Physics and Imaging in Radiation Oncology 32 (2024) 100685

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

Definition of a framework for volumetric modulated arc therapy plan quality assessment with integration of dose-, complexity-, and robustness metrics

Tina Orovwighose [a,b,*], Bernhard Rhein [a,b,c], Oliver Schramm [a,b], Oliver Jäkel [a,b,c,d], Vania Batista [a,b]

[a] Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Germany
[b] Heidelberg Institute of Radiation Oncology (HIRO), Heidelberg, Germany
[c] Heidelberg Ion-Beam Therapy Center (HIT), Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Germany
[d] Dep. Medical Physics in Radiation Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany

ARTICLE INFO

Keywords:
Dose metrics
Plan complexity metrics
Plan robustness metrics
Statistical process control
Robustness prediction

ABSTRACT

Background and purpose: Conventionally, the quality of radiotherapy treatment plans is assessed through visual inspection of dose distributions and dose-volume histograms. This study developed a framework to evaluate plan quality using dose, complexity, and robustness metrics. Additionally, a method for predicting plan robustness metrics using dose and complexity metrics was introduced for cases where plan robustness evaluation is unavailable or impractical.

Materials and methods: The framework and prediction models were developed and validated using 103-bronchial Volumetric Modulated Arc Therapy (VMAT)-plans. The application of the framework was demonstrated using 25-VMAT-plans. To identify significant metrics for plan evaluation, 122-metrics were analysed and narrowed down using multivariate Spearman correlation. Metric limits were set with Statistical process control (SPC). Robustness metrics were predicted using multivariable or single linear regression models based on dose-and complexity-metrics.

Results: Twenty-five-metrics were selected based on the amount and strength of correlations. R_{95}(dose coverage) and $HI_{95/5}$(homogeneity index) stood out among the dose-metrics, while the complexity-metrics showed similar correlations. Average scenarios dose at 95 % Clinical Target Volume $D95_{mean}$(CTV) and Errorbar-based Volume-Histograms (EVH) were notable for robustness metrics. Approximately 99 % of evaluated metrics fell within established SPC limits. The prediction model for $D95_{mean}$(CTV) showed good performance (adjusted $R^2 = 0.88$, mean squared error (MSE) $= 3.84 \times 10^{-6}$), while the model for EVH demonstrated moderate reliability (adjusted $R^2 = 0.52$, MSE $= 0.2$). No statistically significant differences were found between the predicted (using dose-and complexity-metrics) and calculated robustness metrics (EVH (*p-value* $= 0.9$) and $D95_{mean}$(CTV) (*p-value* $= 1$)).

Conclusions: The developed framework enables early detection of sub-optimal, complex and non-robust treatment plans. The predictive model can be used when robustness evaluations are impractical.

1. Introduction

Traditionally, a treatment plan (TP) is evaluated based on a visual examination of the calculated dose distribution and dose-volume histograms (DVH) [1,2]. However, the expected differences between planned and delivered plans should be addressed during the evaluation. Accurate plan quality assessment should consider qualitative and quantitative measures rather than relying solely on clinical protocols, historical practices, or personal preferences [2]. The literature [3] recommends evaluating plan quality using dose, plan complexity, and robustness metrics and implementing these metrics into the clinical treatment planning system (TPS). Dose metrics (dose coverage, homogeneity, conformity, and gradient index) used to evaluate the dose distribution are mainly based on the DVH data [3]. Plan complexity metrics quantify the degree of complexity associated with machine parameters, TPS properties, and plan characteristics, which can cause discrepancies

* Corresponding author at: Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Germany.
E-mail address: tina.orovwighose@med.uni-heidelberg.de (T. Orovwighose).

https://doi.org/10.1016/j.phro.2024.100685
Received 12 June 2024; Received in revised form 26 November 2024; Accepted 27 November 2024
Available online 29 November 2024

between delivered and calculated dose distributions, potentially affecting robustness [3–5]. In photon therapy, uncertainties are typically addressed by adding a margin to the Clinical Target Volume (CTV) to create a planning target volume (PTV) to ensure adequate CTV coverage despite discrepancies [6]. The classical PTV- or margin-approach has some shortcomings, such as high-dose irradiation of the organ-at-risk (OAR) or static dose cloud approximation, which assumes that the dose distribution is unaffected by uncertainties [7,8]. Conversely, a robust planning approach accounts for uncertainties during the optimisation or evaluation [8,9]. A robust optimisation produces a dose distribution with less variability and greater resilience to uncertainties, while robust evaluation only assesses how well the dose distribution handles uncertainties [7,10]. Robust optimisation methods include minimax [7,11], worst-case [7,11], conditional value at risk [9], and probabilistic planning [7] optimisation. Robust evaluation approaches include scenario-based (individual scenarios); aggregated dose distribution (combination of multiple scenarios, e.g., voxel-wise minimum, mean, and maximum dose) [12]; worst-case scenario [12,13]; and probabilistic robust evaluation [13].

Statistical process control (SPC) aims to monitor process quality by comparing current parameters with previous data [14]. The classical Shewhart control chart is derived under the normality assumption [15]. For non-parametric data (e.g. skewed data), new generation methods (heuristic, transformation, or percentile methods) could be used [16–18]. In radiotherapy, SPC has been used in previous studies to set limits for patient-specific measurements [16,19,20], machine quality assurance tolerances [14,21], and dose comparisons [22].

This study has three parts. First, dose-, complexity-, and robustness-metrics were calculated, and multivariate Spearman correlations were used to identify metrics for evaluating plan quality. SPC was used to determine the undefined limits of the metrics, which vary according to clinical standards, radiation techniques, and treatment sites. Second, regression models were introduced to predict plan robustness using dose and complexity metrics. The prediction could be used when robust evaluation is unavailable or impractical because robust evaluation calculations could be time-consuming, especially during the planning process. The third part demonstrates the application of the framework and the prediction model. The aggregation and visualisation of the results through a user interface are necessary for proper clinical implementation, but this is beyond the scope of this study.

2. Methods and materials

The Ethics Committee of the Medical Faculty, Heidelberg University (S-192/2022), approved this study.

2.1. Treatment plans

Volumetric Modulated Arc Therapy (VMAT) plans from 2020 for patients with bronchial carcinoma were retrospectively analysed. Prescribed doses ranged from 45 to 66 Gy (15–33 fractions). All retrospective plans were generated in the TPS (RayStation-v.10A, RaySearch Lab., Sweden), with 6MV-photon using a collapsed cone algorithm for an Elekta VersaHD LINAC (Elekta Ltd, Crawley, UK). The dose calculation (dose-to-different density water) [23] was performed using a (2 × 3 × 2) mm dose grid, and the plans were normalised to the target volume median dose (D50 %). The baseline and test plans applied in this study are as follows.

Baseline plans: The framework (including SPC limits) and prediction model were established using 103 bronchial VMAT plans (TP1-TP103).

Test plans: The framework application was demonstrated on 25 additional bronchial VMAT plans (TP104-TP128). The original plans were copied and recalculated using an updated TPS version (RayStation-v.11B) with a re-commissioned machine due to a clinical TPS update.

