On the new metrics for IMRT QA verification

Alejandro Garcia-Romero,[a] Araceli Hernandez-Vitoria, Esther Millan-Cebrian, Veronica Alba-Escorihuela, Sonia Serrano-Zabaleta, and Pablo Ortega-Pardina
Servicio de Fisica y Proteccion Radiologica, Hospital Clinico Universitario "Lozano Blesa" de Zaragoza, Avenida San Juan Bosco 15, Zaragoza E-50009, Spain

(Received 27 April 2016; revised 27 September 2016; accepted for publication 30 September 2016; published 21 October 2016)

Purpose: The aim of this work is to search for new metrics that could give more reliable acceptance/rejection criteria on the IMRT verification process and to offer solutions to the discrepancies found among different conventional metrics. Therefore, besides conventional metrics, new ones are proposed and evaluated with new tools to find correlations among them. These new metrics are based on the processing of the dose–volume histogram information, evaluating the absorbed dose differences, the dose constraint fulfillment, or modified biomathematical treatment outcome models such as tumor control probability (TCP) and normal tissue complication probability (NTCP). An additional purpose is to establish whether the new metrics yield the same acceptance/rejection plan distribution as the conventional ones.
Methods: Fifty eight treatment plans concerning several patient locations are analyzed. All of them were verified prior to the treatment, using conventional metrics, and retrospectively after the treatment with the new metrics. These new metrics include the definition of three continuous functions, based on dose–volume histograms resulting from measurements evaluated with a reconstructed dose system and also with a Monte Carlo redundant calculation. The 3D gamma function for every volume of interest is also calculated. The information is also processed to obtain ΔTCP or ΔNTCP for the considered volumes of interest. These biomathematical treatment outcome models have been modified to increase their sensitivity to dose changes. A robustness index from a radiobiological point of view is defined to classify plans in robustness against dose changes.
Results: Dose difference metrics can be condensed in a single parameter: the dose difference global function, with an optimal cutoff that can be determined from a receiver operating characteristics (ROC) analysis of the metric. It is not always possible to correlate differences in biomathematical treatment outcome models with dose difference metrics. This is due to the fact that the dose constraint is often far from the dose that has an actual impact on the radiobiological model, and therefore, biomathematical treatment outcome models are insensitive to big dose differences between the verification system and the treatment planning system. As an alternative, the use of modified radiobiological models which provides a better correlation is proposed. In any case, it is better to choose robust plans from a radiobiological point of view. The robustness index defined in this work is a good predictor of the plan rejection probability according to metrics derived from modified radiobiological models. The global 3D gamma-based metric calculated for each plan volume shows a good correlation with the dose difference metrics and presents a good performance in the acceptance/rejection process. Some discrepancies have been found in dose reconstruction depending on the algorithm employed. Significant and unavoidable discrepancies were found between the conventional metrics and the new ones.
Conclusions: The dose difference global function and the 3D gamma for each plan volume are good classifiers regarding dose difference metrics. ROC analysis is useful to evaluate the predictive power of the new metrics. The correlation between biomathematical treatment outcome models and the dose difference-based metrics is enhanced by using modified TCP and NTCP functions that take into account the dose constraints for each plan. The robustness index is useful to evaluate if a plan is likely to be rejected. Conventional verification should be replaced by the new metrics, which are clinically more relevant. [http://dx.doi.org/10.1118/1.4964796]

Key words: dose reconstruction, IMRT-QA, verification metrics, ROC analysis, plan robustness

1. INTRODUCTION

Prior to its delivery to a patient, any IMRT treatment designed and calculated by a treatment planning system (TPS) must be verified with an alternative system. To that end, calculation-based or measurement-based approaches may be employed. From the beginning of IMRT, a pretreatment approach has been used in most of the cases, comparing measured dose in selected planes or points with TPS predicted dose. Dedicated phantoms have been used for this purpose, as well as several

detectors such as ionization chamber arrays, diode arrays, ionization chambers, and radiochromic or radiographic films.[1] To decide whether the plan is accepted, several metrics have been traditionally used, being one of the most popular the gamma analysis of a planar dose in a significant plane inside a phantom.[2,3] The applied tolerance has usually been 3%–3 mm. This gamma analysis can be performed for a whole plan containing the sum of all the beams or individually, for a particular beam. In addition, sometimes a composite plan with all the beams perpendicular to the measurement plane has been employed. Keeping in mind that this kind of gamma analysis applies to relative dose (where the normalization point has to be chosen carefully), the verification is usually completed with absolute dose measurements in one or several points with an IMRT-dedicated ionization chamber. The cutoff point to reject a plan is set by analyzing if the percentage of points with gamma >1 is bigger than a certain amount (5% or 10%) or if the absolute dose measured at one point differs more than 3% from the dose predicted by the TPS. In this work, this kind of verification will be referred to as "conventional verification (CONV)." If some cutoff value is exceeded, a subsequent investigation of the cause in order to accept the plan is necessary. From a clinical point of view, this procedure can lead to the so-called false positives (FP) and false negatives (FN) because as it has been stated, there may be cases in which the measured and calculated dose distribution differ in a significant number of points but this has no clinical relevance and, in the opposite way, sometimes there may be cases with few points out of tolerance but located in a clinically relevant region of the patient (i.e., the spinal cord).[4–7] The first circumstance leads to an unnecessary drain of resources (replan and new verification) and the second one could be potentially risky for the patient, due to a clinical constraint out of tolerance. In addition, when the dose distributions are studied in 3D, the 3D gamma criterion has to be used instead, and this may produce different results to those obtained using planar dose distributions.[4,7,8]

In recent years, several systems have incorporated the estimation of dose–volume histograms (DVH) coming from the pretreatment measurements, by means of an appropriate algorithm or a measurement-based dose perturbation.[9–12] Simultaneously, several groups have investigated the use of new metrics for IMRT plan verification with the aim of getting clinically relevant results.[11–16] Since the verification DVH provides information about discrepancies in plan constraints, the result of the plan verification can be considered more relevant. With the differential DVH, it is possible to obtain biomathematical treatment outcome models as equivalent uniform dose (EUD), tumor control probability (TCP), and normal tissue control probability (NTCP). The differences between verification and TPS calculation in these radiobiological functions can also be used as a valid metric to investigate plan acceptance.[15] To date, it remains unclear what metrics are the most appropriate ones, and the cutoffs are not well established or explained.[14] The DVH comparison can be performed and complete reports generated, but it would be very useful to rely on additional metrics to make a decision, since dose discrepancies may not have any clinical impact, even if they are evaluated from the DVH. In this work, new metrics are proposed in an effort to account for this clinical impact. Showing how the different verification metrics are related and how new single parameters summarize the verification outcome can provide confidence to the user. Additionally, the generation of a cutoff value based on statistics is a common procedure in clinical diagnosis, which is a very similar process to the IMRT verification process. This allows the user to effectively discern which plans would benefit the most from further analysis.

In this paper, a "metric" is considered as a function that arises from the comparison between two dose distributions whose output is a continuous or binary variable.

After the IMRT commissioning at our institution, our experience, using different verification methods available over time, has shown that the results of plan acceptance or rejection were contradictory in some cases, even if the same metric was used.

Therefore, our purpose is to compare different metrics, either dose distribution-based, DVH-based, or TCP/NTCP-based, to establish valid cutoffs in plan acceptance and to explore the possibility of a combined analysis, including the use of indicators or a few parameters that can summarize the whole verification process. Some of these metrics proposed here are new.

2. METHODS

2.A. Treatment unit and TPS calculation

In our hospital, step-and-shoot IMRT treatments are carried out with a Siemens Oncor Impression Plus linac equipped with the Optifocus multileaf collimator (MLC). This collimator has 41 pairs of opposed tungsten leaves. The width at isocenter of the 39 inner pairs is 1 cm, while that of the two outer pairs is 0.5 cm, so that the maximum field size in the direction perpendicular to the leaf movement is 40 cm. The leaves are doubly focused, their thickness is 7.6 cm, and they have a straight end. To reduce the interleaf transmission, the MLC has been endowed with a tongue-and-groove structure (TGS): each leaf has a tongue on one side and a groove on the other, except for the pair of central leaves, which have a tongue on either side. In a previous work,[17] we developed a component module for the Optifocus MLC compatible with the BEAMnrc code. The behavior of the Optifocus MLC has been studied previously.[18,19]

The TPS used to design the IMRT plans was PCRT3D 6.02 (Tecnicas Radiofisicas SL). The calculation algorithm was the collapsed cone superposition algorithm (noted hereafter as TPS). The TPS calculations were validated using Monte Carlo (MC) simulations performed with the BEAMnrc code.[20]

2.B. Plan selection

Fifty eight IMRT plans were selected, including several locations: neck, brain, prostate with lymph nodes, and gynecological pelvis. The plan distribution is shown in Table I.

TABLE I. Distribution of plans per location and prescribed dose.

