuracy of dose calculation and delivery [3]. Therefore, it is useful to assess the level of plan complexity as it is relevant to avoid future pre-treatment QA failures during the treatment process and to have an overall high-quality plan with a certain degree of robustness against most common uncertainties during treatment delivery.

In order to quantify the complexity of the treatment plan, several plan complexity metrics (PCMs) have been proposed [1,4]. These metrics emphasize multiple aspects of the treatment plan by focusing on for instance beam aperture modulation (e.g., size and irregularity) and machine parameter modulation (e.g., variations in gantry speed, etc.). During the optimization process, PCMs can assist the dosimetrist or medical physicist expert (MPE) to interactively evaluate the trade-offs between dosimetric performance and plan complexity.

Before the actual delivery of the plan to the patient, measurement-based pre-treatment quality assurance (QA) is an established routine practice to detect plans that are too complex to be delivered as planned by comparing the calculated dose with the delivered dose, commonly evaluated using the percent dose difference (DD) and distance to agreement (DTA) (cf. gamma analysis) [5]. As this procedure is considered as a resource-intensive and time-inefficient task, Artificial Intelligence (AI) model architectures have been proposed by multiple institutions to automatically correlate different PCMs to the pre-treatment QA results, in which measurements could be reduced or completely bypassed by accurate predictions [6–8]. Such implementation can have high potential during (real-time) adaptive RT, where measurement-based QA verifications of adapted plans are not possible. Some AI proposals focused on the direct value prediction of gamma passing rate (GPR) based on the complexity features, whereas others focused on a classification between failed and passed plans based on this metric [9,10]. In addition, accelerator parameters can be included in the training data to take into account the day-to-day performance of the machine [7]. More recently, deep learning (DL) algorithms have been proposed to establish GPR prediction models based on planar dose distribution as input [10,11]. Such models can classify whether the beam passes the QA, predict the GPRs of different gamma criteria and predict the trend and position of the dose difference [11].

However, there are still challenges in this RT field and the clinical implementation of methods to control and evaluate plan complexity is very heterogeneous across different RT centres according to the results of a 2020 ESTRO survey [2]. Different PCMs that focus on different plan parameters give different results and there is no clear consensus which ones should be used, and which tolerance level should be assigned. This latter issue shows similarities with heterogeneous measure interpretation for quantitative evaluation of contouring in RT [12] with no straightforward harmony of correct usage. To overcome this issue, it is reasonable to train ML models on PCMs information that has been validated across different centres with the intention to determine the need for pre-treatment verifications based on PCMs and facilitate the use of plan complexity in clinical practice.

Therefore, an in-house ML model was initially trained to correlate PCMs to pre-treatment electronic portal imaging devices (EPID)-based QA failure in the context of multi-site VMAT beams for different treatment sites (i.e., pathologies). Afterwards, the ML model's generalizability and robustness was tested in two steps. Firstly, it was intra-institutionally validated on beams of another machine type to see if models are transferrable to other machine types. Secondly, an inter-institutional validation was conducted on three independent datasets, collected from different RT centres, consisting of VMAT plans with their associated QA results that showed different degrees of similarity in clinical practice in comparison to the centre where the ML was trained.

2. Material and methods

2.1. Proposed QA workflow

A full-integrated web-based software was implemented in 2018 to provide QA data for all intensity modulated radiotherapy (IMRT)/VMAT/dynamic conformal arcs (DCA) plans based on EPID measurements. In Fig. 1, an additional, automated procedure is proposed to complement the current QA workflow. The purpose of the prediction model was not intended to completely replace the established QA flow, but to help reduce the measurement burden of pre-treatment QA. Different PCMs (cf. Section 2.3) were calculated per arc and used as input features for a ML classifier to solve a binary classification problem distinguishing beams that will fail or pass the local QA procedure. The predicted output can automatically trigger a reduction in the level of complexity of the treatment plan by alerting the physicist/dosimetrist investigation (cf. orange arrow). In Fig. S1I (Supplementary materials), an illustration is given of a failed VMAT beam in the QA software.

2.2. In-house data collection

The clinical database was queried for patients that were treated in 2022 for the following treatment sites: prostate, head and neck (H&N), and lung. All plans in this dataset were delivered by the same linear accelerator type (cf. Linac type 1). In a previous study, our institution determined for all different treatment sites and prescriptions the most

Table 1
Overview of the different plan complexity metrics calculated per beam used in the creation of the model. Note that mean DR and mean GS are actually plan parameters, which were computed to fully characterize the treatment plan. For ease of use, they are considered as PCMs in this study.