Table 1
Summary of the evaluated metrics grouped by dose, plan complexity, and plan robustness evaluation. In $HI_{95/5}$/$HI_{98/2}$ (95; 98; 5; or 2) refer to the $D_{x\%}$ (dose at x% volume). The criteria 10 mm, 15 mm, and 20 mm were used for SAS and 10 mm × 10 mm, 15 mm × 15 mm, and 20 mm × 20 mm for SA/CP. ROI: The regions of interest include the target volume (CTV) and the OARs (right and left lung, heart and spinal cord). Acronym RTOG: Radiation Therapy Oncology Group. MU is the unit for PMU, cm^2 for MFA and PA, and 1/cm for C/A. The other metrics/parameters listed are unitless.

a) Dose metrics	b) Plan complexity metrics	c) Plan robustness evaluation metrics
– TC – Target Coverage [37] – QC_{RTOG} – RTOG Quality of coverage [38] – R_{95} – Dose Coverage [39] – $CI_{paddick}$ – Paddick Conformity Index [40] – CI_{RTOG} – RTOG Conformity Index [38] – HI_{RTOG} – RTOG Homogeneity Index [38] – $HI_{95/5}$ – Homogeneity Index [41] – $HI_{98/2}$ – Homogeneity Index [39,41] – GI – Gradient Index [2] – DGI – Dose Gradient Index [42] – PQI – Plan Quality Index [25]	– PMU – Plan normalised MU [43] – MFA – Mean Field Area [44] – PA – Plan averaged beam Area [43,45] – PI – Plan averaged beam Irregularity [43,45] – PM – Plan averaged beam Modulation [43,45] – LT – Leaf Travel [46] – MCS – Plan Modulation Complexity Score [46] – LTMCS – LT x MCS [46] – C/A – Circumference/Area [47,48] – $SAS_{(10mm;\ 15mm;\ 20mm)}$ – Small Aperture Score [44] – SA/ $CP_{(10mm\times10mm;15mm\times15mm;\ 20mm\times20mm)}$ – Segment Area per Control Point (CP) [49]	**Target-based:** – EVH – Errorbar-based Volume Histograms (EVH) [50] – RVH – Root Mean Square Dose (RMSD) Volume Histograms (RVH) [51] – DVHB(ROI) – Dose-Volume histogram (DVH)-band of target (DVHB) [52] – $DVHBW_{Dx\%}$ – DVH BandWidths (BW) at $D_{x\%}$ [53] – $D_{x\%}$(ROI) – dose at x% ROI volume as range – scenario dose evaluation – $Passrate_{Dx\%}$(ROI) – ROI clinical goal pass rate **OAR-based:** – DVHB(OAR) [52] (similar to target) – $D_{x\%}$(OAR) (similar to target) – $Passrate_{Dx\%}$(OAR) (similar to target) **Robust-dose-metrics:** – $Dosemetrics_{robust}$ – Robust-dose-metrics. All dose metrics listed in column a) were calculated for the scenarios and recorded as minimum, mean, and maximum.

2.2. Plan quality evaluation metrics

Dose ($n = 11$), complexity ($n = 15$), and robustness metrics ($n = 96$) were calculated using a Python script (v.3.8.7) compiled in the TPS to retrieve plan parameters. All the metrics ($n = 122$) were calculated and examined using the baseline plans. These metrics are listed in Table 1, and their equations are provided in the Supplementary material (Table A.1).

CTV was considered the target volume for the plan quality evaluation (clinical goals: $D_{95} \geq 0.95$; $D_{max} < 1.07$). The OARs clinical goals are from QUANTEC (right/left lung: $V_{20Gy} = 30$ %, $D_{mean} = 7$ Gy; heart: $V_{30Gy} = 46$ %, $D_{mean} = 3$ Gy, $D_{max} = 30$ Gy; spinal cord: $D_{max} = 45$ Gy) [24]. For comparison purposes, the dose values were normalised to the prescribed dose.

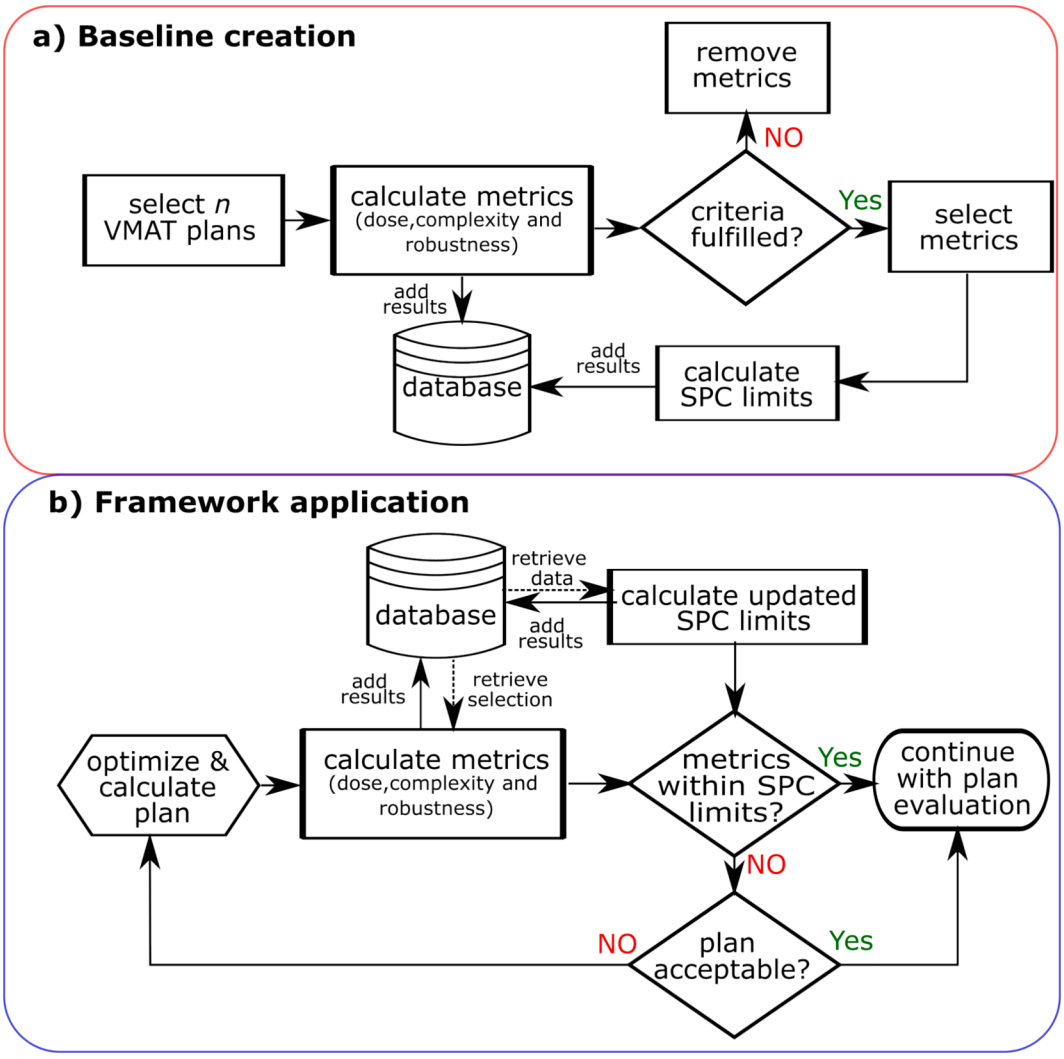

Fig. 1. Established workflow used to assess plan quality. This workflow has two processes: (a) baseline creation (red framed) and (b) an example of the framework application (blue framed). The "baseline creation" refers to the selection of metrics and the calculation of SPC limits needed for plan evaluation. The baseline is created with *n* VMAT treatment plans (e.g., n = 103 plans). The baseline results are saved in a shared database. The "framework application" uses an existing baseline, saved in a database, to evaluate a treatment plan. The step 'optimise & calculate plan" is not part of the framework in this study. Note: The treatment plan can refer to any plan requiring evaluation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.2.1. Dose metrics

Eleven dose metrics were evaluated (Table 1). Plan Quality Index (PQI) combines the target dose coverage and the dose received by OARs [25]. The above-listed OARs were used.