Location	Total number of plans	Prescribed dose (Gy) (number of plans)
Neck	32	70(3), 68.8(2), 67.9(1), 66(25), 60(1)
Brain	10	46(1), 42(2), 60(3), 54(1), 52(1), 50(2)
Prostate	13	76(1), 78(4), 74(2), 56(2), 50(1), 32(1), 28(1), 46(1)
Pelvis	3	50(1), 51(2)

This plan set is a representative of the IMRT treatments delivered at our center from the moment we started to monitor the MLC behavior until the experimental part of this study was concluded. The chosen patients have a variable number of PTVs and different dose prescription at each PTV. In neck treatments, 1–3 PTVs are irradiated with an equivalent scheme of 70, 60, and 50 Gy given in 2 Gy/fraction. These three volumes are not always present depending on the tumor extension. In prostate treatments, there are two volumes: one involving lymph nodes with an equivalent dose prescription of 46 or 50 Gy in 2 Gy/fraction, and the prostate itself, which receives up to 78 Gy in 2 Gy fractions. In brain tumors, the dose prescription is 60 Gy, with an external volume irradiated up to 42 Gy in some cases. For the gynecological pelvis, lymph nodes are irradiated up to 46–50 Gy in 2 Gy fractions and the solid tumor is irradiated up to 60 Gy. The contoured organs at risk are parotid glands, spinal cord, optical nerves, chiasm, brain stem, lenses, larynx, oral cavity, mandible, and thyroid gland for head and neck treatments, and rectum, bladder, femoral heads, kidneys, and spinal cord for pelvic and abdominal regions.

2.C. IMRT conventional verification

All the selected plans were measured initially with a CC01 pin-point ionization chamber (IBA dosimetry) calibrated in dose to water. The chamber was always placed in a significant location inside the IMRT phantom (IBA dosimetry), usually corresponding to a point inside the PTV in the original plan, surrounded by a homogenous dose area (i.e., less than 3% coefficient of variation in a radius of 2 mm). The measured dose was compared to the dose predicted by the TPS, usually the prescription dose per fraction, and a cutoff of ±3% was established as permitted deviation.

In addition to the ion chamber measurements, planar analysis was performed using three different methods over time, due to changes in available equipment. In all cases, the fields were irradiated at their original gantry angles. The first method was a field-by-field analysis comparing TPS prediction with an EPID image transformed to dose using in-house developed software.[21,22] The second method was the complete plan irradiation with a 2D ionization chamber array (MatriXX evolution detector, IBA dosimetry) placed in the Multicube phantom (IBA dosimetry). The third method was the complete plan irradiation with a radiochromic film placed inside the IMRT phantom, transforming the gray level of the scanned film to dose following published methods.[23,24] A gamma analysis with a 3%/3 mm agreement criterion was used when comparing the measured planes to the TPS prediction. The test was passed when the gamma passing rate was higher than 90% (with a 10% dose threshold). In the first method, field-by-field EPID analysis, the normalization dose was fixed according to a significant dose inside the field map (flat area inside the map with a significant dose, typically over half the maximum dose). In the second and third cases, planar MatriXX dose measurement and radiochromic film, the normalization dose was the prescribed dose per fraction because the whole plan was delivered to the measurement system.

2.D. IMRT verification tools

The COMPASS system (IBA dosimetry) is a commercially available solution which uses a 2D ionization chamber matrix (the MatriXX, same as in Sec. 2.C). It is able to reconstruct 3D doses in a phantom or a patient CT using the measured signal to reconstruct a measured fluence. This reconstructed fluence leads to a reconstructed 3D dose by means of a beam model and a collapsed cone superposition algorithm. Further details on this process can be found in Godart *et al.*[10] Hereafter the acronym CMPM refers to this measurement-based calculation. In addition, the system can be used separately from the detector: a dose calculation can be performed without measurement to provide an independent dose calculation, starting from a theoretical fluence constructed using the RTPlan information. The acronym CMPC refers to this RTPlan-based dose calculation matrix. The COMPASS software also contains several tools to analyze the data in the patient anatomy and extract information from the DVHs. The calculated or reconstructed 3D dose matrix can be exported, and therefore it is possible to import these CMPM and CMPC matrices in our TPS. The COMPASS system has to be commissioned as it contains its own calculation engine and experimental beam data must be introduced. This commissioning process was done at our institution prior to the beginning of this study. All the 58 plans mentioned in Sec. 2.C have been measured with COMPASS retrospectively.

2.E. Monte Carlo calculations

To use MC as intended, it has been necessary to adapt the BEAMnrc package to our particular MLC, developing a new component module, and a new software application to perform the patient dose calculation, using the RTPlan file and the patient CT image set.[17] This application generates a dose matrix file that can be imported from the current TPS. Therefore, MC produces a DVH set for every plan that is subsequently compared with the TPS DVHs using the metrics defined below. Our purpose is also to explore the possibility of incorporating the daily linac variability in the MC calculations. To this end, a proportional scaling in

the MC differential DVHs is defined as follows:

$$\Delta D_i = D_i \cdot \frac{\bar{D}_{\mathrm{pPTV,CMPM}} - \bar{D}_{\mathrm{pPTV,CMPC}}}{D_{\mathrm{prescribed}}}, \quad (1)$$

where D_i are the MC differential histogram dose values, and $\bar{D}_{\mathrm{pPTV,CMPM}}$ and $\bar{D}_{\mathrm{pPTV,CMPC}}$ are the average doses to the primary PTV obtained using CMPM and CMPC, respectively. This is an approximation that takes into account different effects including the daily MLC leaves position, which is the most important variability factor in our case, and allows incorporating this effect in the MC redundant calculation. We refer to the whole process as Monte Carlo proportional scaling (MCPS), and the DVHs coming from this process are treated as if they were coming from a measurement-based calculation algorithm.

2.F. Biomathematical treatment outcome models and robustness index (IR)

For TCP and NTCP calculation, the same models proposed in Zhen *et al.*[15] are employed, where TCP is calculated by means of a logit expression and NTCP is calculated using the de Lyman–Kutcher–Burman model. The generalized equivalent uniform dose (gEUD) can be calculated for the PTVs and for the OARs. This is done by means of the following equation:

$$\mathrm{gEUD} = \left(\sum_i v_i D_i^{\frac{1}{n}} \right)^n, \quad (2)$$

where v_i is the fractional volume that receives a dose level of D_i and n is the volume effect parameter. The specific values for the parameters used in this study to calculate gEUD are taken from Zhen *et al.*[15] and can be found in Table II.

m is the slope parameter specific for a tissue, inversely proportional to the slope of the dose–response curve. In the same table are also listed: D_{50}, the dose which yields a tumor control probability of 50%, γ_{50}, the normalized slope of the TCP curve at D_{50}, and TD_{50}, the dose that would cause a 50% chance of normal tissue complication, if the entire organ is uniformly irradiated to that dose. They are employed as in Zhen *et al.*[15] for the TCP and NTCP calculations. Here, the robustness index concept is introduced. It is associated to the plan robustness initially described in Zhen *et al.*[15] as a quantitative way to integrate the concept of plan robustness into the treatment planning process to improve the quality of radiotherapy plans and make them robust to perturbations as far as possible. We suggest a mathematical expression for this index, developed in several steps. First of all, it would be necessary to adapt the dose–response curves to the imposed dose constraints, introducing a shift on them until they are situated in a dose variation sensitive region. The motivation for that is that TCP and NTCP are not sensitive to dose changes if D_{50} and TD_{50} are not similar to the dose constraint established for the evaluated organ. Sometimes the prescribed dose is not the complete treatment dose, for instance, when IMRT is applied as a boost or in a plan that is going to be completed with brachytherapy. In addition, a 50% complication for some organs at risk (i.e., spinal cord and larynx) should not be assumed, which implies the constraint is always situated far away from a significant complication dose. To increase the TCP and NTCP sensitivity, D_{50} and TD_{50} parameters can be adapted to the plan constraints. This is an artificial and mathematical trick, and it leads to new biomathematical functions that we call modified TCP and NTCP, $\mathrm{TCP}_{\mathrm{MOD}}$ and $\mathrm{NTCP}_{\mathrm{MOD}}$, respectively, built using conveniently chosen values for $D_{50,\mathrm{MOD}}$ and $\mathrm{TD}_{50,\mathrm{MOD}}$, respectively. For $\mathrm{TD}_{50,\mathrm{MOD}}$, the chosen values, noted as $D_{v,c}$, can be found in Table III and correspond to the dose plan constraint (c) at each particular volume of interest (VOI) (v).

For $D_{50,\mathrm{MOD}}$, the chosen value is the 80% of the prescribed dose. That locates the $\mathrm{TCP}_{\mathrm{MOD}}$ dose–response curve in an adequate region for each plan.

Modified TCP and NTCP functions lose, of course, their radiobiological meaning but are considered to be better predictors of the plan viability. The curve slopes are calculated using gEUD. Taking the same equations for the TCP and NTCP as in Zhen *et al.*,[15] the partial derivative of $\mathrm{TCP}_{\mathrm{MOD}}$ for every PTV is evaluated. First, gEUD is calculated following Eq. (2), and then, the TCP logit expression when the fractional

TABLE II. List of parameters used in radiobiological calculations.

Organ	D_{50}/TD_{50} (cGy)	γ_{50}	n $(1/a)$	m
H&N PTV	5044	1.73	−0.1	
Prostate PTV	7050	2.66	−0.1	
Brain PTV	4800	2	−0.1	
Lung PTV	5100	1.6	−0.1	
Rectum	7200		0.06	0.15
Bladder	6200		0.13	0.11
Femoral head	6500		0.25	0.12
Cord	6650		0.05	0.175
Brain stem	7000		0.05	0.18
Parotid	2840		1	0.18
Lens	1250		0.3	0.27
Optical nerve	6500		0.25	0.14
Heart	4800		0.35	0.1
Kidney	4000		0.7	0.1
Chiasm	6500		0.25	0.14
Small bowel	5500		0.15	0.16

TABLE III. IMRT plan constraints ($D_{v,c}$) extracted from the DVH analysis.