Plan complexity metric	Characteristics
Modulation Complexity Score [14] (MCS)	This score integrates two contributions to complexity: variability in the shape of segments and variations in their area. The value ranges from 0 to 1 (cf. maximum to no complexity).
Modulation Index Total [15] (MITotal)	This index evaluates the variations in speed and acceleration of the MLC as well as variations of the gantry speed and the dose rate. It is the only complexity index that takes into account the modulation of the dose rate and the gantry speed.
Plan Irregularity [16] (PI)	The metric describes the deviations of aperture shapes from a circle, being 1 for a perfect circle.
Beam Modulation [16] (BM)	It indicates to what extent the beam is delivered into smaller apertures. The values range from 0 to 1 (cf. the higher the value, the more modulated the plan).
EdgeMetric [17] (EM)	This metric computes the complexity of MLC apertures based on the ratio of MLC side edge length within the beam aperture and aperture area.
Leaf Travel [18] (LT) / ArcLength (AL)	This index indicates the average distance travelled by the moving leaves, divided by the AL.
Leaf interdigitation	This refers to the end of a trailing leaf extending past the end of an adjacent leading leaf. Namely, opposing leaves of adjacent rows can overlap [19].
Mean MLC gap	This represents the average leaf pair opening (in mm)
Q1 gap	First quartile of the distribution of leaf gap sizes (mm).
Mean Tongue and Groove index (TGi)	Mean value for the ratio of the distance between adjacent leaves and their MLC gap size. The value ranges from 0 (all leafs aligned) to 1 (full interdigitation and maximum T&G).
Mean MLC Speed	This metric represents the mean speed of all in-field leaves (cm/s).
MLC Speed Modulation	This represents the mean variation of MLC speeds. It is computed as the sum of MLC speed variations divided by the total leaf travel.
Mean Dose Rate (DR)	This defines the Mean Dose Rate (MU/min).
DR modulation	Total Dose Rate (MU/min) variations divided by the arc length.
Mean Gantry Speed (GS)	Average GS (degree/s).
GS Modulation	This metric represents the Total Gantry Speed variation (cf. sum of variations divided by the arc length).

'useable' parameters for pass/fail criteria, i.e. those with a good balance between detection of clinically relevant problems and the total number of false positive results [13]. For the considered pathologies, a gamma passing rate of 3 %/2 mm was applied with a passing tolerance of 98 %. All beams that did not meet these criteria were selected for this study and classified as 'failed'. Further analysis was performed of these individual QA results to avoid the presence of false positives (FP). The most prominent causes for FP were 1) unexpected interruptions of beams, 2) the need for EPID imager calibration, and 3) a wrong set-up position of the EPID imager. As a result, a total of 567 VMAT beams (cf. 477 passed, 90 failed) were collected.

2.3. Plan complexity metrics calculation

An overview of the used PCMs is given in Table 1. These 16 features were calculated by a dedicated MATLAB script for every separate beam, which only needed the DICOM RTplan (cf. dicom file) for analysis. These metrics were subsequently used as input features for an AI model to make correlations with the failure degree of the different beams of VMAT plans.

2.4. Failure prediction

An in-house random forest (RF) model was trained based on PCMs (cf. Section 2.3) of plans delivered on Linac 1. It was used to make a classification between acceptable and non-acceptable beams according to the local QA tolerances (cf. Table 2). The original dataset was divided into a training 453 beams) and test set (114 beams) in a stratified way, where the latter contained the same ratio of classes (96 passed beams and 18 failed beams). During the training phase, k-fold cross-validation ($k = 10$) was used, and various hyperparameter combinations were exhausted by grid search. During every fold, 10 % of the dataset was selected as validation set with the preservation of the relative class frequencies and scored based on balanced accuracy. After cross-validation, the model architecture with highest mean balanced accuracy was re-trained on the whole training set. The performance of the RF model on the test set was characterised by the confusion matrix.

In addition, feature importance was investigated to determine which PCMs had the highest clinical relevance to detect potential QA failure. This was calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability was calculated by the number of samples reaching the node divided by the total number of samples. The higher the value, the more important the feature. An additional decision threshold tuning strategy was performed to fully optimize the decision function. Scripting of these models was performed in Python using dedicated ML libraries (cf. Tensorflow/scikit-learn).

Table 2
Overview of the most important similarities and differences between the different validation sets.

	Study Institution	Institution 1	Institution 2	Institution 3
Number of test beams	114	62	80	40
Anatomy Localisation	Pelvis Chest & Abdomen H&N	Prostate Lung Breast	Prostate Lung H&N	Prostate H&N
TPS algorithm	Eclipse AAAv16.1	Eclipse AAA v16.1	Eclipse AAA v16.1	Eclipse AAA v16.1
Dose resolution	3 mm	2.5 mm	2 mm	2 mm
Complexity reduction strategy used?	No	No	Yes	Yes
Linac Type	Linac type 1	Linac type 1	Linac type 1	Linac type 1
QA device	EPID	EPID	EPID/Point dosimetry	Delta4
(Absolute) DD/DTA	3 %/2 mm	3 %/2 mm	2 %/2 mm	2 %/2 mm
Low threshold	10 % (H&N) 20 % (Chest & Abdomen)	10 %	10 %	20 %
Acceptable (global) GPR	98 %	98 %	95 %	95 %

2.5. Different machine validation

The in-house RF model was only trained on PCMs that were calculated on treatment plans for delivery on Linac type 1. Besides this type of machine, our institution has one 'older generation' device: the Linac type 2, with a Millennium 120 MLC. To investigate if the Linac type 1-dedicated correlation between the PCMs and the QA failure status can be translated to a different machine, a random validation set of 101 Linac type 2 beams (i.e. 42 passed beams and 59 failed beams) was composed. The same pathologies (i.e., prostate, H&N, and lung) were incorporated with the same optimisation/calculation engine and pre-treatment verification criteria. The performance of the RF model was characterised by the confusion matrix.

