ONLINE FIRST

This is a provisional PDF only. Copyedited and fully formatted version will be made available soon.

REPORTS OF PRACTICAL ONCOLOGY AND RADIOTHERAPY

ISSN: 1507-1367

e-ISSN: 2083-4640

Forecasting model for qualitative prediction of the results of patient-specific quality assurance based on planning and complexity metrics and their interrelations. Pilot study

Authors: Tomasz Piotrowski, Adam Ryczkowski, Petros Kalendralis, Marcin Adamczewski, Piotr Sadowski, Barbara Bajon, Marta Kruszyna-Mochalska, Agata Jodda

DOI: 10.5603/rpor.101093

Article type: Research paper

Published online: 2024-06-18

Forecasting model for qualitative prediction of the results of patient-specific quality assurance based on planning and complexity metrics and their interrelations. Pilot study

Running title: Forecasting model for qualitative prediction of the PSQA

10.5603/rpor.101093

Tomasz Piotrowski[1–3], Adam Ryczkowski[1, 2], Petros Kalendralis[4], Marcin Adamczewski[3], Piotr Sadowski[1], Barbara Bajon[2], Marta Kruszyna-Mochalska[1, 2], Agata Jodda[2]

[1]*Department of Electroradiology, Poznan University of Medical Sciences, Poznan, Poland*
[2]*Department of Medical Physics, Greater Poland Cancer Centre, Poznan, Poland*
[3]*Department of Biomedical Physics, Adam Mickiewicz University, Poznan, Poland*
[4]*Department of Radiation Oncology (Maastro), GROW School for Oncology and Reproduction, Maastricht University Medical Centre, Maastricht, The Netherlands*

Corresponding Author: Tomasz Piotrowski, PhD, Department of Medical Physics, Greater Poland Cancer Centre, Garbary 15, 61–866 Poznan, Poland. e-mail: tomasz.piotrowski@me.com

Abstract

Background: The purpose was to analyse the interrelations between planning and complexity metrics and gamma passing rates (GPRs) obtained from VMAT treatments and build the forecasting models for qualitative prediction (QD) of GPRs results.

Materials and method: 802 treatment arcs from the plans prepared for the head and neck, thorax, abdomen, and pelvic cancers were analysed. The plans were verified by portal dosimetry and analysed twice using the gamma method with 3%|2mm and 2%|2mm acceptance criteria. The tolerance limit of GPR was 95%. Red, yellow, and green QDs were established for GPR examination. The interrelations were examined, as well as the analysis of effective differentiation of QD. Three models for QD forecasting based on discriminant analysis (DA), random decision forest (RDF) methods, and the hybrid model (HM) were built and evaluated.

Results: Most of the interrelations were small or moderate. The exception is correlations of the join function with the average number of monitor units per control point (R = 0.893) and the beam aperture with planning target volume (R = 0.897). While many metrics allow for the effective separation of the QDs from each other, the study shows that predicting the values of the QD is possible only through multi-component forecasting models, of which the HM is the most accurate (0.894).

Conclusion: Of the three models explored in this study, the HM, which uses DA methods to predict red QD and RDF methods to predict green and yellow QDs, is the most promising one.

Key words: complexity; plan metrics; PSQA; machine learning; forecasting models

Introduction

For intensity-modulated radiation therapy (IMRT) or volumetric modulated arc therapy (VMAT), once the parameters of beam geometry are established and the treatment energy is chosen, the dose distribution is inversely optimised by the radiation field shaping devices, such as multi-leaf collimator. While inverse optimisation allows for the streamlined creation of advanced treatment plans, its trial-and-error nature can result in sub-optimal and inconsistent treatment plan quality. In some situations, the optimisation process leads to obtaining plans with a high level of complexity. This complexity can approach the limit of the accuracy of the dose calculation model, the precision of the treatment delivery device, or both [1].