Organs at risk	Constraint
Brain stem	$D1 < 53$ Gy
Optical nerves and chiasm	$D1 < 49.25$ Gy
Spinal cord	$D1 < 44.5$ Gy
Parotid glands	Average dose < 26 Gy
Lenses	$D1 < 8$ Gy
Rectum	$D20 < 70$ Gy
Bladder	$D30 < 70$ Gy
PTV	**Constraint**
Constraint 1	$D95 > 95\%$ prescribed dose
Constraint 2	Average dose > 0.97* prescribed dose
Constraint 3	$D2 < 110\%$ prescribed dose

volume is 1 is derived, as all voxels are supposed to have the gEUD,

$$\frac{\partial \mathrm{TCP}_{\mathrm{MOD},i}}{\partial \mathrm{gEUD}_i} = \frac{4 \cdot D_{50,\mathrm{MOD}} \cdot \left(\frac{D_{50,\mathrm{MOD}}}{\mathrm{gEUD}_i}\right)^{4\cdot\gamma_{50}-1} \cdot \gamma_{50}}{\left(1+\left(\frac{D_{50,\mathrm{MOD}}}{\mathrm{gEUD}_i}\right)^{4\cdot\gamma_{50}}\right)^2 \cdot \mathrm{gEUD}_i{}^2}. \quad (3)$$

The same process is applied to NTCP and a partial derivative can be defined using $\mathrm{TD}_{50,\mathrm{MOD}}$,

$$\frac{\partial \mathrm{NTCP}_{\mathrm{MOD},i}}{\partial \mathrm{gEUD}_i} = \frac{0.399 \cdot e^{-\frac{\left(\mathrm{gEUD}_i - \mathrm{TD}_{50,\mathrm{MOD}}\right)^2}{2\cdot m^2 \cdot \mathrm{TD}^2_{50,\mathrm{MOD}}}}}{m \cdot \mathrm{TD}_{50,\mathrm{MOD}}}. \quad (4)$$

The $\mathrm{TCP}_{\mathrm{MOD}}$ and $\mathrm{NTCP}_{\mathrm{mod}}$ curves reach their maximum slope when $\mathrm{gEUD} = D_{50,\mathrm{MOD}}$ for the PTV and $\mathrm{gEUD} = \mathrm{TD}_{50,\mathrm{MOD}}$ for the OARs, respectively. These values are

$$\left.\frac{\partial \mathrm{TCP}_{i,\mathrm{MOD}}}{\partial \mathrm{gEUD}_i}\right|_{\max} = \frac{\gamma_{50}}{D_{50,\mathrm{MOD}}}$$

and

$$\left.\frac{\partial \mathrm{NTCP}_{i,\mathrm{MOD}}}{\partial \mathrm{gEUD}_i}\right|_{\max} = \frac{0.399}{m \cdot TD_{50,\mathrm{MOD}}}.$$

First, gEUD is evaluated for every organ for a given plan, and this value is used to evaluate the TCP and NTCP curve slope at the considered point. Then, the ratio to the maximum slope is calculated and all the components for every volume i are multiplied, being N the total number of organs, including PTV. The robustness index is defined as

$$\mathrm{RI} = \prod_{i=1}^{N}\left(1 - \frac{\frac{\partial \mathrm{XCP}_{i,\mathrm{MOD}}}{\partial \mathrm{gEUD}_i}}{\left.\frac{\partial \mathrm{XCP}_{i,\mathrm{MOD}}}{\partial \mathrm{gEUD}_i}\right|_{\max}}\right). \quad (5)$$

Note that the first initial of the modified radiobiological index has been noted as "X," which indicates that may be substituted by "T" or "NT" depending on the considered volume, PTV or OAR. RI is zero if one of the considered volumes has a gEUD that equals $D_{50,\mathrm{MOD}}$ or $\mathrm{TD}_{50,\mathrm{MOD}}$. In this case, a plan is classified as "barely robust." A plan is "robust," i.e., RI close to 1, if none of the calculated slopes for the modified TCP or NTCP curves approaches to the $D_{50,\mathrm{MOD}}$ or $\mathrm{TD}_{50,\mathrm{MOD}}$ values, respectively. In addition, a plan with an organ receiving a dose that is significantly higher than the dose constraint would also be a robust plan. For instance, if one of the two parotid glands is close to the tumor, accepting its loss enables the average dose to be considerably higher than the dose constraint (usually 26 Gy). In this case, RI will not be affected by significant dose variations involving this particular parotid gland.

2.G. Proposed metrics

IMRT verification is similar to a diagnosis problem. One has to evaluate if the result of the applied test is positive or negative, i.e., if the considered plan is rejected or accepted. When studying a set of cases and two metrics to classify them, contingency tables can be established, if it is assumed for convenience that one of the metrics is the reference metric. In this work, as stated in the Introduction, a metric is given always by a variable, and this variable can be continuous or binary.

TABLE IV. Generic contingency table.

		Reference variable		
Contingency table		Positive 0	Negative 1	Total
Tested variable	Positive 0	TP	FP	TP + FP
	Negative 1	FN	TN	FN + TN
Total		Positives	Negatives	Total cases

In Table IV, a generic contingency table is shown between binary variables (0 is taken as positive or, in our case, plan rejection and 1 is taken as negative or plan acceptance). In this kind of table, there are two binary variables compared, one set as reference and the other set as tested. Every variable has its own positives and negatives and the reference variable is the one that classifies the results of the tested variable into false results (FN, FP) or true results (TN, TP, true negatives or positives).

For metrics that are represented by continuous variables, it is always possible to discretize the variable in finite steps, once a cutoff value is fixed. For metrics that are represented by binary variables, one may vary the cutoff value that defines the acceptance/rejection criteria and build a set of contingency tables based on different cutoffs. A binary variable is constructed based on the cases that have a value under or over the cutoff.

In this work, we propose and define new metrics that can be calculated from the DVHs and the IMRT plan constraints. These metrics are based on continuous or binary variables. First, three new continuous functions are defined. The purpose of these functions is to synthesize in a few parameters the differences found between the DVHs coming from verification dose matrix and those coming from the TPS algorithm. The first two functions are defined for every algorithm separately. To evaluate them, the constraints in Table III and the values obtained for the same evaluation points in the DVH for the considered algorithm are used. The third function does not use the constraints but directly compares the DVH values obtained from two different algorithms. Let N be the total number of constraints.

Therefore, we define these three functions as follows:

— The weighted noncompliance function, WNCF, defined for a particular algorithm with an output dose matrix, as the quadratic sum of the differences between the evaluation doses at DVH points defined in Table III and the dose constraints (c) that are shown in the same table for each volume v. The difference contributes only if the constraint is not fulfilled,

$$\mathrm{WNCF} = \frac{100 \cdot \sqrt{\sum_{v=1}^{N} w_v \delta_{v,c} \left(D_{v,\mathrm{eval}} - D_{v,c}\right)^2}}{D_{\mathrm{prescribed}}}, \quad (6)$$

where
$\delta_{v,c} = 1$ if $D_{\mathrm{PTV,eval}} < D_{\mathrm{PTV},c}$ (average dose and $D95$) or $D2_{\mathrm{PTV,eval}} > D2_{\mathrm{PTV},c}$ or $D_{\mathrm{OAR,eval}} > D_{\mathrm{OAR},c}$

$\delta_{v,c} = 0$ if $D_{\text{PTV,eval}} > D_{\text{PTV},c}$ (average dose and $D95$) or $D2_{\text{PTV,eval}} < D2_{\text{PTV},c}$ or $D_{\text{OAR,eval}} < D_{\text{OAR},c}$. The subindex v stands for a PTV or for an OAR. It can take three values in the case of a primary PTV ($D95$, average dose, and $D2$) or two values ($D95$ and average dose), if it is a secondary PTV. The sum is normalized to the primary PTV prescription dose. The constant w_v is employed as a weight that takes into account if one constraint is to be considered more restrictive than another. An adequate set of w_v is going to be investigated in this work. Six different sets are proposed (see Table V). These sets are guessed sets where different possibilities of PTV-OAR relative importance are explored. For instance, set 1 and set 2 change the relative importance of the clinical constraints among OARs, while the other four sets explore the PTV-OAR balance. The results of this method may depend on the selection of the weighting set.

As mentioned above, WNCF is calculated independently for every algorithm; therefore, the subindex eval is employed, which stands for the evaluation method, and it can be the one used for the verification (verif) or the TPS algortihm.

Consequently, WNCF will exist for both systems, being noted as $\text{WNCF}_{\text{verif}}$ and WNCF_{TPS}. The difference $\Delta\text{WNCF}_{\text{verif}} = \text{WNCF}_{\text{verif}} - \text{WNCF}_{\text{TPS}}$ measures the deterioration found in the verification due to the nonfulfillment of the constraints. The subindex "verif" can be substituted by the name of the verification algorithm employed, thus giving rise to two variables, namely, ΔWNCF_CMPM and ΔWNCF_MCPS.