2.2.2. Plan complexity metrics

Fifteen complexity metrics were calculated (Table 1). Small aperture score (SAS) and segment area per control point (SA/CP) were computed using three criteria each (10 mm, 15 mm and 20 mm) for SAS and (10 mm × 10 mm, 15 mm × 15 mm and 20 mm × 20 mm) for SA/CP.

2.2.3. Plan robustness evaluation metrics

The robustness evaluation, essential for calculating robustness metrics, was limited to setup uncertainties for simplicity. A setup uncertainty of 5 mm in all directions was used, which is appropriate for daily imaging with cone-beam computed tomography. The scenario-based evaluation method [12] available in the TPS was used, where the setup error was sampled in different scenarios, and the dose was calculated using the nominal plan beam sets. The setup error was re-sampled by applying isocenter shifts to the diagonal endpoints, referring to points from the centre to the cube's corners formed by the given uncertainties. This resulted in eight dose distribution scenarios used to calculate the robustness metrics. These metrics were divided into three groups (Table 1): target-based, OAR-based and robust-dose-metrics (dose metrics calculated from the scenarios dose). For the robustness based on target (CTV), the following parameters were calculated: i) robustness metrics, as errorbar volume-histogram (EVH), root-mean-square volume-histogram (RVH), DVH-band (DVHB) and DVH band width (DVHBW), ii)Scenarios CTV coverage ($D_{x\%}$(CTV)), $D_{x\%}$ represents D_{mean}, D_{min}, D_2, D_5, D_{50}, D_{95}, D_{98}, D_{max}; and iii)clinical goal pass-rate ($Passrate_{Dx\%}$(CTV)), $D_{x\%}$ represents D_{95}, D_{max}. The pass-rate refers to the number of scenarios that fulfil the clinical goal. Regarding the OARs-based DVHB, clinical goal dose and pass-rate were calculated. For the robust-dose-metrics ($Dosemetrics_{robust}$), all dose metrics in Table 1 were calculated for each scenario. The robust-dose-metrics and the CTV coverage were reported as minimum, mean, and maximum.

2.3. Statistical analysis method

The Shapiro-Wilk test assessed data normality ($\alpha = 0.05$) [15,26]. The Mann-Whitney *U* test was used to determine the significance of median differences between subsets of metrics ($p < 0.05$) [27].

Table 2
Selected metrics, counts from the multivariable Spearman correlation, normality check, skewness value, and corresponding SPC results. The Shapiro-Wilk test ($\alpha = 0.05$) checks for normality, and results are indicated as 'pass' for p-values greater than 0.05 and 'fail' for those smaller than 0.05. Calculated distribution skewness is shown as a negative value for left-skewed and a positive value for right-skewed distributions. The calculated CL (centre line), UCL (upper control limit), and LCL (lower control limit) are provided according to Eqs. (2.1) and (2.2) for the baseline plans. * marked the negative limits set to 0.

Variable name	Number of counts	Normality check	Data Skewness	CL – centre line	LCL – lower control limit	UCL – upper control limit
	Dose metrics					
R_{95} [u.a.]	11	fail	−0.60	0.98	0.95	0.99
$CI_{paddick}$ [u.a.]	6	pass	0.01	0.51	0.22	0.80
$HI_{95/5}$ [u.a.]	12	pass	0.37	0.05	0.01	0.10
DGI [u.a.]	8	pass	0.40	0.09	0.02	0.17
PQI [u.a.]	5	fail	4.40	5.44	0*	106.20
	Plan complexity					
PMU [MU]	9	pass	0.24	498.40	91.23	905.56
PA [cm^2]	6	fail	0.55	60.23	20.06	124.52
PI [u.a.]	10	pass	0.28	1.21	0.02	2.40
PM [u.a.]	8	fail	−1.07	0.71	0.23	0.92
MCS [u.a.]	7	fail	0.75	0.32	0.12	0.68
LTMCS [u.a.]	7	fail	0.89	0.14	0.04	0.35
C/A [1/cm]	5	pass	0.11	0.47	0.13	0.81
SAS_{15mm} [u.a.]	7	pass	0.07	0.28	0*	0.58
$SA/CP_{15mm\times15mm}$ [u.a.]	6	pass	−0.04	0.36	0.00	0.73
	Plan robustness (target-based)					
DVHB(CTV) [u.a.]	4	fail	0.91	0.01	0.00	0.03
$D95_{mean}$(CTV) [u.a.]	16	fail	−1.27	0.98	0.95	0.99
EVH [u.a.]	18	fail	0.82	3.49	1.52	7.32
	Plan robustness (OAR-based)					
DVHB(right lung) [u.a.]	1	fail	0.81	0.04	0*	0.13
$Dmean_{max}$(right lung) [u.a.]	1	fail	0.48	0.29	0*	0.83
DVHB(left lung) [u.a.]	6	fail	1.51	0.03	0*	0.13
$Dmean_{max}$(left lung) [u.a.]	9	fail	0.83	0.28	0*	0.98
DVHB(heart) [u.a.]	1	fail	0.42	0.05	0.00	0.14
$Dmax_{max}$(heart) [u.a.]	15	fail	−2.62	1.05	0.23	1.33
DVHB(spinal cord) [u.a.]	7	fail	0.68	0.02	0*	0.07
$Dmax_{max}$(spinal cord) [u.a.]	8	fail	−0.51	0.69	0.04	1.11

Spearman's correlation coefficient was used to assess correlations between metrics ($p < 0.05$) [28]. Absolute coefficients were categorised as very strong/high (0.90–1.00), strong/high (0.70–0.89), moderate (0.50–0.69), weak/low (0.30–0.49), and no/little (0.00–0.29) [29]. Corrections for multiple testing (e.g. Bonferroni, Benjamini-Hochberg) were not applied.

2.4. Statistical process control (SPC) method

Eq. (2.1) was used for normally distributed data [15]. For non-normal data with skewed distribution, the skewness correction method (Eq. (2.2)) was used [16,18].

$$UCL \Big/ LCL = CL_{mean} \pm 3 \bullet \frac{\overline{MR}}{d_2 \bullet \sqrt{n}} = CL \pm 3 \bullet \frac{\overline{MR}}{1.128} \tag{2.1}$$

$$UCL/LCL = CL_{median} \pm \left(3 + \frac{\frac{4}{3}k_3}{1 + 0.2k_3^2} \right) \bullet \frac{\overline{MR}}{d_2^{sc}} \tag{2.2}$$

where, CL: centre line; UCL/LCL: upper/lower control limit; CL_{mean}/CL_{median} : mean/median of value ($\bar{x}$); $\overline{MR} = |x_i - x_{i-1}|$) is the average of the moving range, the difference between two individual measurements (x_i) for a subgroup (n); The bias correction factors ($d_2 = 1.128$) for $n = 2$ can be used for $n = 1$ [21]; d_2^{sc} is the control chart constant according to k_3(skewness), taken from [16], which were interpolated if k_3 is not given and extrapolated for $k_3 < 4$. CL, LCL, and UCL were calculated from the baseline plans and are continuously updated as new plans are added.