2.6. Multi-institutional validation

During the commissioning phase, the RF model was exclusively trained and tested on historical patient data. To gain insight in the

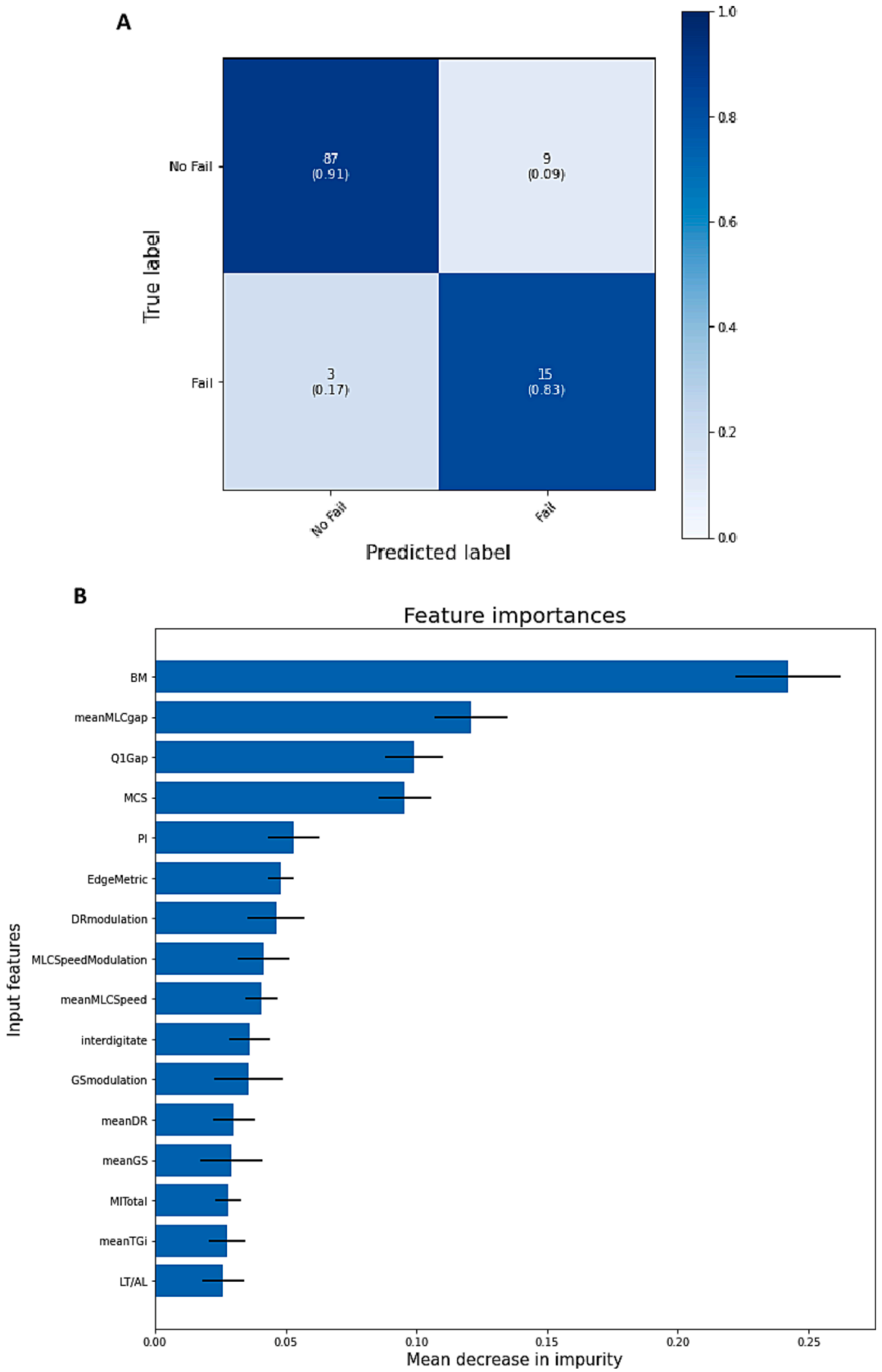

Fig. 2. A) Confusion matrix representing the classification results on the independent test set of the study-specific institution. B) Mean and standard deviation of the relative feature importance of 16 PCMs to distinguish acceptable and non-acceptable beams by the overall RF model.