— The dose difference constraints-based global function, CGF, is defined as the sum of the differences between the dose values presented in Table III and the dose values taken from the DVH at the same evaluation points. This function is different from WNCF because the constraints do not have to be necessarily unfulfilled to obtain a contribution in every evaluation point; therefore, this contribution can be positive or negative,

$$\text{CGF} = \frac{100 \cdot \sum_{v=1}^{N} k \cdot (D_{v,c} - D_{v,\text{eval}})}{D_{\text{prescribed}}}, \tag{7}$$

where $k = 1$ if v corresponds to an OAR or a $D2$ from the primary PTV and $k = -1$ if v corresponds to an average dose or a $D95$ from a PTV.

CGF is defined for every separated algorithm, in the same way as WNCF. CGF is normalized to a percentage relative to the prescribed dose. Again, $\Delta\text{CGF}_{\text{verif}} = \text{CGF}_{\text{verif}} - \text{CGF}_{\text{TPS}}$ is a measurement of the discrepancy between verification system and TPS calculation. The subindex verif can be substituted by the name of the verification algorithm employed giving rise to two variables, ΔCGF_CMPM and ΔCGF_MCPS.

— Dose difference global function, GF, is defined as the sum with its sign of the differences extracted from the DVH obtained with the verification algorithm (MCPS or CMPM) and the TPS algorithm. These differences are taken into account only if they correspond to the deterioration of the ability to meet constraints in Table III, i.e., decreased PTV coverage, increased PTV hotspots, or increased dose to the OARs at the constraint levels listed in Table III. This function directly compares the results from two algorithms; therefore, it is the metric that measures the differences found in the verification. GF does not use the dose constraint values, but only the evaluation points defined in Table III (i.e., $D95$, average dose, and $D2$ for the PTV, and $D1$, $D20$, $D30$, or average dose depending on the evaluated OAR),

$$\text{GF} = \frac{100 \cdot \sum_{v=1}^{N} \delta_{\text{verif,TPS}} (D_{v,\text{verif}} - D_{v,\text{TPS}})}{D_{\text{prescribed}}}. \tag{8}$$

$\delta_{\text{verif,TPS}} = -1$ if $D_{\text{PTV,verif}} < D_{\text{PTV,TPS}}$ for PTV average dose and $D95$

$\delta_{\text{verif,TPS}} = 1$ if $D_{\text{OAR,verif}} > D_{\text{OAR,TPS}}$ or $D2_{\text{PTV,verif}} > D2_{\text{PTV,TPS}}$

$\delta_{\text{verif,TPS}} = 0$ if $D_{\text{OAR,verif}} < D_{\text{OAR,TPS}}$ or $D_{\text{PTV,verif}} > D_{\text{PTV,TPS}}$ for PTV average dose and $D95$

GF can be applied to the CMPM-TPS comparison or to the MCPS-TPS comparison giving rise to two variables: GF_CMPM and GF_MCPS.

TABLE V. Different weighting sets for the WNCF. Every row corresponds to a different constraint. The order is as follows: PTV (average dose, $D95$, and $D2$ with the same weight), brain stem ($D1$), spinal cord ($D1$), rectum ($D20$), bladder ($D30$), parotid gland (average dose), lens ($D1$), optical nerve ($D1$), and chiasm ($D1$).

	Set1	Set2	Set3	Set4	Set5	Set6
wmptv	1	1	0.001	1000	0,5	1
wbsd1	0.04	0.4	1	0.1	2	1
wcordd1	0.075	0.75	1	0.1	2	1
wrd20	0.05	0.5	1	0.1	2	1
wbd30	0.05	0.5	1	0.1	2	1
wparad	0.05	0.5	1	0.1	2	1
wlend1	0.025	0.25	1	0.1	2	1
wond1	0.04	0.4	1	0.1	2	1
wchid1	0.04	0.4	1	0.1	2	1

These three functions, WNCF, CGF, and GF, are clinically relevant. For instance, if the primary PTV has a higher $D95$ in the verification DVH set than in the one coming from the TPS, this fact is not penalized because this kind of dose difference is not considered as detrimental.

Apart from these three functions, other metrics are going to be applied to the plan verification process. The first two of the following metrics will be used as convenient reference metrics in this study:

— Dose difference metric (DDM). It is a binary variable that depends on the evaluation of the TPS DVH and the verification DVH at the points established in Table III. The plan is accepted (negative → 1) when the deterioration, if produced, is less than 3% of

the prescribed dose at all these points. Otherwise, the plan is rejected (positive → 0). In this context, "deterioration" means a lower verification dose at the PTVs for $D95$ and average dose, and a higher verification dose at $D2$ for the PTVs and at the OARs for any defined dose constraint. If the PTV is a secondary one, the $D2$ constraint is not taken into account. The binary variable associated is called DDM_CMPM and DDM_MCPS, depending on the algorithm which is being compared with the TPS.

— Constraint-based dose difference metric (CDDM). The previous metric may yield a rejection for plans that, however, might be acceptable from a clinical point of view. For instance, let us consider that the plan verification is for a brain tumor treatment, and a 5% global dose increase has been obtained in the $D1$ parameter for the brain stem. If the $D1$ was 35 Gy according to the TPS data, the verification $D1$ would be around 37 Gy. This means plan rejection under the DDM metric but, unless there is any other verification issue, the consequences to the patient would be negligible because the dose constraint is $D1 < 50$ Gy. In this perspective, a metric taking into account the plan constraints can be defined, and that means that the plan is rejected only when the dose differences lead to the nonfulfillment of the constraints. The associated binary variables are CDDM_CMPM and CCDM_MCPS. They should produce as many accepted plans as DDM_CMPM and DDM_MCPS or more, but never less.

— Volumes of interest 3D gamma function-based metric. This metric is built by evaluating the 3D global gamma function (3%–3 mm) for PTVs, OARs, and BODY contours, and calculating the percentage gamma rejection rate for each volume. It is only available for the CMPM algorithm. When one VOI or more have a rejection rate over a fixed cutoff X, the metric yields the plan rejection. The normalization dose value in the gamma test is assigned to the prescribed dose.

— ΔTCP and ΔNTCP are the increments that evaluate the difference found in the verification between the verification algorithm and TPS from a radiobiological point of view. For the modified TCP and NTCP, the metric is evaluated in the same way. The binary metric based on these increments works this way: a plan is accepted if $\Delta TCP > -X$ and, at the same time, $\Delta NTCP < X$ for all the PTVs and organs analyzed, X being a certain value, expressed in percentage. The value X can be varied to construct a discrete function to be used as a classifier in the receiver operating characteristics (ROC) analysis (Sec. 2.H). Two binary variables are defined, ΔTCP/NTCP and $\Delta TCP_{MOD}/NTCP_{MOD}$ corresponding to nonmodified or modified functions, respectively. They can be applied to CMPM-TPS differences or to MCPS-TPS differences. The notation ΔTCP/NTCP indicates that ΔTCP and ΔNTCP are evaluated at the same time for the plan volumes.

— For the robustness index, ΔRI is defined as a continuous metric that quantifies RI difference between the verification algorithm and the TPS.

— There is also a binary variable associated to the conventional verification, which takes into account the procedure explained above (Sec. 2.C). CONV is 1 (acceptance) when the dose measured with the ionization chamber is within ±3% of the TPS predicted dose and when the gamma passing rate in the measured planes is over 90%, while CONV is 0 (rejection) if any of these conditions are not met. This planar analysis includes only one of the three methods described in Sec. 2.C depending on the measurement date.

A complete summary of all the proposed metrics is presented in Table VI.

DDM and CDDM try to mimic the medical physicist process of assessing the DVH-based plan verification. Both variables are based on the dose comparison at certain DVH critical points. This is the reason why they are used for convenience as reference variables in the subsequent statistical analysis.

2.H. Statistical analysis

A set of contingency tables representing different cutoff levels can be treated with ROC analysis, which measures

TABLE VI. Summary of metrics and notation.

Metric	Description	Associated binary variables
ΔWNCF	Weighted noncompliance function	ΔWNCF_CMPM, ΔWNCF_MCPS
ΔCGF	Dose difference constraints-based global function	ΔWNCF_CMPM, ΔWNCF_MCPS
GF	Dose difference global function	GF_CMPM, GF_MCPS
DDM	Dose difference metric	DDM_CMPM, DDM_MCPS
CDDM	Constraint-based dose difference metric	CDDM_CMPM, CDDM_MCPS
VOI 3Dgamma	Volumes of interest 3D gamma function	VOI_3Dgamma
ΔTCP/NTCP	Differences in TCP and NTCP between verification algorithm and TPS	ΔTCP/NTCP_CMPM, ΔTCP/NTCP _MCPS, $\Delta TCP_{MOD}/NTCP_{MOD}$_CMPM, $\Delta TCP_{MOD}/NTCP_{MOD}$_MCPS
ΔRI	Difference in robustness index	ΔRI _CMPM, ΔRI_MCPS
CONV	Conventional verification	CONV

the predictive power of a variable in relation to the acceptance/rejection criteria of a certain reference classification variable. This has been done for the metrics defined in Table VI in two senses, continuous variables predicting binary variable results and binary variables predicting binary variable results. Each of the continuous functions defined above, ΔWNCF, ΔCGF, and GF, is used as input in the ROC analysis proposed, evaluating their performance when predicting the results of the binary variables DDM_MCPS, DDM_CMPM, CDDM_MCPS, and CDDM_CMPM. As a consequence, a geometrically optimal cutoff point (OCP) value and the area under the curve (AUC) are obtained. For a given ROC curve, the geometrically OCP corresponds to the point along the curve farthest from the diagonal random guess line. AUC measures the probability that a positive (under the reference variable) randomly chosen from the sample will have a higher value of the classifying continuous variable than a negative chosen in the same way (assuming a positive yields a higher value than a negative because the opposite may occur). Common statistical software, the statistical package for the social sciences (SPSS) for Windows v15.0, is employed to do this kind of ROC analysis.