2.5. Plan quality assessment workflow

Fig. 1a) illustrates the baseline creation, with the selection of the metrics using the baseline plans and corresponding calculation of SPC limits. Metrics listed in Table 1 were calculated and saved in a database. The metrics were selected based on the following criteria and steps: i) Statistically significant Spearman coefficients (>0.3, indicating weak correlation) were counted for each metric, ignoring correlations within the same category. E.g. for $HI_{95/5}$, the correlation with other dose metrics are not considered, and for DVHB(heart), the correlation with other OAR-based robustness metrics are excluded; ii)Metrics with less than 5 counts were excluded; iii)For metrics with equivalent information, the metric with the highest correlation count was selected. Examples of metrics that are considered to provide equivalent information include dose metrics calculated using different equations (e.g.Gradient Index (GI) and Dose Gradient Index (DGI) or Paddick Conformity Index ($CI_{paddick}$) and RTOG Conformity Index (CI_{RTOG})), plan complexity metrics calculated using various criteria (SAA or SA/CP), and robustness metrics where results are recorded as a minimum, mean and maximum ($Dosemetrics_{robust}$, $D_{x\%}$(CTV), $D_{x\%}$(OAR), $Passrate_{Dx\%}$(OAR), or $Passrate_{Dx\%}$(CTV)); iv) For equal counts, the metric with the highest coefficient was selected. SPC limits were calculated for the selected metrics.

Fig. 1b shows an example of the framework application, where the pre-calculated control limits are used to evaluate a plan. If the plan falls outside the SPC limits, the user can re-optimise or proceed if it is acceptable.

2.6. Prediction of plan robustness using dose and complexity metrics

The forward selection method was used to determine the best

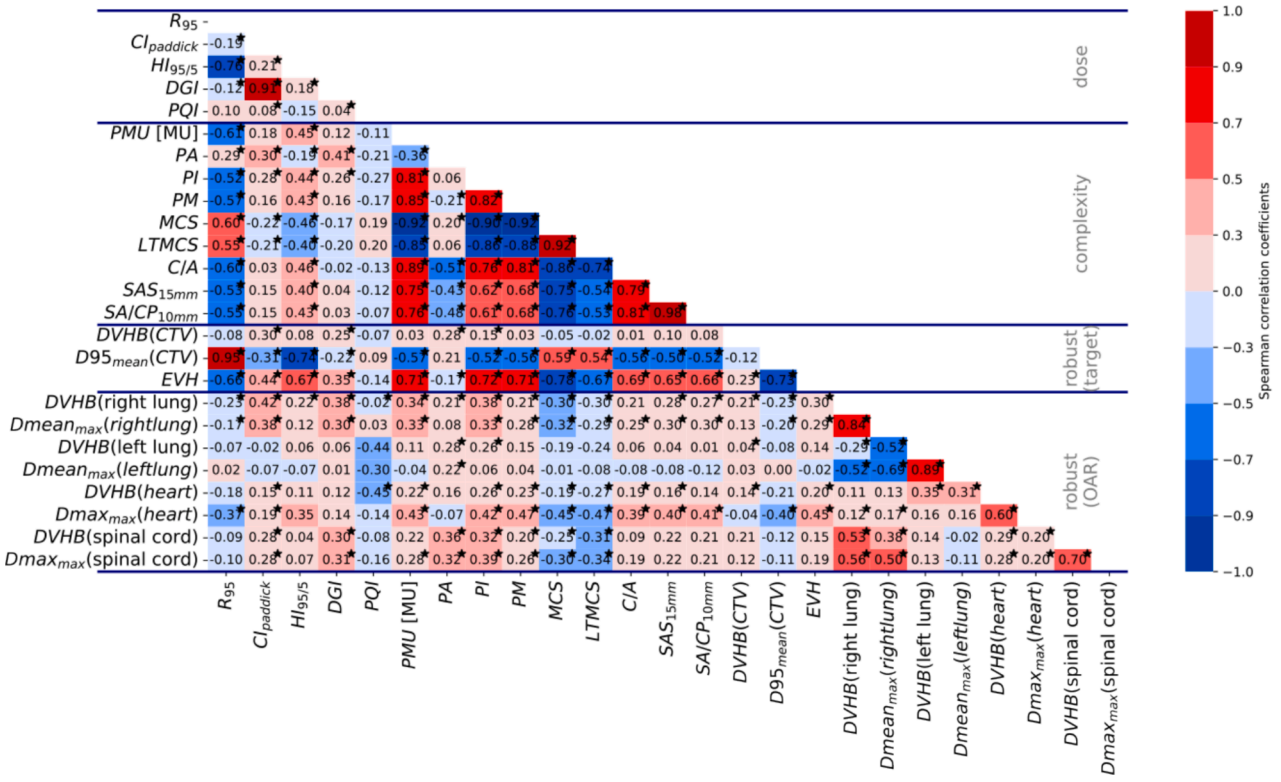

Fig. 2. Multivariable Spearman correlation analysis of the selected dose, plan complexity, and plan robustness metrics based on the CTV and the OAR. Each cell in the plot shows the correlation between the corresponding row and column variables. An asterisk (*) indicates statistically significant correlation coefficients (α = 0.05). Blue indicates a negative correlation, while red indicates a positive correlation. The lightest blue/red colours represent coefficients with no correlation, and the colour intensity increases according to the categorisation of the coefficients: very strong/high (0.90–1.00), strong/high (0.70–0.89), moderate (0.50–0.69), weak/low (0.30–0.49), and no/little (0.00–0.29). The number of coefficients >0.3 in Figure is lower than those listed in Table 2 because some significant coefficients were counted but not selected. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

statistically significant metrics for prediction [30]. Only dose- and complexity-metrics with at least a weak significant correlation with the target-based robustness metrics were used as input variables for the forward selection method. The prediction model used single or multi-variable linear regression (ordinary least squares), depending on the number of selected variables. The 103 baseline plans were randomly divided into training (80 % = 82 plans) and validation (20 % = 21 plans) datasets. The prediction models were validated with the validation datasets calculating the Residual Sum of Squares (RSS), coefficient of determination (R^2), Adjusted R^2, Mean Squared Error (MSE), and Akaike Information Criterion (AIC) [30]. Goodness of fit was classified as high or good ($R^2 \geq 0.7$), moderate ($0.7 > R^2 \geq 0.5$), or poor ($R^2 < 0.5$) [31]. High RSS and MSE indicate low prediction accuracy [30,32]. The 95 % confidence interval (range of the true means) and prediction interval (range of new observations) were calculated for the validation dataset [30,33]. The percentage of this data within these intervals and the difference between predicted and calculated values were assessed.

2.7. Application of framework and prediction model using test plans

2.7.1. General framework

As illustrated in Fig. 1b, the selected metrics were calculated and evaluated for compliance with SPC limits for each test plan. Plans outside the SPC limits were not reoptimised.

2.7.2. Prediction model

As described in Section 2.6, the test plans' robustness metrics were predicted from dose and complexity metrics. Confidence and prediction intervals were determined for predicted values, and the percentage of calculated metric values within these intervals was assessed. The calculated and predicted robustness metrics were compared using the Mann-Whitney test.

3. Results

3.1. Selection of metrics using multivariate correlation

From the initial 122 metrics, 25 were selected for further use (Table 2 and Fig. 2). In summary, out of the 122 metrics, 45 were removed due to counts of less than five, and 52 provided equivalent information to other metrics and had fewer counts or lower coefficients. All robust-dose-metrics were deselected because they yield similar results as the dose metrics calculated with the nominal plan. Despite having fewer than five significant correlations, the following metrics were included: DVHB (CTV), DVHB(left lung), $Dmean_{max}$(left lung), DVHB(heart), and $Dmean_{max}$(left lung). The DVHB metrics were selected to evaluate the robustness of all relevant volumes, and $Dmean_{max}$(left lung) complements $Dmean_{max}$(right lung). Although $Dmax_{min}$(spinal cord) had the highest count (12), $Dmax_{max}$(spinal cord) with eight counts was selected as it reflects the worst-case scenario for spinal cord dose.