Patient-specific quality assurance (PSQA) is an essential clinical step to ensure the treatment plans can be delivered as intended and to verify the treatment planning systems (TPS) dose computation. The PSQA protocols employ a physical measurement device to compare this measurement with the TPS-calculated dose. The gamma index, which combines criteria of both per cent dose difference (DD) and distance-to-agreement (DTA), is the most common method of evaluating the concordance of the measured and calculated dose [2]. The prevalent method for evaluating PSQA is assessing the gamma passing rate (GPR). The GPR signifies the percentage of measurement points that successfully meet the specified gamma index criterion. The American Association of Physicists in Medicine (AAPM) TG 218 report recommended 95% of GPR as the tolerance limit under a 3%|2 mm gamma criterion checked globally [3]. While the AAPM recommendations concern conventional fractionation schemes for the stereotactic or radiosurgery treatment, when the tumour sizes are significantly smaller, and the higher fraction doses are delivered, there are no clear recommendations for the

gamma criterion [4, 5]. Only a suggestion to tighten the gamma criterion for verification of this kind of treatment is posted in the AAPM report [3]. Due to this, our institute uses a 2%|2 mm criterion measured in the local mode.

Advantages of VMAT relative to traditional IMRT include significantly faster and more efficient treatment delivery, though these advantages come at the expense of additional plan complexity [6]. Decreasing the delivery time of treatment sessions for every patient treated by VMAT saves extra time on the accelerator during the day, thus increasing the number of patients treated daily. However, the increased number of patients also means more PSQA verifications which, in the classic form, require access to the accelerator for gathering the measurement data to compare them with the planned data. Freeing time on the accelerator needed for PSQA measurements justifies the search for software-based QA protocols that could replace the traditional PSQA procedures. These studies focus on searching complexity metrics of the treatment and constructing artificial intelligence or machine learning models containing planning and complexity data to forecast the potential failure of PSQA results [7, 8].

While complexity metrics add to the understanding of the complexity of treatment plans, the current perception is that PSQA scores cannot be predicted based on a single complexity metric [9]. The earlier studies focused on models incorporating planning and complexity metrics, demonstrating the viability of employing machine learning algorithms to predict PSQA outcomes [10-13]. However, each machine learning model depends on the characteristics and quality of available data, and each PSQA prediction involves the combination of technologies, the choice of machine learning model, and clinical protocols used for optimising VMAT treatment plans, which can vary across institutions. Current studies where machine learning models were developed tried to forecast the GPR results of PSQA in a quantitative form. In our opinion, qualitative information acquired in the planning stage is also a helpful tool to inform the dosimetrist whether the constructed plan meets the gamma criteria set according to the technique used (i.e. conventional or stereotactic fractionation) that, regarding our institutional protocols, are 3%|2 mm measured in the global mode and 2%|2 mm measured in the local mode.

Therefore, this work explored the interrelations between planning and complexity metrics and GPR results obtained from routinely realised VMAT treatments in our institution. Additionally, three multicomponent models were tested for further modelling GPR results in the qualitative form.

Materials and methods

The study is based on the retrospective anonymised analysis approved by the local Bioethics Committee at the Poznan University of Medical Sciences. All examinations have been performed following the Committee guidelines and the Declaration of Helsinki [14]. The study includes the original studies conducted upon patients' informed consent in writing due to the standard institution protocol. The study is based on unsponsored, single-institutional studies using the database collected from January to May 2022. All data have been anonymised, and the examined patients cannot be identified. There were 802 treatment arcs extracted from 378 volumetric modulated arc therapy (VMAT) treatment plans. Forty-six plans contained three arcs, and 332 plans had two arcs. The plans were created and realised for patients with cancer localised in the head and neck (HN; 192 arcs), thorax (THX; 191 arcs) and abdomen and pelvic (AP; 419 arcs) regions. Detailed locations are provided in the supplementary data (Tab. S1).

All plans were prepared using the 6 MeV photon energy and met our institutional clinical guidelines for dose distribution. The plans were based on conventional as well as stereotactic fractionation schemes. Three hundred and twenty-three plans (671 arcs) were realised conventionally with a flattening filter (6X), and the remaining 55 plans (131 arcs) were realised without a flattening filter (6X-FFF). The maximum planned dose rate (DR) was 600 [MU/min] for 6X and 1400 [MU/min] for 6X-FFF. The dose distribution calculations were performed on CT scans (Somatom Definition AS scanner; Siemens Medical Solution, Erlangen, Germany) using the analytical anisotropic algorithm (AAA) v.16.1.0 implemented in the Eclipse v.16.0 treatment planning system (Varian Medical Systems, Palo Alto, USA).