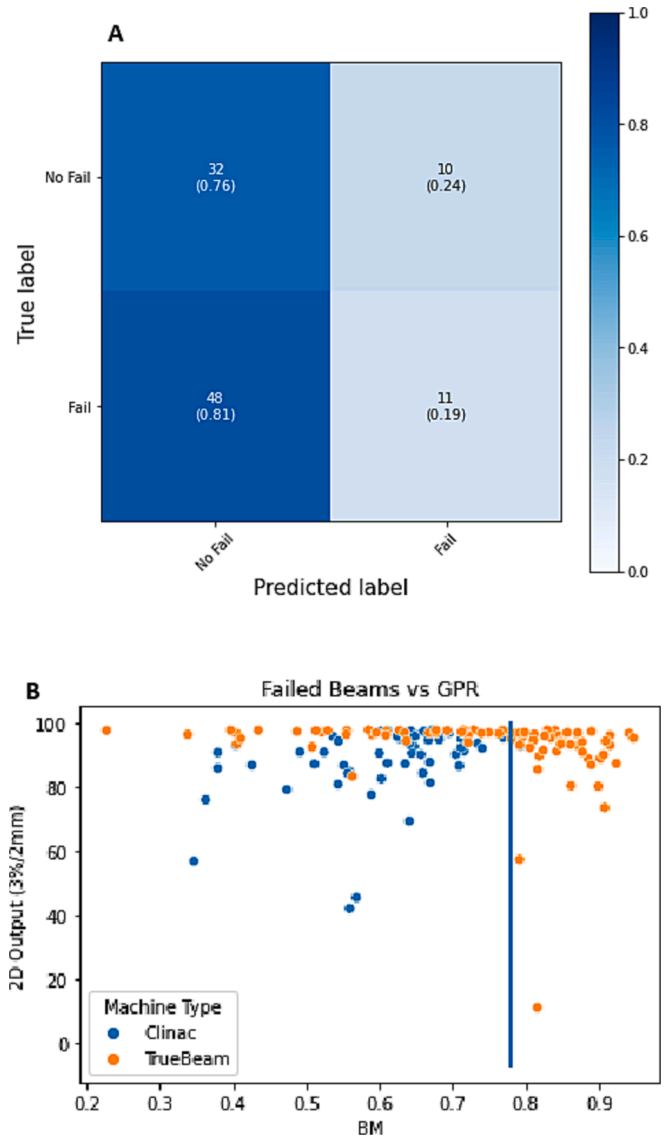

Fig. 3. A) Confusion matrix with the classification results of the RF model for the Linac type 2 validation set. B) Representation of The GPR (3 %/2 mm) vs BM value for the failed Linac type 2 beams (blue) and Linac type 1 beams (orange). In case of the latter, the two beams (both lung) showed a high divergent GPR (11,12 % and 57,41 %). According to the criteria mentioned in Section 2.2, no technical reason could be found to consider these as false positives and were thus maintained in the original dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

general applicability of the correlations between the plan complexity and QA pass/fail status across different centres, treatment plans from three independent RT institutions were collected with their centre-specific QA results. Centres were specifically selected in a way that they showed increased dissimilarity in QA procedures and tolerances in comparison to our institution. Institution 1 has the same QA platform/software with EPID as QA device and 3 %/2 mm as absolute DTA/DD for VMAT plans. Institution 2 gradually deviates from this similarity by using another software in combination with EPID and point dosimetry as QA device and an absolute DTA/DD of 2 %/2 mm. Institution 3 mainly differs from the latter by using Delta4 (*IBA*) as main QA device. Institution 2 and 3 also quantify and minimize plan complexity during plan optimization. In Table 2, an overview of the composition of different validation sets at the level of plan creation, QA procedure and tolerances per centre are given.

Considering the planning and verification processes as a single process could be useful to understand the complex relationship between plan quality, plan complexity, plan deliverability and pre-treatment verification results, especially in a multi-centre environment where multiple planning strategies, anatomical localisations, and TPSs can be presented [20].

3. Results

3.1. Intra-institutional validation

The performance on the independent test set of the overall (random forest) RF model after k-fold (k = 10) cross-validation is shown in the confusion matrix in Fig. 2. The dataset contained 96 passed and 18 failed beams based on the local QA criteria. Fig. 2A shows the classification accuracy of the model, where 83 % of the failed beams (15/18) were assigned to the correct class with 17 % of false negatives (3/18). The ranking of the PCM features used to train the RF model is shown in Fig. 2B. In comparison to the other features, the beam modulation showed the highest correlation to the EPID-based local institute QA results, followed by meanMLCgap, Q1gap and MCS.