Fig. 2 displays the correlations among the final selected metrics. Dose coverage (R_{95}) and Homogeneity Index ($HI_{95/5}$) have several weak to moderate correlations with complexity and target-based robustness metrics. The plan complexity metrics (Plan normalised MU (PMU), Plan averaged beam Irregularity (PI), Plan averaged beam Modulation (PM), Modulation Complexity Score (MCS), Leaf Travel MCS (LTMCS), Circumference/Area (C/A), SAS_{15mm}, and $SA/CP_{15mm\times15mm}$) show moderate to strong correlations with EVH and $D95_{mean}$(CTV).

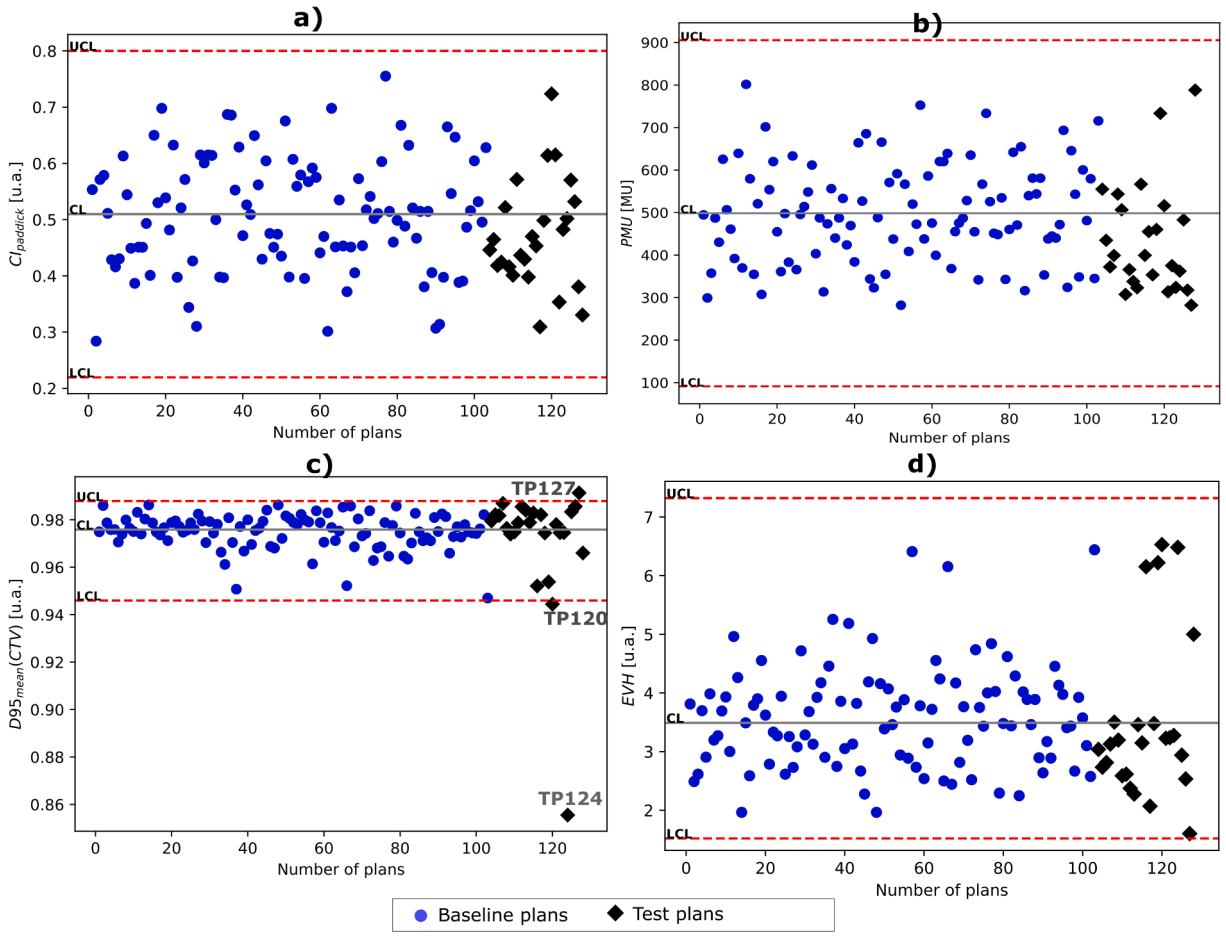

Fig. 3. Statistical Process Control (SPC) chart of four calculated metrics. Blue circles represent baseline treatment plans, the filled black diamonds indicate test plans, red dashed lines mark the upper and lower control limits (UCL/LCL), and the black solid line indicates the centre line (CL). The control chart for the dose metrics "$CI_{paddick}$" (a) and complexity metrics "PMU" (b) are in control and are normally distributed. The left-skewed robustness metric "$D95_{mean}$(CTV)" (c) shows plans TP120, TP124, and TP127 outside the limits. The right-skewed robustness metric "EVH" (d) values are within the limits. TP116, TP119, TP120, and TP124 are the calculated test plan metrics between the EVH values 6–7. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2. Statistical process control (SPC)

The Shapiro-Wilk test shows that 9 out of 25 metrics follow a normal distribution (Table 2). Overall, 99.5 % of the metrics were within the SPC limits displayed in Table 2. Outliers included 11 metrics (1 DGI, 4 PQI, 1 DVHB(CTV), 1 DVHB(heart), and 4 $Dmax_{max}$(heart)). SPC limits may be broad and occasionally have negative values, e.g., OARs DVH-Band(DVHB), $Dmean_{max}$(right lung), and $Dmean_{max}$(left lung)). The negative SPC limits were set to zero (Table 2).

3.3. Prediction of plan robustness using dose and plan complexity metrics

The input variables for the forward selection are taken from Fig. 2. The input variables were: a) for $\underline{D95_{mean}(CTV)}$ – R_{95}, $CI_{Paddick}$, $HI_{95/5}$, PMU, PI, PM, MCS, LTMCS, C/A, SAS_{15mm}, and $SA/CP_{15mm\times15mm}$; b) for $\underline{EVH}$ – the same as for $D95_{mean}$(CTV) plus DGI, and c) for $\underline{DVHB(CTV)}$ – $CI_{Paddick}$ and DGI. The selected metrics for prediction were a) R_{95} and $CI_{Paddick}$ for $D95_{mean}$(CTV); b) LTMCS, $CI_{Paddick}$, and $SA/CP_{15mm\times15mm}$ for EVH; and c) $CI_{Paddick}$ for DVHB(CTV).

The validation of the prediction models indicated that $D95_{mean}$(CTV) (RSS = 8.1×10^{-5}, $R^2 = 0.883$, $R^2_{adjusted}$=0.880, MSE = 3.84×10^{-6}) was a good model (Fig. 4a), EVH (RSS = 4.13, $R^2 = 0.53$,$R^2_{adjusted} = 0.52$, MSE = 0.20) a moderate (Fig. 4b) and DVHB(CTV) ($R^2 = 0.065$, $R^2_{adjusted}$=0.053) a poor model. Therefore, DVHB(CTV) won't be predicted. The respective equations for these models are:

$$D95_{mean}(CTV) = -0.04 + 1.04 \bullet R_{95} - 0.01 \bullet CI_{Paddick}$$

$$EVH = 1.70 + 2.52 \bullet CI_{paddick} - 4.88 \bullet LTMCS + 2.84 \bullet SA/CP_{15mm\times15mm}$$

The Mann-Whitney p-values for $D95_{mean}$(CTV) (*p-value* = 0.8) and EVH (*p-value* = 0.9) indicated no significant difference between predicted and calculated metrics of the validation dataset. The corresponding percentages of validation data within the confidence/prediction intervals were 28.6 %/100 % for $D95_{mean}$(CTV) and 52.4 %/100 % for EVH, respectively.