The plans were realised on the six TrueBeams accelerators (Varian Medical Systems, Palo Alto, USA), four of which were equipped with an electronic portal imaging device (EPID) aS1200 and two with EPID aS1000. Patient-specific quality assurance for every plan was performed using the gamma analysis method. The planned doses were compared with those measured by EPIDs. In general, 521 arcs were measured by EPID aS1200 and 281 arcs by EPID aS1000. For both EPIDs, the same performance algorithm (PDIP) v.16.1.0 was used.

Each arc has been verified twice: in the global mode with criteria of dose differences (DD) equal to 3% and the distance-to-agreement (DTA) 2 mm and in the local mode with DD = 2% and DTA = 2 mm. For both verifications, the threshold was 5% and was normalised to the maximum planned dose. Based on the gamma passing rates (GPR) from both verifications, a three-level qualitative descriptor (QD) was established to score the result of verification (Tab. 1).

Figure 1 shows examples of the comparisons for which, as a result of gamma analysis, three different QDs were granted, i.e. (a) green, (b) yellow and (c) red. Regarding Figure 1, each comparison was performed between the predicted dose (from the treatment plan) and the delivered dose, gathered on the same type of portal, i.e. aS1200. Moreover, every comparison was performed for the doses obtained from the 6X arcs with a 600 [MU/min] dose rate, and the examples included patients with the same location of the treatment area (PA) and comparable planning target volume (PTV).

Figure 2 shows the relations between the GPRs obtained through gamma analyses based on two different criteria for the DD and DTA and realised in two different modes (global and local).

The study's first phase includes an analysis of the interdependence between the selected metrics of the treatment plans, the selected plans' complexity metrics, and the results of its dosimetry verification presented in the form of qualitative descriptors (QD). Mann-Whitney, Kruskal-Wallis with Dunn multiple pairwise comparisons and Spearman tests were used to check these relations with a 0.05 significance level.

The plan metrics included in the study were:

- D_{arc} [Gy] — the part of the fraction dose delivered during the arc irradiation;
- PTV [L] — planning target volume in litres;
- energy (6X or 6X-FFF) — energy, type of radiation and beamforming technology;
- area — the PTV location: HN, THX and AP.

The complexity metrics used in the study were:

- BA, BI and BM — beam aperture, intensity, and modulation, respectively [15];
- MU/Gy – monitor units [16];
- aMU/CP and sdMU/CP — the average number of monitor units in Gy per control point during the arc irradiation (aMU/CP) and the corresponding standard deviation (sdMU/CP) [17];
- aDR and sdDR — the average normalised dose rate during the arc irradiation (aDR) and the corresponding standard deviation (sdDR) [18];
- aGS and sdGS — the average normalised speed of the gantry movement during the arc irradiation (aGS) and the corresponding standard deviation (sdGS) [18];
- Join function (ϑ) — empirically determined function representing the relationship between aDR and aGS.

All plan and complexity metrics listed above were extracted automatically from the plans dicom files by our script written in Python using the SciPy library [19].

In contrast to complexity metrics listed from (a) to (e) that were first introduced by other authors [15–18], the join function (ϑ) is our empirically determined function by the nonlinear estimation method that describes the relation between the dose rate and the gantry speed for volumetric modulated arc therapy.

The relations visualised in Figure 3 may be expressed by the formula:

$$\vartheta = \left[aDR + \left(\frac{1}{aGS} \right)^{\frac{1}{2.6}} \right] - 1$$

The join function ranges from 0 to 2. For the values from 0 to 1 of the function, aGS is near the maximum available speed (~1), and aDR that ranges from 0 to 1 plays a predominant role in the function. When aDR obtains 1, which is equal to the maximum available planned dose rate, the proper dose delivery starts to be controlled by aGS, decreasing from 1 to 0, and as a result, aGS starts to play a predominant role in the function.