In Fig. 3A, the confusion matrix shows that only 19 % of the failed Linac type 2 beams could be detected with a high false negative rate (81 %). In Fig. 3B, according to the high misclassification results, the feature value distribution BM is shown in function of the GPR (cf. 3 %/2 mm) for all the failed Linac type 2 and Linac type 1 beams respectively. As mentioned in Table 2, a VMAT beam passed the local QA with a strict GPR of 98 %. For the Linac type 2 data, failed beams were randomly distributed over the total BM value range. All Linac type 2 beams were under a cut-off value of 0.78 (cf. blue line, Fig. 3.2B). In contrary, the Linac type 1 BM values range up to the maximum value of 1.0 with a majority of failed Linac type 1beams across the 0.78 value. Beneath this specific value, the majority of the Linac type 1 beams had a GPR value close to the acceptance level, where a less strict DTA/DDA, for example 3 %/3 mm, will remarkably reduce the number of failed Linac type 1 beams.

3.2. Inter-institutional validation

In Fig. 4, the classification performance for the independent institutions is shown by means of the confusion matrix. For institution 1, with the same QA procedure as our institution, the RF model could identify 86 % of the beams (6/7) that failed the institutional-specific QA procedures with 14 % of false negatives (1/7). For institution 2, the RF model could detect 40 % of the beams (2/5) that failed the institutional-specific QA procedures with 60 % of false negatives (3/5). For institution 3, the RF model could detect 25 % of the beams (1/4) that failed the institutional-specific QA procedures with 75 % of false negatives (3/4).

As the similarity in clinical practice and QA procedures decreased from institution 1 to 3, the ability of the model to predict accurately the failing status of a VMAT beam dropped simultaneously. In Fig. 5 and Supplementary materials Fig. 2, a comparison is given of the distribution of the four most dominant features across the different centres. Violin plots were used for this comparison to simultaneously illustrate summary statistics and kernel density estimation. It should also be mentioned that the number of failed beams for institution 1, 2 and 3 is respectively seven, five and four. Institution 1 had a higher mean BM compared to our institution and the high predictive power of BM compensates for smaller deviations comparing meanMLCgap, Q1gap and MCS (cf. Fig. 5 and Supplementary materials Fig. 2). For institution 2, small differences for all four parameters decreased model's accuracy. For institution 3, the mean BM values is too low, compared to our institution and did not allow good prediction. A summary of the classification results for the intra- and inter-institutional validation can be found in Supplementary materials Table S1.