3.4. Application of framework and prediction model using test plans

3.4.1. General framework

The Mann-Whitney test comparing baseline plans with test plans showed significant differences in six metrics: PMU, PI, PM, MCS, LTMCS, and C/A. The analysis showed that 98.9 % of the calculated test plan metrics fall within the control limits. The outliers were R_{95} for test plan TP124, $D95_{mean}$(CTV) for TP119, TP124, and TP127 (Fig. 3c), and $Dmax_{max}$(heart) for TP124. Test plan TP124 exceeded the limits for several metrics. Analysis revealed that the patient had a pacemaker positioned at the same level as the target volume, resulting in a suboptimal plan that was not robust yet clinically desirable. Fig. 3a–d show

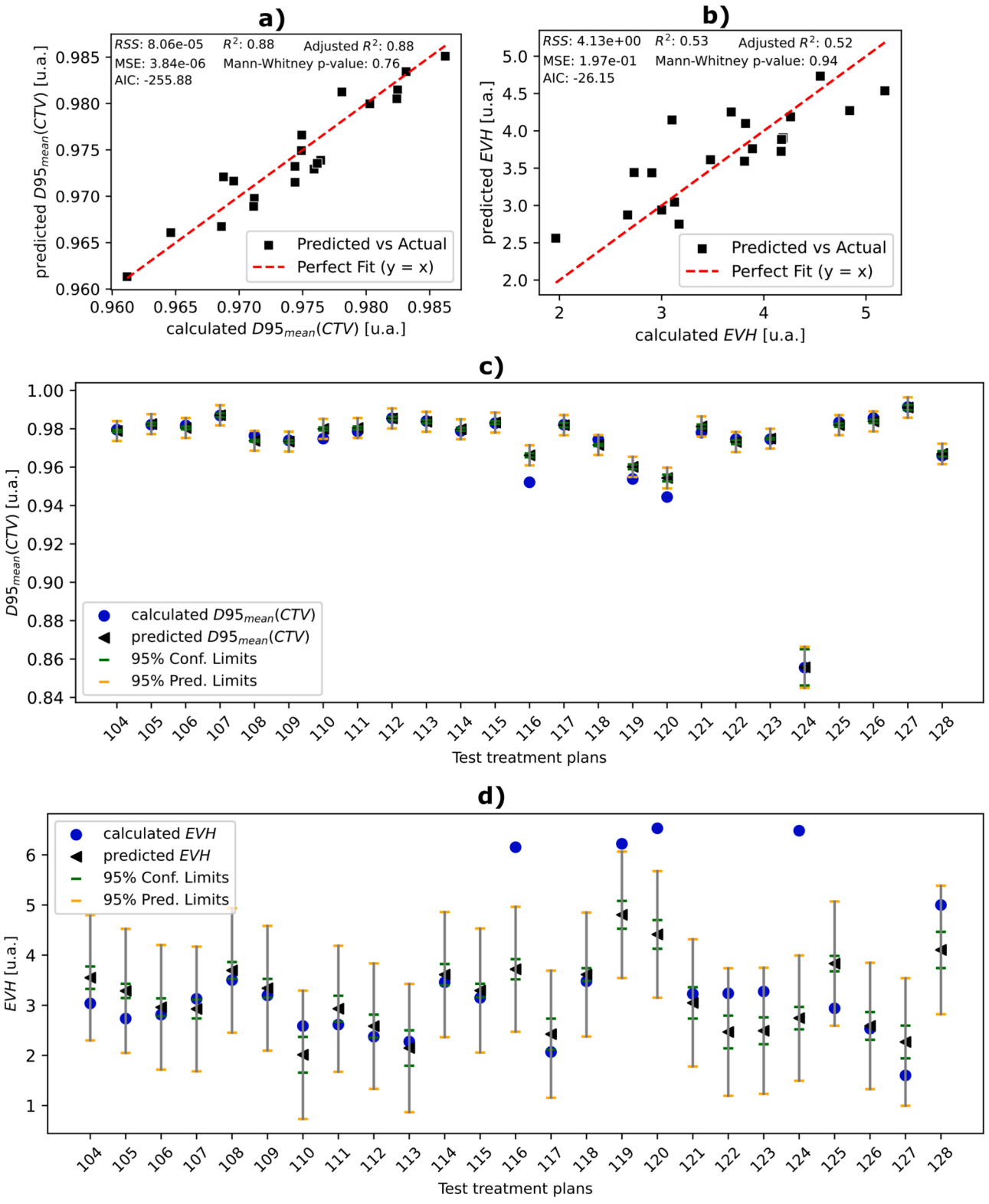

Fig. 4. The black squares show the comparison between the calculated and predicted values from the validation dataset (20 % of the randomly split baseline dataset) for $D95_{mean}$ (CTV) (a) and EVH (b), used for model validation. The red fit line in (a) and (b) represents a perfect fit, where calculated equals predicted. Important metrics to validate the prediction models are stated in the plot. The calculated robustness metrics (blue circles) and predicted robustness metrics (black triangles), along with confidence and prediction intervals, are shown for the test plans. The robustness metrics $D95_{mean}$(CTV) (c) and EVH (d) are presented. A low EVH indicates a more robust plan, and $D95_{mean}$(CTV) should be high. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

some examples of SPC plots. Although TP116, TP119, TP120, and TP124 in Fig. 3d were within the EVH metrics limit, their EVH values are higher than the other test plans.

3.4.2. Prediction model

The calculated EVH was within the confidence/prediction limits of 28 % / 84 % (Fig. 4c), and the $D95_{mean}$(CTV) was within 44 % / 88 % (Fig. 4d), respectively. The calculated metrics of plan TP116, TP119, and TP120 are outside the prediction interval for EVH and $D95_{mean}$(CTV) and TP124 for $D95_{mean}$(CTV). The Mann-Whitney test for EVH (*p-value* = 0.9) and $D95_{mean}$(CTV) (*p-value* = 1) indicated no statistically significant differences between the calculated and predicted values.

4. Discussion

Plan quality assessment included dose-, plan complexity-, and plan robustness metrics [3]. From an initial 122 metrics, 25 were selected based on multivariate Spearman correlations, including dose, complexity, and robustness metrics. Yaparpalvi et al. [34] used a similar method to evaluate dose metrics for SBRT lung plans but did not include complexity and robustness metrics. Table 2 lists the recommended metrics based on our analysis. R_{95} and $HI_{95/5}$ were prominent for dose metrics, while $D95_{mean}$(CTV) and EVH were prominent for robustness metrics. Plan complexity metrics showed similar results, except for Plan Averaged Beam Area (PA).

SPC was primarily used to determine the limits of the metrics. Most metrics for baseline and test plans were within SPC limits, as expected for clinically accepted plans. Some limits are broad and could be tightened to detect more suboptimal plans. Before clinical implementation, limits should be adjusted, e.g., setting negative limits to zero or using an SPC range smaller than $\pm 3\sigma$, to ensure clinically meaningful thresholds and avoid unnecessary workload increases. Alternatively, the quantile method could be used to set metric limits [16].