In the study's second phase, based on the treatment plans and the complexity metrics, the predictive models of the qualitative descriptors of the dosimetry verifications were created and examined. Two methods were chosen. The first was a probabilistic parametric classification technique called discriminant analysis (DA), and the second was a machine learning, random decision forest (RDF) model. The DA is a popular statistical technique to classify observations into nonoverlapping groups based on determining a linear or quadratic equation constructed from one or more continuous or categorical predictor variables to predict which group the case belongs to [20]. The RDF is a classifier that evolves from the decision trees model - a predictive model expressed as a recursive partition of the feature space to subspaces that constitute a basis for prediction. A random forest is an ensemble method that combines multiple decision trees through bagging. Bagging involves creating multiple subsets of the original dataset through random sampling (with replacement) and training a decision tree on each subgroup. The final prediction is an average or majority vote of predictions from individual trees. It is used to overcome the overfitting problem of one decision tree by reducing variance. The RDF enables many weak or weakly correlated classifiers to form a robust classifier [21].

All plan and complexity metrics explored in the study's first phase were included to build DA, RDF, and hybrid models. The hybrid model assumed two steps of the prediction

procedure - the first, where the DA model was used to predict red QD and the second, where the RDF model predicted green and yellow QD. The models were compared by accuracy and the number of correct classifications and misclassifications. The proper classification related to the different QD values for every model was studied, including the sensitivity and specificity of the models to forecast specified QD. The accuracy of the models and the sensitivity/specificity of the models to forecast specified QD were computed by the formulas [22]:

$$Accuracy = (TP+TN)/(TP+TN+FP+FN),$$

$$Sensitivity = TP/(TP+FN),$$

$$Specificity = TN/(TN+FP),$$

the TP, FP, TN, and FN are true and false positive observations and true and false negative observations, respectively. Both models were constructed and tested using XLSTAT software (Addinsoft, New York, USA). Training and validation groups used for models were the same and contained 642 and 160 treatment arcs, respectively (i.e. 80%/20% split). Data were split using a stratified technique based on the distribution of QD of GPRs to guarantee that the testing set was representative of the overall population of QD of GPRs (Fig. 4).

Results

Figure 5 shows the percentage of observations grouped by QDs (green, yellow, and red) and related to (a) the area of the irradiation, (b) detector type, and (c) energy used. The distribution of the QDs was different for the pelvis and abdomen (PA) area from that for the thorax (THX) or head and neck (HaN) areas (Kruskal-Wallis, $p < 0.001$). Better results of the QDs distribution were observed for the newest EPID type (aS1200) than for the aS1000 type (Mann-Whitney, $p < 0.001$). Almost all QDs for 6X-FFF were classified as green. Different distribution was for 6X (Mann-Whitney, $p < 0.001$), where yellow and red QDs were noted, too.

The 6X-FFF arcs were characterised in general by a high fraction dose per arc (D_{arc}) and were used mainly in stereotactic treatment (75.5% of all 6X-FFF arcs). The requirements of the stereotactic treatment link these results with the records where small PTV and, consequently, small beam apertures (BA) and a high number of monitor units per control

point (aMU/CP) were used. Examining the interdependence between plan and complexity metrics shows many statistically significant correlations. Nevertheless, it should be noted that many of them are small, fair, or moderate [23]. Almost perfect correlations were observed between PTV and BA ($R = 0.897$, $p < 0.001$), and aMU/CP and results of the joint function (ϑ) ($R = 0.893$, $p < 0.001$). The rest of the detailed results are presented in supplementary data (Tab. S2).

Figure 6 shows the interdependence between aMU/CP and ϑ. The data are presented as two trend lines determined by the energy parameter (6X or 6X-FFF).

Analysis of the proportion of the QDs of dosimetry verification results related to the complexity and quantitative plan metrics values shows that BA, aMU/CP, MU/Gy, aDR, and ϑ effectively differentiate all three QDs. The PTV, aGS, sdGS and sdMU/CP effectively separate green from the yellow and red QDs and do not differentiate the yellow from red. The BI and BM allow separating green from yellow QDs. For the rest of the parameters, the QD differentiation was ineffective. Figure 7 shows the results of QD differentiation for selected parameters. Table 2 shows the p-values obtained from the Dunn multiple pairwise comparisons performed during the Kruskal-Wallis analysis of qualitative descriptor differentiation by plan and complexity metrics.