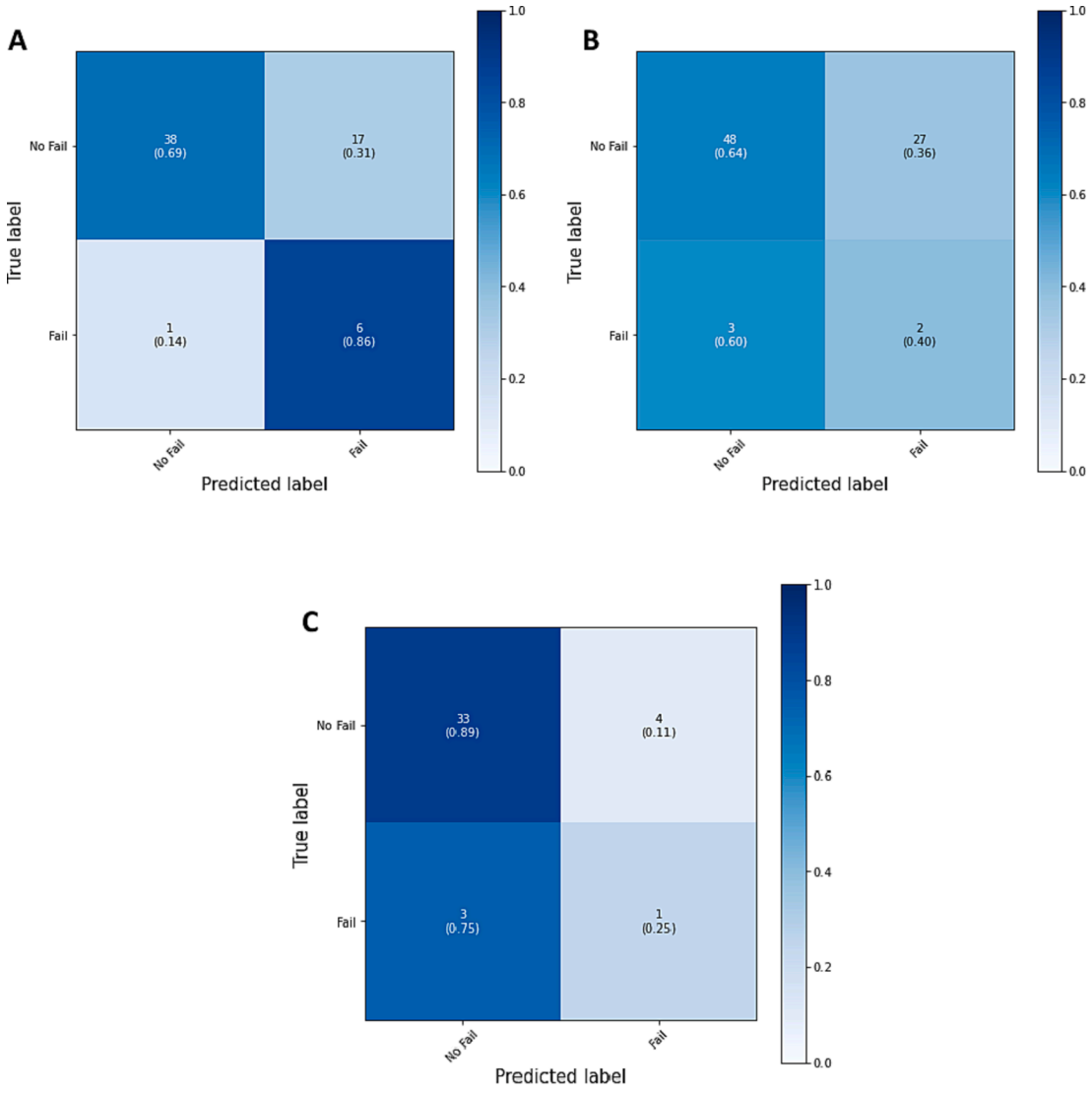

Fig. 4. Confusion matrices representing the classification results of the overall RF model on the three, independent test sets: A) Institution 1, B) Institution 2 and C) Institution 3.

4. Discussion

In this study, we investigated the multi-institutional use of ML model based on PCMs to predict the pre-treatment QA. The robustness of the model decreased when major differences appear in the QA platform or in planning strategies, but that it is feasible to extrapolate institutional-specific trained models between centres with similar clinical practice. Such strategies could be used in combination with secondary dose calculations to trigger further necessary adaptations of the treatment plan and reduce plan-specific QA (PSQA) measurements to streamline the patient's individualized care path in an adaptive RT workflow.

Recently, different studies focused on the use of PCMs to directly assign a degree of delivery uncertainty to each treatment plan [8]. Some complexity indices provided similar information and can be considered equivalent (e.g., MCS, PI and EM). However, indices that focused on different plan parameters yielded different results (cf. MITotal and EM) and there is no general consensus on which tolerance levels should be used for each complexity metric [2,4]. In addition, dedicated evaluations of metrics with the purpose to increase information on the dosimetric uncertainty of a plan beyond common QC results, are scarce [21].

Given the wide variety of metrics available, the first aim of this study was to identify correlations between department's planning data for three different treatment sites (cf. Pelvis, Chest & Abdomen and H&N) and QA results at beam level considering the most relevant published, correlated PCMs (cf. Section 2.3) [2]. Intra-institutionally, Fig. 3B showed that BM, mean MLC gap, Q1 gap, and MCS were top ranked and most correlated to the local institution's QA results, which were not highlighted in other ML-based studies [6,7,9].

To get more insight into the model's general applicability to detect failing beams related to over-modulation, the linear accelerator (Linac) type 1 ML model was tested for another machine type, as well as for treatment plans from different institutions. According to the results in Fig. 3A, the ML model had a low detection rate for failing Linac type 2 beams. This could be explained by the fact that Linac type 1 models are less subject to beam failures than Linacs type 1which has been demonstrated by an institution-specific follow-up study, where on the 'new generation machines', 97 % of measurements passed with the standard 3 %/2 mm analysis, compared to 77 % on the 'old generation machines' [13]. Additional proof was given by plotting the BM distribution for failing Linac type 1 and 2 beams (cf. Fig. 3B). A highest value of approximately 0.78 can be seen for Linac type 2 beams, which shows that the maximum level of Linac type 2-specific over-modulation is lower than Linac type 1, making the model neither applicable nor transferable to Linac type 2 beam QA [20,21].

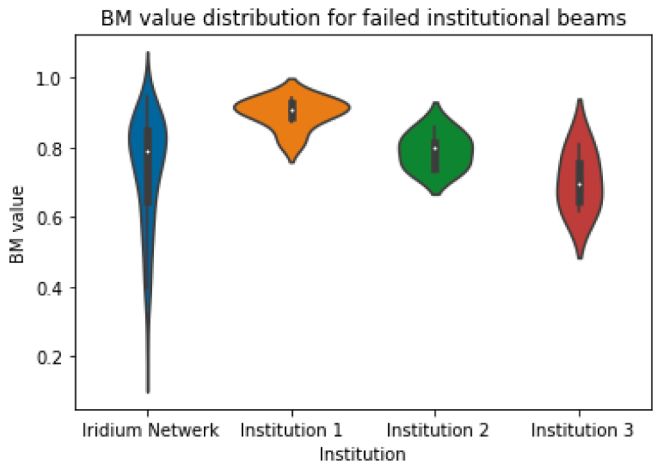

Fig. 5. The comparison of the most important feature (BM) distribution for failing fractions between our institution and the corresponding institution was represented by violin plots. The plot displays separate densities along the y-axis, meaning that there is no overlap between distributions. Wider area of the violin plot represents a higher probability of the values that the PCMs take on, whereas the thinner area corresponds to a lower probability. Note that the negative values in the violin plot are estimations of values of the PCM data caused by the use of kernel density estimation.