On the other hand, a set of contingency tables with a binary variable can be generated if the cutoff is changed progressively. A ROC curve can be built where every point corresponds to a different cutoff value. This has been done with VOI_3Dgamma, ΔTCP/NTCP, and $\Delta TCP_{MOD}/NTCP_{MOD}$. For this curves, an OCP is also extracted. In the case of binary variables with varying cutoff, the ROC curve is built manually with a spreadsheet. The AUC can be integrated subsequently.

Note that once an OCP value has been obtained for the ROC curve, a derived binary variable can be defined classifying plans with values over/under the OCP (see Sec. 3.E) for the metric used as input in the ROC analysis. Consequently, the following binary variables are defined: GF_CMPM_OCP, GF_MCPS_OCP, VOI_3Dgamma_OCP, $\Delta TCP_{MOD}/NTCP_{MOD}$_CMPM_OCP, $\Delta TCP_{MOD}/NTCP_{MOD}$_MCPS_OCP, ΔRI_CMPM_OCP, and ΔRI_MCPS_OCP. They will be used with the explicit OCP value obtained in the ROC analysis.

The correlation between binary variables is as well investigated. The chi-squared test for categorical data with one degree of freedom is used. If the p-value of the chi-squared test is less than 0.05 (associated with a high chi-squared value), the variables are not considered to be independent in relation to the plan classification into accepted or rejected plans. On the other hand, if p-value is higher than 0.05, the two binary variables considered classify the plans in a very different way.

3. RESULTS

3.A. Best weights (w_v) for WNCF

As mentioned before, the ROC analysis has been employed to determine the best set of weights of those listed in Table V. The reference metric role can be played by DDM_CMPM or, alternatively, DDM_MCPS. In this case, the continuous variables investigated are ΔWNCF_CMPM and ΔWNCF_MCPS as described in Sec. 2.G.

The results are shown in Fig. 1. The weighting set that better predicts the behavior of the dose difference metrics is the number 2, as it presents the best balanced AUC considering both investigated cases. This optimal weighting set might be dependent on the chosen patient data set; therefore, it cannot be generalized to other cases.

3.B. Best continuous function

The predictive power of ΔWNCF (with weighting set number 2), ΔCGF, and GF, the continuous functions defined in Sec. 2.G, has been investigated. The ROC analysis for these

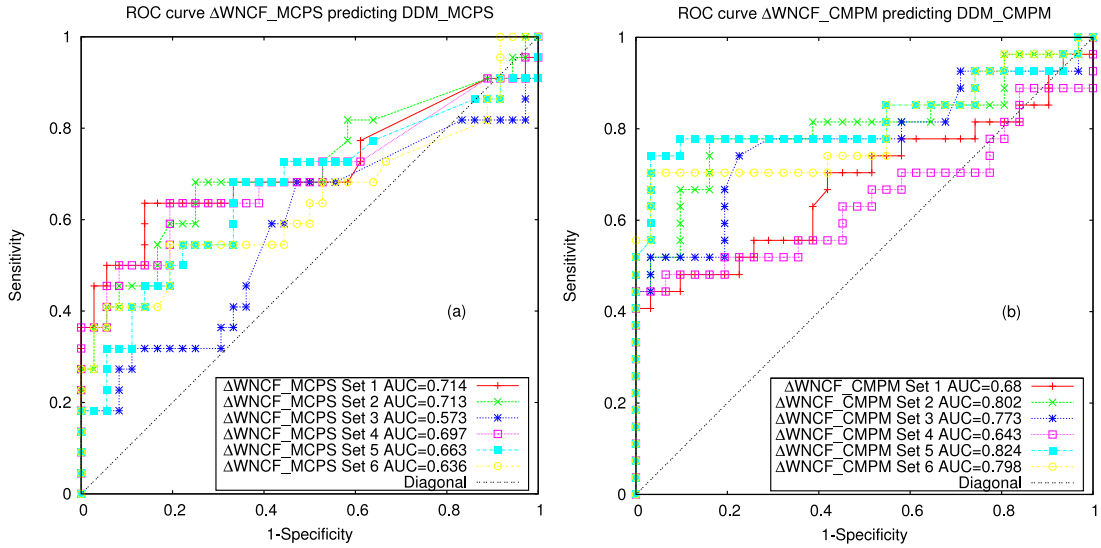

FIG. 1. ROC curves for the increment in ΔWNCF depending on the chosen weighting set. Two binary variables are used as reference, DDM_MCPS (a) and DDM_CMPM (b), and the respective continuous functions (ΔWNCF_MCPS and ΔWNCF_CMPM) are used as predictors. The best balanced set is the number 2, based on AUC values.

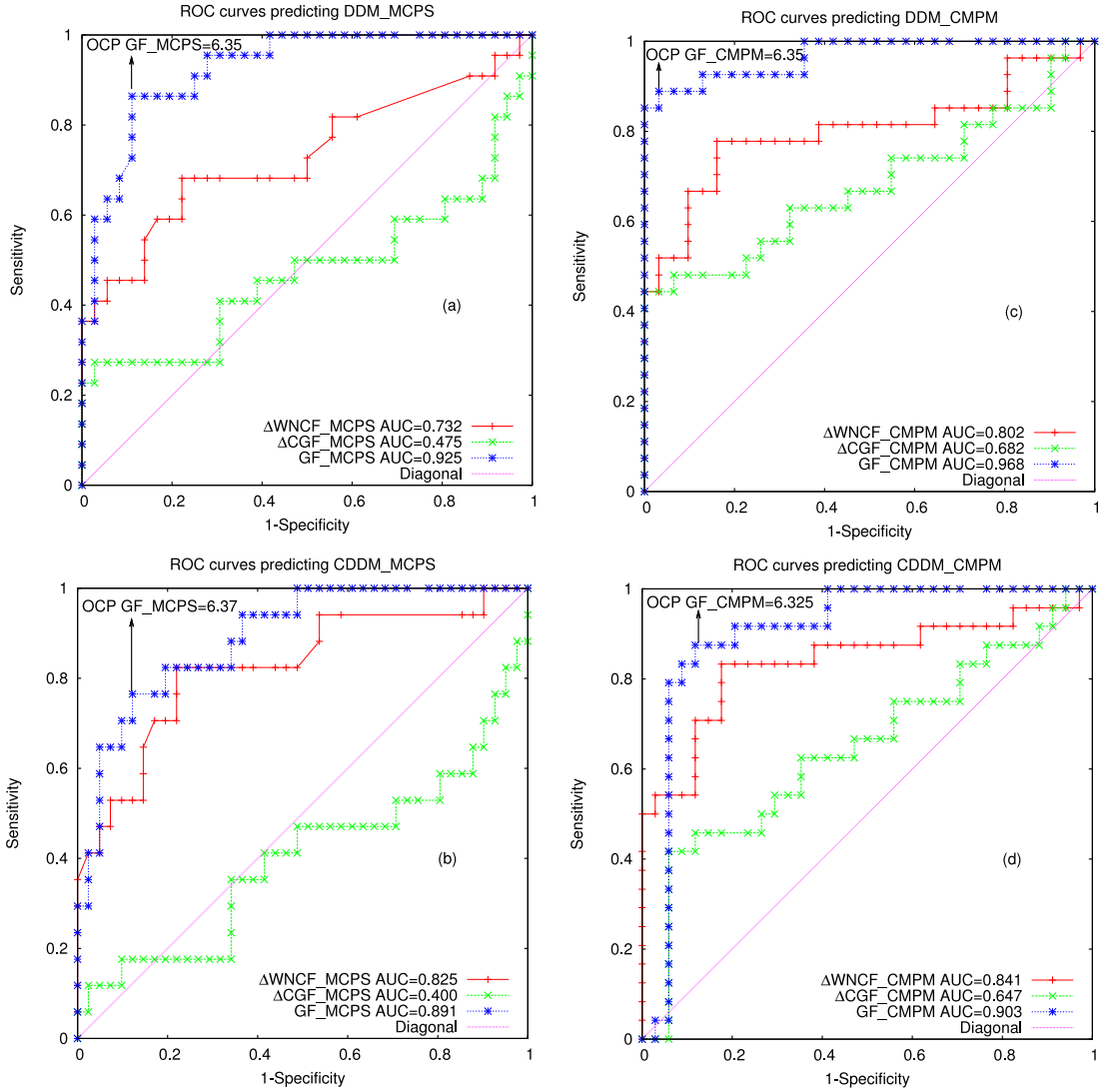

FIG. 2. ROC curves for the three continuous functions proposed as predictors, ΔWNCF, ΔCGF, and GF. Four binary variables are used as reference, DDM_MCPS (a), CDDM_MCPS (b), DDM_CMPM (c), and CDDM_CMPM (d). AUC is shown for all the predictors, and OCP is shown in every graph only for the best predictor, GF.

three continuous functions when used to predict the dose difference metrics and the constraint-based dose difference metrics for both algorithms, CMPM and MCPS (see also Sec. 2.G), is displayed in Fig. 2.