Higher accuracy of the model was observed when the RDF method was used rather than the DA method (0.875 vs. 0.550). The wrong prediction of the green and yellow QDs caused the relatively weak accuracy of the DA model. As many as 70 green QDs (from all 108 green QDs in the validation set) were classified by the DA model as yellow. It causes weak results in the sensitivity of the green QD prediction (0.352) and the specificity of the yellow QD prediction (0.381). While the prediction of the green and yellow QDs by the RDF model was better than the DA model, the prediction of red QD was better for the DA model. While the DA model correctly predicted all five red QDs, the RDF model did it only for two, which strongly affected the sensitivity of prediction for these QDs (1.000 for DA vs. 0.400 for RDF). We introduce a hybrid model in which, in the first phase, the DA model is used to predict red QDs, and then, in the second phase, the prediction of green and yellow QDs is based on the RDF model. The constructed hybrid model has the highest accuracy and the best average sensitivity and specificity values (Table 3). The confusion matrices obtained for training validation sets are presented in the supplementary data (Tables S3-S7).

Discussion

It is known that the quality of dose distributions in plans is frequently independent of planning complexity [24], and comparable dose distributions can be attained through treatment plans of varying levels of complexity due to the potential introduction of unnecessary intricacy through inverse optimisation [25, 26]. For these rationales, numerous researchers have advocated the integration of complexity metrics into the cost function utilised by optimisation algorithms [26–28]. In this study, we selected complexity metrics that are easy to extract from the TPS at the dose optimisation and calculation stage. By examining the correlations between the complexity metrics, plan metrics and the PSQA scores, we confirmed previous literature findings [29, 30] that many complexity metrics correlated. Multiple metrics can account for different uncertainties and sources of plan complexity. As we have shown, the complexity metrics also correlated with the plan metrics, e.g., the intercorrelations presented in Figure 4, between the join function, average monitor units per control point and the energy/beamforming technology that is strictly related in our data to the fractionation scheme (stereotactic/conventional) that is represented by D_{arc} — the fraction dose delivered during the arc irradiation. Nevertheless, as we have shown, predicting PSQA results based on one specified predictor is impossible. Therefore, in contrast to the ideas that assumed the usage of these indices on the optimisation stage to reduce plan complexity, we used them with plan metrics to construct the forecasting model that provides qualitative information on the planning stage on further results of PSQA. While other works that focused on the forecasting models show the results of quantitative model development [10–13] that are intended to replace the PSQA procedures, our concept assumes the introduction of a support tool that will provide qualitative information during the treatment plan preparation about its feasibility by the treatment machine. Our study shows that the most effective forecasting of the QD of the GPR results was obtained for the hybrid model based on the DA and RDF models. When implemented commercially, such a solution will enable the effective use of information generated during the treatment planning process to finally create a plan that can be implemented on the therapeutic machine with the accuracy adopted in the institution. This solution should be pre-configured and dependent on the institution-specific data. It means that the team developing the model should decide which DD|DTA criteria of gamma analysis will be included to generate green, yellow, and red descriptors. Moreover, the data on which the model will be trained should be gathered in this institution for specific dose development and PSQA methods. As shown, while we used one PSQA method (EPID dosimetry), the GPR results differed by the EPID model. Therefore, a specific characteristic of the dosimetry tool used during the PSQA should also be included.

The presented study is of a pilot nature. Our findings provide the basis for further model development to increase its accuracy, which currently allows correct QD prediction at 89.4%.

Conclusion

While we found a lot of statistically significant interrelations between metrics describing the plan and its complexity, they were small, fair or moderate. Only the correlations between ϑ and aMU/CP and the BA and PTV were almost perfect (R = 0.893 and R = 0.897, respectively).

Analysis of the proportion of the QDs related to the values of the complexity and plan metrics shows that a lot of these features allow for the effective separation of each of the descriptors (BA, aMU/CP, MU/Gy, aDR and ϑ) or to separate one descriptor from two other descriptors (PTV, aGS, sdGS, sdMU/CP, BI, BM).

The study shows that predicting GPR results based on one specified predictor is problematic. However, multi-component forecasting models became possible. Analysis of the efficacy of the DA, RDF and hybrid models shows that a hybrid model, which uses DA methods to predict red QD and RDF methods to predict green and yellow QDs, is the most accurate (0.894 compared to 0.875 for the RDF model and 0.550 for the DA model).