A robust evaluation may not always be available, so a multiple-variable regression model was fitted for EVH and $D95_{mean}$(CTV). The performance of the EVH model was moderate, and the one for the $D95_{mean}$(CTV) model was good. This prediction model aimed not to provide exact values but to help planners identify more robust plans even without performing the robust evaluation. The observed similarity in trends between the calculated and predicted values supports this aim. The calculated robustness metrics for test plans T116, T119, and T120 were not within the prediction interval for $D95_{mean}$(CTV) and EVH, and their values can be identified as outliers, mainly observed in Fig. 3d. The robustness metrics can always be calculated if robust evaluation is available.

Significant differences were found between baseline and test plan distributions for some complexity metrics, likely due to optimisation or workflow changes. The updated TPS version and machine should not impact complexity, as plans were recalculated and not re-optimized. No visual changes were observed between the test plans' original and recalculated dose distributions.

The robust evaluation and optimisation only included patient position uncertainty. The impact of setup errors on dose distribution uncertainty is reduced by increasing the number of fractions [35,36], and therefore the results may be overestimated. However, we acknowledge that other important uncertainties, such as density and organ motion, could affect the accuracy of dose delivery, and we plan to include these uncertainties in our method in the future. Existing co-dependencies between metrics weren't evaluated because conventional statistical methods were used; perhaps a deep learning algorithm could determine co-dependencies.

In conclusion, this study represents the first step in developing a tool to assess plan quality through dose, plan complexity, and robustness metrics within an all-in-one framework that can be applied during the treatment planning or evaluation process. The introduced method could be integrated into TPS through scripting and an intuitive GUI. Although this framework was established for normal fractionated lung cancer cases, it can easily be applied to different radiation sites/types and to determine the metrics' unknown limits. It allows the comparison of multiple plans for the same or other patients using SPC. It can detect suboptimal plans and variations in the planning process (e.g., new algorithms, new margin concepts, different optimisation strategies). This enables planners to systematise and standardise the treatment planning process and the decision-making during plan quality assessment.

CRediT authorship contribution statement

Tina Orovwighose: Conceptualization, Methodology, Software, Validation, Investigation, Formal analysis, Writing – original draft, Writing – review & editing, Data curation, Visualization, Project administration, Supervision. **Bernhard Rhein:** Conceptualization, Methodology, Resources, Writing – review & editing. **Oliver Schramm:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision. **Oliver Jäkel:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Project administration. **Vania Batista:** Conceptualization, Methodology, Formal analysis, Writing – review & editing, Validation, Supervision, Project administration.

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.phro.2024.100685.

References

[1] Cheng C-W, Das IJ. Treatment plan evaluation using dose–volume histogram (DVH) and spatial dose–volume histogram (zDVH). Int J Radiat Oncol 1999;43:1143–50. https://doi.org/10.1016/S0360-3016(98)00492-1.
[2] Kaplan LP, Korreman SS. A systematically compiled set of quantitative metrics to describe spatial characteristics of radiotherapy dose distributions and aid in treatment planning. Phys Med 2021;90:164–75. https://doi.org/10.1016/J.EJMP.2021.09.014.
[3] Hernandez V, Hansen CR, Widesott L, Bäck A, Canters R, Fusella M, 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.
[4] Chiavassa S, Bessieres I, Edouard M, Mathot M, Moignier A. Complexity metrics for IMRT and VMAT plans: a review of current literature and applications. Br J Radiol 2019;92:20190270. https://doi.org/10.1259/bjr.20190270.
[5] Kamperis E, Kodona C, Hatziioannou K, Giannouzakos V. Complexity in radiation therapy: it's complicated. Int J Radiat Oncol 2020;106:182–4. https://doi.org/10.1016/j.ijrobp.2019.09.003.
[6] Vanherk M. Errors and margins in radiotherapy. Semin Radiat Oncol 2004;14:52–64. https://doi.org/10.1053/j.semradonc.2003.10.003.
[7] Unkelbach J, Alber M, Bangert M, Bokrantz R, Chan TCY, Deasy JO, et al. Robust radiotherapy planning. Phys Med Biol 2018;63:22TR02. https://doi.org/10.1088/1361-6560/aae659.
[8] Wagenaar D, Kierkels RGJ, Free J, Langendijk JA, Both S, Korevaar EW. Composite minimax robust optimization of VMAT improves target coverage and reduces non-target dose in head and neck cancer patients. Radiother Oncol 2019;136:71–7. https://doi.org/10.1016/j.radonc.2019.03.019.
[9] Fredriksson A. A characterization of robust radiation therapy treatment planning methods-from expected value to worst case optimization. Med Phys 2012;39:5169–81. https://doi.org/10.1118/1.4737113.
[10] McGowan SE, Albertini F, Thomas SJ, Lomax AJ. Defining robustness protocols: a method to include and evaluate robustness in clinical plans. Phys Med Biol 2015;60:2671–84. https://doi.org/10.1088/0031-9155/60/7/2671.
[11] Fredriksson A, Forsgren A, Hårdemark B. Minimax optimization for handling range and setup uncertainties in proton therapy. Med Phys 2011;38:1672–84. https://doi.org/10.1118/1.3556559.
[12] Korevaar EW, Habraken SJM, Scandurra D, Kierkels RGJ, Unipan M, Eenink MGC, et al. Practical robustness evaluation in radiotherapy – a photon and proton-proof alternative to PTV-based plan evaluation. Radiother Oncol 2019;141:267–74. https://doi.org/10.1016/j.radonc.2019.08.005.