The results indicate clearly that GF has the better performance at predicting the dose difference metrics. GF has in any case the biggest AUC and can be employed as an excellent classifier from the point of view of the dose difference metrics. The OCP (defined in Sec. 2.H) found for GF is 6.35, and this is valid for the four ROC curves where GF is involved, given the fact that their AUC differs only slightly among them. The performance of ΔWNCF and ΔCGF is poorer and the reason for this might be that they use the constraints $D_{v,c}$ to calculate the contributions for the verification and TPS algorithms. Additionally, CGF gets a contribution from every evaluation point in Table III, and this means that, whenever there is an improvement in the comparison between verification and TPS, the deterioration in other points may happen to cancel out.

3.C. VOI 3D gamma function

The prediction of the 3D gamma function evaluated in every patient volume of interest with respect to the dose difference metrics binary variables is shown in Fig. 3. The gamma rejection rate values are identified with the variable cutoff to build the ROC curves. The values are 1%, 5%, 7.5%, 10%, 15%, 25%, 30%, and 100%. There is a good correlation between VOI 3D gamma function and DDM_CMPM or CDDM_CMPM. Moreover, the optimal cutoff point for VOI 3Dgamma, in both cases, corresponds to a rejection rate of 10%. At this point, the sensitivity is around 80% and the specificity is around 90%.

3.D. Biomathematical treatment outcome models and robustness index

In Fig. 4, the results regarding the predictive power of ΔTCP/NTCP and $\Delta TCP_{MOD}/NTCP_{MOD}$, defined in Sec. 2.G,

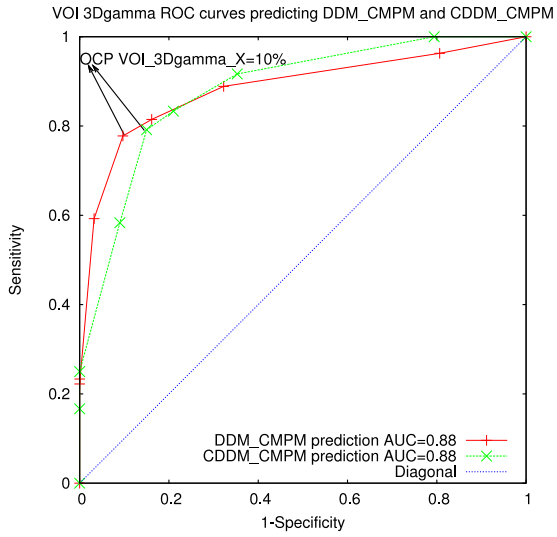

FIG. 3. ROC curves for the VOI 3Dgamma function, built by changing the acceptance threshold. Two binary variables are used as reference, DDM_CMPM and CDDM_CMPM. The OCP for the two curves is 10% (rejection rate for the VOI 3Dgamma function).

are presented. The variables taken as reference are the constraint-based dose difference metrics, CDDM_MCPS [Fig. 4(a)] and CDDM_CMPM [Fig. 4(b)], corresponding to the reference algorithms MCPS and CMPM, respectively. The chosen values for the cutoff to build the ROC curves are 1%, 2%, 3%, 5%, 7%, 9%, 12%, and 15%. The percentages refer to allowed ΔTCP or ΔNTCP.

The performance of ΔTCP/NTCP for both algorithms is poor and the AUC is below 0.7 in both cases. This was expected from the lack of sensitivity of TCP and NTCP to certain dose changes as described in Sec. 2.E. It is also difficult to establish a useful cutoff value. If TCP_{MOD} and $NTCP_{MOD}$ are used instead, the result is better, which suggests that the correlation between biomathematical models and dose differences is improved by the modification. AUC is 0.73 for $\Delta TCP_{MOD}/NTCP_{MOD}$_MCPS predicting CDDM_MCPS and 0.79 for $\Delta TCP_{MOD}/NTCP_{MOD}$_CMPM predicting CDDM_CMPM. OCP is 2% for both curves. However, although the improvement in correlation is noticeable, it is clear that the behavior of dose difference metrics and biomathematical model-based metrics is not equivalent.

3.E. Correlations

As stated above, certain OCP have been obtained for several continuous variables used as input in the ROC analysis. For GF, OCP = 6.35; for VOI_3Dgamma, OCP = 10%; and for $\Delta TCP_{MOD}/NTCP_{MOD}$, OCP = 2%. Therefore, five binary variables are obtained based on these thresholds: GF_CMPM_6.35, GF_MCPS_6.35, VOI_3Dgamma_10, $\Delta TCP_{MOD}/NTCP_{MOD}$_CMPM_2, and $\Delta TCP_{MOD}/NTCP_{MOD}$_MCPS_2. Note that VOI 3D gamma is only calculated for the CMPM algorithm.

The correlations between each pair of binary variables can be found in Table VII. This table presents a lot of condensed information regarding the plan classification comparison process. Every figure in bold in the table reflects a discrepancy that has to be investigated, keeping in mind that significance over 0.05 suggests independence of the plan classification made by a particular couple of variables. Following this criterion, two main discrepancies are observed: first, the conventional verification is mainly independent from the new metrics defined and second, some of the metrics coming from MCPS do not correlate properly with those coming from CMPM.

3.F. Robustness index, RI

The increment in the robustness index between the verification algorithm (CMPM or MCPS) and the TPS algorithm, ΔRI, has also been used to predict the outcome of CDDM_CMPM and CDDM_MCPS. The associated binary variable is ΔRI

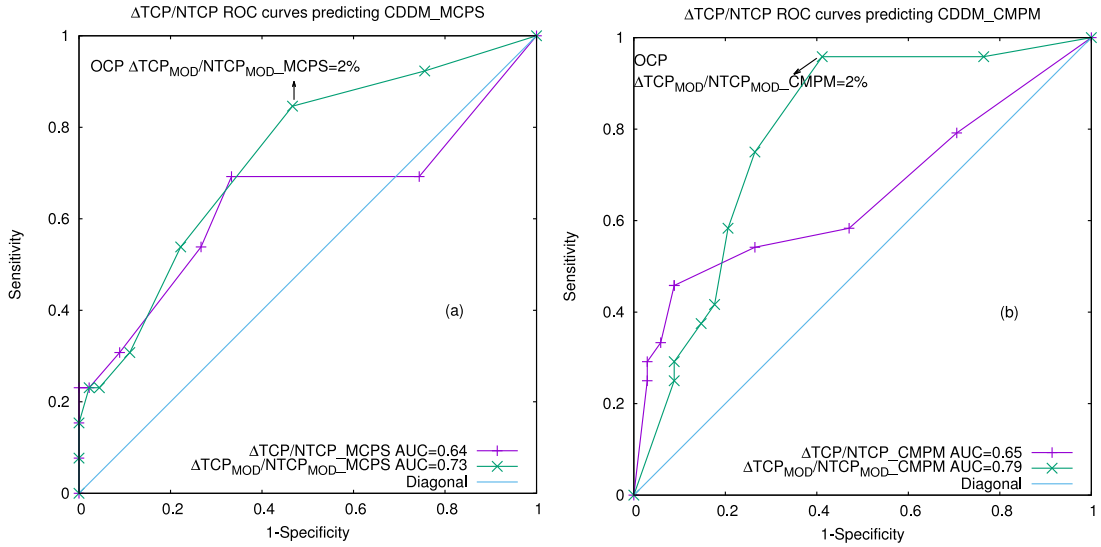

FIG. 4. ROC curves for the ΔTCP/NTCP and $\Delta TCP_{MOD}/NTCP_{MOD}$ functions, built by changing the acceptance threshold. Two binary variables are used as reference, CDDM_MCPS (a) and CDDM_CMPM (b). The OCP for $\Delta TCP_{MOD}/NTCP_{MOD}$ is shown, being 2% for both algorithms (MCPS and CMPM).

TABLE VII. Crossed table with chi-squared test signification and chi-squared value in brackets. Significations above 0.05 are in bold, indicating variables that can be considered independent, and therefore the plan classification arising from them is not correlated.