Also, the multi-institutional applicability and robustness was tested. Valdes et al., who predicted the GPR value directly, recommended using different models for different combinations of delivery systems and energies [6]. In accordance to our study, Lambri et al. also performed a validation of a XGBoost regression model based on PCMs in a multi-centric scenario [22]. Regression models can provide more quantitative information, whereas classification (cf. our methodology) has the advantage of providing a quick, unambiguous, and actionable result, similar to routine machine QC results. Additionally, regression analyses could typically underestimated large dose discrepancies, possibly due to the relative infrequency of such occurrences [7]. In addition, Li et al. developed both a regression and classification model based on fifty-four complexity metrics and reported that a classifier with high sensitivity was preferred [23]. A quantitative comparison between our model and published models is visualised in Supplementary materials Table 2. In our study, the main focus was put on the differences in clinical practice and QA measurements between the centres. Therefore, three institutions were selected, with increasing dissimilarity in clinical practice and QA procedures with respect to the our institution (cf. Table 2): a) 'Institution 1′ with the same QA platform and equipment (i.e., EPID dosimetry), b) 'Institution 2′ with a different QA platform, but same equipment, and c) 'Institution 3′ with different QA platform and equipment. The main concept of the inter-institutional validation was to create a scenario from narrow to broad generalization (so called domain shift or drift). In addition to the previously defined differences with the local institute, centres 2 and 3 evaluate and minimize plan complexity during the treatment planning process in combination with a strict DD/DTA (cf. 2 %/2 mm), resulting in higher detection rate for institution 1 in comparison to institutions 2 and 3. Ideally, Institution 2 and 3 should use the same DD/DTA criteria as used in the training dataset to have more reliable predictions, but the rationale of the study focused on maintaining the clinical practice without changing the QA procedures. Our results demonstrate that the robustness of the model decreases when major differences appear in the QA platform or in planning strategies (e.g., guided by PCM analysis), but that it is feasible to transfer institutional-specific trained models between centres with similar clinical practice to predict failing beams. Future work in the context of multi-institutional analysis can be to collect a set of plans and measure them in all different institutions. That would isolate completely the QA platform/device component, since the PCMs distributions would be identical for each plan.

It should be mentioned that EPID-based QA results with a predefined DD/DTA and GPR were used as output for the ML models to make a proper distinction between acceptable and failed beams based on PCMs. These QA results are depending on the imager panel characteristics for sensitivity and stability and could result in day-to-day variations, which may affect the GPR results. However, as indicated by Wolfs et al., who compared different dose comparison methods for DL-based QA prediction models, a relative DD/DTA is still beneficial for DL performance in detecting errors [24]. In addition, not all PCMs were analysed in this study.

To conclude, based on the current external validation of its generalizability, it is advisable to train predictive models using local QA results or combining results from other centres with the same QA platform and similar clinical procedures. Based on these results, it seems too early to consider that PCMs can replace PSQA for general VMAT cases. Nevertheless, the approach can help identifying cases that require attention to increase efficiency and streamline the PSQA process and could help to reduce the heterogeneous usage of the PCMs in the RT community.

Ethics approval and consent to participate

The patient data was anonymised before the start of the research, so that they no longer relate to identifiable persons. Data that no longer relate to identifiable persons or data that have otherwise been rendered anonymous so that the data subject cannot be re-identified, are not personal data and are therefore outside the scope of data protection law. Due to this reason, this research did not require ethical approval of the ethics committee.

Availability of data and materials

The dataset used and analysed during the current study are available from the corresponding author on reasonable request.

Funding

Michaël Claessens was supported by a grant of the Flemish League against Cancer, Belgium (ref: 000019356).

CRediT authorship contribution statement

Michaël Claessens: . **Geert De Kerf:** . **Verdi Vanreusel:** . **Isabelle Mollaert:** . **Victor Hernandez:** . **Jordi Saez:** . **Núria Jornet:** . **Dirk Verellen:** Investigation.

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

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.phro.2023.100525.

References

[1] Antoine M, Ralite F, Soustiel C, Marsac T, Sargos P, Cugny A. Use of metrics to quantify IMRT and VMAT treatment plan complexity: a systematic review and