[13] Kennedy AC, Douglass MJJ, Santos AMC. Being certain about uncertainties: a robust evaluation method for high-dose-rate prostate brachytherapy treatment plans including the combination of uncertainties. Phys Eng Sci Med 2023;46:1115–30. https://doi.org/10.1007/s13246-023-01279-8.
[14] Binny D, Lancaster CM, Kairn T, Trapp JV, Crowe SB. Radiotherapy quality assurance using statistical process control. IFMBE Proc 2019;68:437–42. https://doi.org/10.1007/978-981-10-9023-3_78.
[15] Montgomery DC. Introduction to statistical quality control. 7th ed. John Wiley & Sons, Inc; 2013.
[16] Xiao Q, Bai L, Li G, Zhang X, Li Z, Duan L, et al. A robust approach to establish tolerance limits for the gamma passing rate-based patient-specific quality assurance using the heuristic control charts. Med Phys 2022;49:1312–30. https://doi.org/10.1002/mp.15346.
[17] Xiao Q, Li G. Application and challenges of statistical process control in radiation therapy quality assurance. Int J Radiat Oncol 2024;118:295–305. https://doi.org/10.1016/j.ijrobp.2023.08.020.
[18] Chan LK, Cui HJ. Skewness correction $\bar{X}$ and R charts for skewed distributions. Nav Res Logist 2003;50:555–73. https://doi.org/10.1002/nav.10077.
[19] Breen SL, Moseley DJ, Zhang B, Sharpe MB. Statistical process control for IMRT dosimetric verification; statistical process control for IMRT dosimetric verification. Med Phys 2008;35:4417–25. https://doi.org/10.1118/1.2975144.
[20] Xiao Q, Bai S, Li G, Yang K, Bai L, Li Z, et al. Statistical process control and process capability analysis for non-normal volumetric modulated arc therapy patient-specific quality assurance processes. Med Phys 2020;47:4694–702. https://doi.org/10.1002/mp.14399.
[21] Pawlicki T, Whitaker M, Boyer AL. Statistical process control for radiotherapy quality assurance. Med Phys 2005;32:2777–86. https://doi.org/10.1118/1.2001209.
[22] Nordström F, Wetterstedt S, Johnsson S, Ceberg C, Bäck SÅJ. Control chart analysis of data from a multicenter monitor unit verification study. Radiother Oncol 2012;102:364–70. https://doi.org/10.1016/j.radonc.2011.11.016.
[23] RaySearch Laboratories AB. RayStation 11B-Reference Manual. Work Main Version A697, Skribenta Version 54033; 2021.
[24] Bentzen SM, Constine LS, Deasy JO, Eisbruch A, Jackson A, Marks LB, et al. Quantitative analyses of normal tissue effects in the clinic (QUANTEC): an introduction to the scientific issues. Int J Radiat Oncol 2010;76:S3–9. https://doi.org/10.1016/j.ijrobp.2009.09.040.
[25] Jornet N, Carrasco P, Beltrán M, Calvo JF, Escudé L, Hernández V, et al. Multicentre validation of IMRT pre-treatment verification: comparison of in-house and external audit. Radiother Oncol 2014;112:381–8. https://doi.org/10.1016/j.radonc.2014.06.016.
[26] Schlegel W, Bille J. In: Medizinische Physik. Berlin, Heidelberg: Springer Berlin Heidelberg; 2018. https://doi.org/10.1007/978-3-662-54801-1.
[27] May L, Hardcastle N, Hernandez V, Saez J, Rosenfeld A, Poder J. Multi-institutional investigation into the robustness of intra-cranial multi-target stereotactic radiosurgery plans to delivery errors. Med Phys 2024;51:910–21. https://doi.org/10.1002/mp.16907.
[28] 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.
[29] Asuero AG, Sayago A, González AG. The correlation coefficient: an overview. Crit Rev Anal Chem 2006;36:41–59. https://doi.org/10.1080/10408340500526766.
[30] James G, Witten D, Hastie T, Tibshirani R, Taylor J. In: An introduction to statistical learning. Cham: Springer International Publishing; 2023. https://doi.org/10.1007/978-3-031-38747-0.
[31] Wang H, Zhou Y, Gan W, Chen H, Huang Y, Duan Y, et al. Regression models for predicting physical and EQD2 plan parameters of two methods of hybrid planning for stage III NSCLC. Radiat Oncol 2021;16:119. https://doi.org/10.1186/s13014-021-01848-9.
[32] Grégoire G. Multiple linear regression. EAS Publ Ser 2014;66:45–72. https://doi.org/10.1051/eas/1466005.
[33] Douglas C, Montgomery G, Geoffrey Vining EAP. Introduction to Linear Regression Analysis. 5th ed. Wiley Series in Probability and Statistics; 2012.
[34] Yaparpalvi R, Garg MK, Shen J, Bodner WR, Mynampati DK, Gafar A, et al. Evaluating which plan quality metrics are appropriate for use in lung SBRT. Br J Radiol 2018;91:20170393. https://doi.org/10.1259/bjr.20170393.
[35] Lowe M, Albertini F, Aitkenhead A, Lomax AJ, MacKay RI. Incorporating the effect of fractionation in the evaluation of proton plan robustness to setup errors. Phys Med Biol 2016;61:413–29. https://doi.org/10.1088/0031-9155/61/1/413.
[36] Lowe M, Aitkenhead A, Albertini F, Lomax AJ, MacKay RI. A robust optimisation approach accounting for the effect of fractionation on setup uncertainties. Phys Med Biol 2017;62:8178–96. https://doi.org/10.1088/1361-6560/aa8c58.
[37] Jin X, Yi J, Zhou Y, Yan H, Han C, Xie C. A new plan quality index for nasopharyngeal cancer SIB IMRT. Phys Med 2014;30:122–7. https://doi.org/10.1016/j.ejmp.2013.03.007.
[38] Feuvret L, Noël G, Mazeron J-J, Bey P. Conformity index: a review. Int J Radiat Oncol 2006;64:333–42. https://doi.org/10.1016/j.ijrobp.2005.09.028.
[39] Zhang X, Rong Y, Morrill S, Fang J, Narayanasamy G, Galhardo E, et al. Robust optimization in lung treatment plans accounting for geometric uncertainty. J Appl Clin Med Phys 2018;19:19–26. https://doi.org/10.1002/acm2.12291.
[40] Paddick I, Lippitz B. A simple dose gradient measurement tool to complement the conformity index. J Neurosurg 2006;105:194–201. https://doi.org/10.3171/sup.2006.105.7.194.
[41] Kataria T, Sharma K, Subramani V, Karrthick KP, Bisht SS, et al. Homogeneity index: an objective tool for assessment of conformal radiation treatments. J Med Phys 2012;37:207–20. https://doi.org/10.4103/0971-6203.103606.
[42] Akpati H, Kim C, Kim B, Park T, Meek A. Unified dosimetry index (UDI): a figure of merit for ranking treatment plans. J Appl Clin Med Phys 2008;9:99–108. https://doi.org/10.1120/jacmp.v9i3.2803.
[43] Du W, Cho SH, Zhang X, Hoffman KE, Kudchadker RJ. Quantification of beam complexity in intensity-modulated radiation therapy treatment plans. Med Phys 2014;41:021716. https://doi.org/10.1118/1.4861821.
[44] Crowe SB, Kairn T, Kenny J, Knight RT, Hill B, Langton CM, et al. Treatment plan complexity metrics for predicting IMRT pre-treatment quality assurance results. Australas Phys Eng Sci Med 2014;37:475–82. https://doi.org/10.1007/s13246-014-0274-9.
[45] Park S-Y, Kim J, Chun M, Ahn H, Park JM. Assessment of the modulation degrees of intensity-modulated radiation therapy plans. Radiat Oncol 2018;13:244. https://doi.org/10.1186/s13014-018-1193-9.
[46] 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;40:071718. https://doi.org/10.1118/1.4810969.
[47] Antoine M, Ralite F, Soustiel C, Marsac T, Sargos P, Cugny A, et al. 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.
[48] Götstedt J, Karlsson Hauer A, Bäck A. Development and evaluation of aperture-based complexity metrics using film and EPID measurements of static MLC openings. Med Phys 2015;42:3911–21. https://doi.org/10.1118/1.4921733.
[49] Shen L, Chen S, Zhu X, Han C, Zheng X, Deng Z, et al. Multidimensional correlation among plan complexity, quality and deliverability parameters for volumetric-modulated arc therapy using canonical correlation analysis. J Radiat Res 2018;59:207–15. https://doi.org/10.1093/jrr/rrx100.
[50] Albertini F, Hug EB, Lomax AJ. Is it necessary to plan with safety margins for actively scanned proton therapy? Phys Med Biol 2011;56:4399–413. https://doi.org/10.1088/0031-9155/56/14/011.
[51] Liu W, Frank SJ, Li X, Li Y, Park PC, Dong L, et al. Effectiveness of robust optimization in intensity-modulated proton therapy planning for head and neck cancers. Med Phys 2013;40:051711. https://doi.org/10.1118/1.4801899.
[52] Sterpin E, Rivas ST, Van Den Heuvel F, George B, Lee JA, Souris K. Development of robustness evaluation strategies for enabling statistically consistent reporting. Phys Med Biol 2021;66:045002. https://doi.org/10.1088/1361-6560/abd22f.
[53] Shang H, Pu Y, Wang W, Dai Z, Jin F. Evaluation of plan quality and robustness of IMPT and helical IMRT for cervical cancer. Radiat Oncol 2020;15:34. https://doi.org/10.1186/s13014-020-1483-x.