	DDM_CMPM	CDDM_MCPS	VOI_3Dgamma_10	CONV	$\Delta TCP_{MOD}/NTCP_{MOD}$ _CMPM_2	$\Delta TCP_{MOD}/NTCP_{MOD}$ _MCPS_2	GF_MCPS_6.35	GF_CMPM_6.35
DDM_CMPM	0	0.02(5.395)	0.0(24.1)	0.014(6.086)	0.0(18.196)	**0.684(0.165)**	0.009(6.880)	0.0(32.905)
CDDM_MCPS		0	**0.082(3.017)**	**0.346(0.888)**	**0.196(1.673)**	0.007(7.184)	0(23.196)	0.032(4.576)
VOI_3Dgamma10			0	**0.157(2.002)**	0.0(13.771)	0.044(4.06)	0.009(6.880)	0.0(32.905)
CONV				0	**0.118(2.438)**	**0.409(0.683)**	0.049(3.88)	**0.182(1.782)**
$\Delta TCP_{MOD}/NTCP_{MOD}$ _CMPM_2					0	0.012(6.348)	0.008(7.117)	0.0(24.939)
$\Delta TCP_{MOD}/NTCP_{MOD}$ _MCPS_2						0	0(16.276)	**0.087(2.923)**
GF_MCPS_6.35							0	0.003(9.024)
GF_CMPM_6.35								0

_CMPM_0.05 or ΔRI _MCPS_0.05, depending on the algorithm. For them, the plan is rejected (positive), if two conditions are met simultaneously for the TPS: RI < 0.05 and ΔRI < 0. Moreover, the plan is also rejected if ΔRI < −0.05. Otherwise the plan is accepted. Due to the poor performance of ΔRI _CMPM or ΔRI _MCPS in the ROC analysis, there was not a consistent OCP value for them; therefore, the cutoff choice has been made by analyzing the RI for the 58 studied plans, which present very low RI values. Equation (5) is used for the evaluation, and when one of the DVH values is close to the established constraint, RI quickly approaches zero. Thus, a plan is considered barely robust when RI < 0.05.

It follows from the above considerations that ΔRI is not a good plan acceptance/rejection predictor. However, RI calculated for the TPS plan seems to be a good indicator of the probability of plan acceptance. For instance, Table VIII shows the ability of RI to predict the output of two binary variables, $\Delta TCP_{MOD}/NTCP_{MOD}$_CMPM_2 and $\Delta TCP_{MOD}/NTCP_{MOD}$_CMPM_5, which are the variables associated to the OCP and the variable associated to a 5% increment, respectively. In the table, the distribution of plans which were accepted or rejected under these metrics is shown, depending on whether they were deemed to be robust or barely robust. For $\Delta TCP_{MOD}/NTCP_{MOD}$_CMPM_5, more plans are accepted, regardless of them being robust or just barely robust, but the rejection probability still continues to be higher for the barely robust plans. Thus, RI has a good sensitivity detecting plans that are likely to fail the test but a poor specificity.

4. DISCUSSION

In this work, new metrics for the evaluation of IMRT verification are proposed and tested. They can be used as plan classifiers when dealing with pretreatment "step and shoot" IMRT verification, or even in any verification that involves DVH comparison, such as 3D dose reconstruction from EPID *in vivo* dosimetry.[25] Moreover, the metrics and their meaning may be applied to other techniques such as dynamic IMRT or VMAT, where verification measurements are commonly employed.

It has been shown that, if a continuous function depends on several parameters, it is possible to use the AUC coming from the ROC analysis to determine the best set for these parameters, provided that the range of the parameters is properly chosen.

Also based on the ROC analysis, it is possible to evaluate the ability of new continuous variables or a set of binary variables to reproduce the plan classification obtained by a reference metric. We have tested whether the verification

TABLE VIII. TPS Plan robustness distribution with respect to the modified TCP/NTCP functions results ($\Delta TCP_{MOD}/NTCP_{MOD}$_CMPM_2 and $\Delta TCP_{MOD}/NTCP_{MOD}$_CMPM_5), see Sec. 2.G for the variables definition. The classification (noted as 0 = rejection or 1 = acceptance for the binary variables) defines a probability for the rejection of the plan. Plans barely robust (RI < 0.05) are likely to be rejected.

	Number of plans	$\Delta TCP_{MOD}/NTCP_{MOD}$ _CMPM_2 = 0	$\Delta TCP_{MOD}/NTCP_{MOD}$ _CMPM_2 = 1	Rejection probability (%)
Barely robust	21	18	3	85.7
Robust	37	19	18	51
	Number of plans	$\Delta TCP_{MOD}/NTCP_{MOD}$ _CMPM_5 = 0	$\Delta TCP_{MOD}/NTCP_{MOD}$ _CMPM_5 = 1	Rejection probability (%)
Barely robust	21	13	8	61.9
Robust	37	8	29	22

process can be condensed in only one parameter, and it was found that a dose difference GF is a good predictor not only of dose difference metrics but also of those coming from biomathematical treatment outcome models. GF is calculated using the DVH information with the evaluation of the verification and TPS algorithm at particular points as those defined in Table III. This can be generalized for a wider constraint set. The GF cutoff value, 6.35 in our case, depends on the threshold employed in the dose difference metrics (3% global dose). GF is very simple to calculate and does not need further data processing like the gamma function or others.

Regarding the use of biomathematical models, when implementing TCP and NTCP as new QA metrics, their uncertainties need to be taken into account based on the uncertainties of the model parameters. In this work, we have found differences between the dose difference metric results and the ΔTCP/NTCP metrics (Sec. 3.D). Although in Zhen *et al.*,[15] in an *in silico* study, the biomathematical treatment outcome models are shown to be sensitive and specific to differences between TPS and measurements, in clinical cases the correlation between both kinds of metrics is weak. There are several reasons for that: the dose difference metrics are based on clinical constraints that may not be in the sensitive region of the dose–response curve, the IMRT plan prescribed dose sometimes is not a curative dose or corresponds to a boost or an initial treatment phase, or a big maximum inside the PTV contributes exactly in the opposite direction when using the dose difference metrics or the ΔTCP/ΔNTCP metrics. Following these considerations, the correlation between the two kinds of metrics is improved by defining the modified TCP and NTCP, although the results are only partially successful when trying to mimic the dose difference metrics. The PTV coverage losses that generate plan rejection in CDDM_CMPM do not always produce the same effect on $\Delta TCP_{MOD}/NTCP_{MOD}$_CMPM_$X$. For instance, a PTV coverage loss over 3% in $D95$ or 5% in $D99$ means a 2% loss in TCP_{MOD}, indicating less sensitivity to this kind of changes. In addition, the OCP of 2% for $\Delta TCP_{MOD}/NTCP_{MOD}$_CMPM possesses a low specificity (60%), resulting in a higher number of false positives when compared to the dose difference metric. The rationale behind this is that now the $NTCP_{MOD}$ dose–response curve is very steep in the proximity of the constraint. Therefore, for the OARs, the dose differences are going to be amplified in $\Delta NTCP_{MOD}$. It should also be noted that the parameters in TCP and NTCP models are fitted from clinical data, and, therefore, are associated with their own uncertainties.

Furthermore, in Zhen *et al.*,[15] a pretreatment quantification of plan robustness is proposed. Here, a mathematical form is given to this concept through the robustness index, and there is a correlation between this parameter and the probability of plan rejection.

Regarding 3D gamma analysis (Sec. 3.C), good predictive power has been found for the VOI 3D gamma using the metric VOI_3Dgamma_10, relative to the dose difference metrics. In principle, some excess of false positives would be expected because the gamma function does not make the difference between dose difference sign while DDM_CMPM and DDM_MCPS do. However, when OARs are receiving lower doses following the verification algorithm, the same often occurs with the PTV coverage. This is because, in our case, the main cause of discrepancies between TPS prediction and verification is the MLC leaves position[19] (real average position of leaves compared to the TPS predicted position), which influences the verification process by simultaneously lowering (closed leaves) or increasing (open leaves) the dose to PTV and OARs. As found by other authors,[20] Optifocus MLC quality assurance for IMRT requires frequent and critical assessment of MLC leaf position calibration errors that may appear in many different ways.

The study of the correlations among binary variables leads to several valuable results (Sec. 3.E). First of all, the algorithm employed in 3D dose reconstruction from measurements influences the verification outcome. As shown in Table VII, MCPS and CMPM lead to a different plan classification depending on the metrics compared. Even those pairs of metrics with a chi-squared significance lower than 0.05 may present this behavior. For instance, GF_CMPM_6.35 and GF_MCPS_6.35 differ in the classification in 29% of the plans, although chi-squared significance is 0.003. One may think that this is due to the partial correction made on MCPS because only daily linac behavior is taken into account, but further investigations have revealed that the origin of the main discrepancies is derived from the MC vs CMPC comparison, i.e., Monte Carlo redundant calculation compared to the COMPASS dose calculation engine, which does not depend on measurements. This result is in agreement with those of other authors.[26] As stated in this study, an independent calculation algorithm in the dose reconstruction might lead to additional discrepancy in the verification results, if the dose reconstruction computation is not accurate enough. Additional uncertainties can be introduced by this kind of algorithms.

In regard to the conventional verification, deep disagreement in plan classification has been found when comparing to the new metrics, either coming from CMPM or from MCPS (Table VII). The new metrics yield always a higher number of rejected plans. The rationale behind this is that DVH-based metrics are more sensitive to delivery problems because they evaluate several points and they are more likely to fail than the conventional verification, where the ionization chamber is usually inside the PTV and the 2D gamma planar verification is not correlated to dose errors in anatomic regions-of-interest.[5,11] When measuring the previously approved treatments with the COMPASS system, we found that several plans had some issues based on the DVH verification, especially involving the PTV coverage. In our case, the old metrics were generating a significant number of false negatives, which means low sensitivity. This can be concluded also from the work of Stasi[6] that showed a low sensitivity if 3%/3 mm global gamma was employed, leading to a disputable predictive power for perpatient IMRT QA. Apart from that, some contribution to the disagreement could be explained by the fact that conventional and COMPASS measurements were not coincident in time.