perspectives. Phys Med 2019;64:98–108. https://doi.org/10.1016/j.ejmp.2019.05.024.
[2] Kaplan LP, Placidi L, Bäck A, Canters R, Hussein M, Vaniqui A. Plan quality assessment in clinical practice: results of the 2020 ESTRO survey on plan complexity and robustness. Radiother Oncol 2022;173:254–61. https://doi.org/10.1016/j.radonc.2022.06.005.
[3] Hernandez V, Hansen CR, Widesott L, Bäck A, Canters R, Fusella M. 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.
[4] Hernandez V, Saez J, Pasler M, Jurado-Bruggeman D, Jornet N. Comparison of complexity metrics for multi-institutional evaluations of treatment plans in radiotherapy. Phys Imaging Radiat Oncol 2018;5:37–43. https://doi.org/10.1016/j.phro.2018.02.002.
[5] Low DA, Harms WB, Mutic S, Purdy JA. A technique for the quantitative evaluation of dose distributions. Med Phys 1998;25:656–61. https://doi.org/10.1118/1.598248.
[6] 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;18:279–84. https://doi.org/10.1002/acm2.12161.
[7] Granville DA, Sutherland JG, Belec JG, La Russa DJ. Predicting VMAT patient-specific QA results using a support vector classifier trained on treatment plan characteristics and linac QC metrics. Phys Med Biol 2019;64:95017. https://doi.org/10.1088/1361-6560/ab142e.
[8] 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.
[9] Lam D, Zhang X, Li H, Deshan Y, Schott B, Zhao T. Predicting gamma passing rates for portal dosimetry-based IMRT QA using machine learning. Med Phys 2019;46:4666–75. https://doi.org/10.1002/mp.13752.
[10] Wang L, Li J, Zhang S, Zhang X, Zhang Q, Chan MF, et al. Multi-task autoencoder based classification-regression model for patient-specific VMAT QA. Phys Med Biol 2020;65:235023. https://doi.org/10.1088/1361-6560/abb31c.
[11] Huang Y, Pi Y, Ma K, Miao X, Fu S, Chen H. Virtual patient-specific quality assurance of IMRT using UNet++: classification, gamma passing rates prediction, and dose difference prediction. Front Oncol 2021;11:700343. https://doi.org/10.3389/fonc.2021.700343.
[12] Gooding MJ, Boukerroui D, Vasquez Osorio E, Monshouwer R, Brunenberg E. Multicenter comparison of measures for quantitative evaluation of contouring in radiotherapy. Phys Imaging Radiat Oncol 2022;24:152–8. https://doi.org/10.1016/j.phro.2022.11.009.
[13] Bossuyt E, Weytjens R, Nevens D, De Vos S, Verellen D. Evaluation of automated pre-treatment and transit in-vivo dosimetry in radiotherapy using empirically determined parameters. Phys Imaging Radiat Oncol 2020;16:113–29. https://doi.org/10.1016/j.phro.2020.09.011.
[14] McNiven AL, Sharpe MB, Purdie TG. A new metric for assessing IMRT modulation complexity and plan deliverability. Med Phys 2010;37:505–15. https://doi.org/10.1118/1.3276775.
[15] Park JM, Park S-Y, Kim H, Kim JH, Carlson J, Ye S-J. Modulation indices for volumetric modulated arc therapy. Phys Med Biol 2014;59:7315–40. https://doi.org/10.1088/0031-9155/59/23/7315.
[16] Du W, Cho SH, Zhang X, Hoffman KE, Kudchadker RJ. Quantification of beam complexity in intensity-modulated radiation therapy treatment plans. Med Phys 2014;41:21716. https://doi.org/10.1118/1.4861821.
[17] Younge KC, Matuszak MM, Moran JM, McShan DL, Fraass BA, Roberts DA. Penalization of aperture complexity in inversely planned volumetric modulated arc therapy. Med Phys 2012;39:7160–70. https://doi.org/10.1118/1.4762566.
[18] Masi L, Doro R, Favuzza V, Cipressi S, Livi L. Impact of plan parameters on the dosimetric accuracy of volumetric modulated arc therapy. Med Phys 2013;40:71718. https://doi.org/10.1118/1.4810969.
[19] Ehrgott M, Güler Ç, Hamacher HW, Shao L. Mathematical optimization in intensity modulated radiation therapy. Ann Oper Res 2010;175:309–65. https://doi.org/10.1007/s10479-009-0659-4.
[20] Glenn MC, Hernandez V, Saez J, Followill DS, Howell RM, Pollard-Larkin JM. Treatment plan complexity does not predict IROC Houston anthropomorphic head and neck phantom performance. Phys Med Biol 2018;63:205015. https://doi.org/10.1088/1361-6560/aae29e.
[21] Götstedt J, Bäck A. Edge area metric complexity scoring of volumetric modulated arc therapy plans. Phys Imaging Radiat Oncol 2021;17:124–9. https://doi.org/10.1016/j.phro.2021.02.002.
[22] Lambri N, Hernandez V, Sáez J, Pelizzoli M, Parabicoli S, Tomatis S. Multicentric evaluation of a machine learning model to streamline the radiotherapy patient specific quality assurance process. Phys Med 2023;110:102593. https://doi.org/10.1016/j.ejmp.2023.102593.
[23] Li J, Wang L, Zhang X, Liu L, Li J, Chan MF. 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.
[24] Wolfs CJA, Verhaegen F. What is the optimal input information for deep learning-based pre-treatment error identification in radiotherapy? Phys Imaging Radiat Oncol 2022;24:14–20. https://doi.org/10.1016/j.phro.2022.08.007.