As a corollary to the DVH-based metrics analysis, there is not a complete equivalence among them, suggesting that there is not a completely reliable metric. This arises from the fact that biomathematical function differences have an intrinsically different behavior than dose difference metrics, and also from the fact that a plan failing under DDM or CDDM could be accepted if other indicators as GF or VOI_GAMMA3D show a value below the cutoff. Moreover, no one can guarantee that a dose difference is always clinically relevant. Nonetheless, the recommendation is not to avoid the use of DDM or CDDM, for they are valid metrics to evaluate the plan deliverability, but other presented metrics can provide additional and valuable information and eventually change an initial acceptance/rejection decision.

Finally, a recommendation of IMRT pretreatment verification can be made based on our findings: first, select the most robust plans that can be obtained from the TPS, then evaluate subsequently the dose difference GF and VOI 3Dgamma (if possible). If the obtained values are under the cutoffs, the plan may be accepted. Otherwise, further investigation of the causes that produce the plan rejection under those functions is needed.

5. CONCLUSION

New IMRT verification metrics have been proposed and their relation and acceptance/rejection predictive power has been assessed. The ROC analysis has been proved especially useful to find classifiers and evaluate their power. The dose difference GF and the VOI 3Dgamma are good classifiers regarding dose difference metrics and a rejection cutoff can be defined for them. The correlation between biomathematical treatment outcome models and the dose difference-based metrics is improved by using modified TCP and NTCP functions which take into account the plan constraints. The robustness index, specifically calculated for each plan can be used to predict how high the rejection probability is under modified TCP and NTCP metrics. Conventional verification has to be replaced by the new metrics, which are more relevant from a clinical point of view.

ACKNOWLEDGMENTS

This work was supported by the Health Investigation Program of Ministerio de Ciencia e Innovacion (Spain), Project No. PI11/01274. The authors would also like to thank the Project of INFRAESTRUCTURA CIENTIFICO TECNOLOGICA (Reference No. UNZA-08-4E-014).

CONFLICT OF INTEREST DISCLOSURE

The authors have no COI to report.

[a]Author to whom correspondence should be addressed. Electronic mail: agarciarom@salud.aragon.es; Telephone: +34976768839.

[1]G. A. Ezzell, J. M. Galvin, D. Low, J. Palta, I. Rosen, M. B. Sharpe, P. Xia, Y. Xiao, L. Xing, and C. X. Yu, "Guidance document on delivery, treatment planning, and clinical implementation of IMRT: Report of the IMRT subcommittee of the AAPM Radiation Therapy Committee," Med. Phys. **30**(8), 2089–2115 (2003).

[2]D. A. Low and J. F. Dempsey, "Evaluation of the gamma dose distribution comparison method," Med. Phys. **30**(9), 2455–2464 (2003).

[3]D. A. Low, W. B. Harms, S. Mutic, and J. A. Purdy, "A technique for the quantitative evaluation of dose distributions," Med. Phys. **25**(5), 656–661 (1998).

[4]J. J. Kruse, "On the insensitivity of single field planar dosimetry to IMRT inaccuracies," Med. Phys. **37**(6), 2516–2524 (2010).

[5]B. E. Nelms, H. Zhen, and W. A. Tomé, "Per-beam, planar IMRT QA passing rates do not predict clinically relevant patient dose errors," Med. Phys. **38**(2), 1037–1044 (2011).

[6]M. Stasi, S. Bresciani, A. Miranti, A. Maggio, V. Sapino, and P. Gabriele, "Pretreatment patient-specific IMRT quality assurance: A correlation study between gamma index and patient clinical dose volume histogram," Med. Phys. **39**(12), 7626–7634 (2012).

[7]B. J. Waghorn, S. L. Meeks, and K. M. Langen, "Analyzing the impact of intrafraction motion: Correlation of different dose metrics with changes in target D95%," Med. Phys. **38**(8), 4505–4511 (2011).

[8]K. B. Pulliam, J. Y. Huang, R. M. Howell, D. Followill, R. Bosca, J. O'Daniel, and S. F. Kry, "Comparison of 2D and 3D gamma analyses," Med. Phys. **41**(2), 021710 (6pp.) (2014).

[9]J. L. Bedford, Y. K. Lee, P. Wai, C. P. South, and A. P. Warrington, "Evaluation of the delta4 phantom for IMRT and VMAT verification," Phys. Med. Biol. **54**(9), N167–N176 (2009).

[10]J. Godart, E. W. Korevaar, R. Visser, D. J. L. Wauben, and A. A. Van't Veld, "Reconstruction of high-resolution 3D dose from matrix measurements: Error detection capability of the COMPASS correction kernel method," Phys. Med. Biol. **56**(15), 5029–5043 (2011).

[11]P. Carrasco, N. Jornet, A. Latorre, T. Eudaldo, A. Ruiz, and M. Ribas, "3D DVH-based metric analysis versus per-beam planar analysis in IMRT pretreatment verification," Med. Phys. **39**(8), 5040–5049 (2012).

[12]M. Bakhtiari, A. Parniani, F. Lerma, S. Reynolds, J. Jordan, A. Sedaghat, M. Sarfaraz, and J. Rodgers, "Evaluation of a software system for estimating planned dose error in patients, based on planar IMRT QA measurements," Radiol. Oncol. **48**(1), 87–93 (2014).

[13]V. Feygelman, G. Zhang, C. Stevens, and B. E. Nelms, "Evaluation of a new VMAT QA device, or the 'X' and 'O' array geometries," J. Appl. Clin. Med. Phys. **12**(2), 146–168 (2011).

[14]H. Zhen, B. E. Nelms, and W. A. Tomé, "Moving from gamma passing rates to patient DVH-based QA metrics in pretreatment dose QA," Med. Phys. **38**(10), 5477–5489 (2011).

[15]H. Zhen, B. E. Nelms, and W. A Tomé, "On the use of biomathematical models in patient-specific IMRT dose QA," Med. Phys. **40**(7), 071702 (10pp.) (2013).

[16]E. M. McKenzie, P. A. Balter, F. C. Stingo, J. Jones, D. S. Followill, and S. F. Kry, "Toward optimizing patient-specific IMRT QA techniques in the accurate detection of dosimetrically acceptable and unacceptable patient plans," Med. Phys. **41**(12), 121702 (15pp.) (2014).

[17]V. Laliena and A. García-Romero, "Monte Carlo modeling of the Siemens optifocus multileaf collimator," Phys. Med. **31**(3), 301–306 (2015).

[18]S. Serrano-Zabaleta, E. Millán-Cebrián, S. Calvo-Carrillo, V. Alba-Escorihuela, A. Garcia-Romero, P. Ortega-Pardina, and M. Canellas-Anoz, "Study ot the MLC leaves positioning by means of an adjacent segments test and its influence in the IMRT clinical dosimetry," *IV Congreso Conjunto SEFM-SEPR Val, 2015.*

[19]J. E. Bayouth, "Siemens multileaf collimator characterization and quality assurance approaches for intensity-modulated radiotherapy," Int. J. Radiat. Oncol., Biol., Phys. **71**(1), S93–S97 (2008).

[20]A. García-Romero, M. Canellas-Anoz, and D. Lardies-Fleta, "Desarrollo y verificación Monte Carlo de un algoritmo de superposición de cono colapsado para cálculo de haces de fotones en radioterapia," Rev. Fís. Méd. **10**(3), 187–198 (2009).

[21]S. M. J. J. G. Nijsten, W. J. C. van Elmpt, M. Jacobs, B. J. Mijnheer, A. L. A. J. Dekker, P. Lambin, and A. W. H. Minken, "A global calibration model for a-Si EPIDs used for transit dosimetry," Med. Phys. **34**(10), 3872–3884 (2007).

[22]O. Ripol-Valentin, A. Garcia-Romero, A. Hernandez-Vitoria, J. Jiménez-Albericio, J. Cortes-Rodicio, E. Millan-Cebrian, P. Ruiz-

Manzano, and M. Canellas-Anoz, "Caracterización dosimétrica de un dispositivo electrónico de imagen portal (EPID) y desarrollo de un modelo simple de dosimetría portal (Dosimetric characterization of an electronic portal imaging device (EPID) and development of a portal dosimetry sim)," Rev. Fís. Méd. **11**(3), 199–210 (2010).

[23]S. Devic, J. Seuntjens, E. Sham, E. B. Podgorsak, C. R. Schmidtlein, A. S. Kirov, and C. G. Soares, "precise radiochromic film dosimetry using a flat-bed document Scanner," Med. Phys. **32**(7), 2245–2253 (2005).

[24]L. Paelinck, W. De Neve, and C. De Wagter, "Precautions and strategies in using a commercial flatbed scanner for radiochromic film dosimetry," Phys. Med. Biol. **52**(1), 231–242 (2007).

[25]M. Wendling, L. N. McDermott, A. Mans, J.-J. Sonke, M. van Herk, and B. J. Mijnheer, "A simple backprojection algorithm for 3D *in vivo* EPID dosimetry of IMRT treatments," Med. Phys. **36**(7), 3310–3321 (2009).

[26]H. Lin, S. Huang, X. Deng, J. Zhu, and L. Chen, "Comparison of 3D anatomical dose verification and 2D phantom dose verification of IMRT/VMAT treatments for nasopharyngeal carcinoma," Radiat. Oncol. **9**(1), 71 (7pp.) (2014).