Data availability

The datasets analysed during the study are available from the corresponding author on request.

Author contributions

T.P. — the concept of the study, literature analysis, writing the manuscript, data analysis and models training and validation; A.R.— the concept of the study, SciPy coding, data export and analysis and writing the manuscript; P.K.— literature analysis and writing the manuscript, supervision of training and validation of the models; M.A.— literature analysis, collecting and export of complexity metrics data; P.S.— collecting and export of plan metrics data; B.B. and M.K-M — collecting and analysis of the PSQA data; A.J. – the concept of the study, supervision of the data collecting and export, manuscript writing.

Conflict of interests

Authors declare no conflict of interests

Funding

None declared.

References

1. Malicki J, Piotrowski T, Guedea F, et al. Treatment-integrated imaging, radiomics, and personalised radiotherapy: the future is at hand. Rep Pract Oncol Radiother. 2022; 27(4): 734–743, doi: 10.5603/RPOR.a2022.0071, indexed in Pubmed: 36196410.

2. Low DA, Harms WB, Mutic S, et al. A technique for the quantitative evaluation of dose distributions. Med Phys. 1998; 25(5): 656–661, doi: 10.1118/1.598248, indexed in Pubmed: 9608475.

3. Miften M, Olch A, Mihailidis D, et al. Tolerance limits and methodologies for IMRT measurement-based verification QA: Recommendations of AAPM Task Group No. 218. Med Phys. 2018; 45(4): e53–e83, doi: 10.1002/mp.12810, indexed in Pubmed: 29443390.

4. Kim JI, Park SY, Kim HJ, et al. The sensitivity of gamma-index method to the positioning errors of high-definition MLC in patient-specific VMAT QA for SBRT. Radiat Oncol. 2014; 9: 167, doi: 10.1186/1748-717X-9-167, indexed in Pubmed: 25070065.

5. Alharthi T, Pogson EM, Arumugam S, et al. Pre-treatment verification of lung SBRT VMAT plans with delivery errors: Toward a better understanding of the gamma index analysis. Phys Med. 2018; 49: 119–128, doi: 10.1016/j.ejmp.2018.04.005, indexed in Pubmed: 29685425.

6. Teoh M, Clark CH, Wood K, et al. Volumetric modulated arc therapy: a review of current literature and clinical use in practice. Br J Radiol. 2011; 84(1007): 967–996, doi: 10.1259/bjr/22373346, indexed in Pubmed: 22011829.

7. Chiavassa S, Bessieres I, Edouard M, et al. Complexity metrics for IMRT and VMAT plans: a review of current literature and applications. Br J Radiol. 2019; 92(1102): 20190270, doi: 10.1259/bjr.20190270, indexed in Pubmed: 31295002.

8. Chan MF, Witztum A, Valdes G. Integration of AI and Machine Learning in Radiotherapy QA. Front Artif Intell. 2020; 3: 577620, doi: 10.3389/frai.2020.577620, indexed in Pubmed: 33733216.

9. Hernandez V, Hansen CR, Widesott L, et al. What is plan quality in radiotherapy? The importance of evaluating dose metrics, complexity, and robustness of treatment plans. Radiother Oncol. 2020; 153: 26–33, doi: 10.1016/j.radonc.2020.09.038, indexed in Pubmed: 32987045.

10. Interian Y, Rideout V, Kearney VP, et al. Deep nets vs expert designed features in medical physics: An IMRT QA case study. Med Phys. 2018; 45(6): 2672–2680, doi: 10.1002/mp.12890, indexed in Pubmed: 29603278.

11. Lam D, Zhang X, Li H, et al. Predicting gamma passing rates for portal dosimetry-based IMRT QA using machine learning. Med Phys. 2019; 46(10): 4666–4675, doi: 10.1002/mp.13752, indexed in Pubmed: 31386761.

12. Granville DA, Sutherland JG, Belec JG, et al. Predicting VMAT patient-specific QA results using a support vector classifier trained on treatment plan characteristics and linac QC metrics. Phys Med Biol. 2019; 64(9): 095017, doi: 10.1088/1361-6560/ab142e, indexed in Pubmed: 30921785.

13. Ono T, Hirashima H, Iramina H, et al. Prediction of dosimetric accuracy for VMAT plans using plan complexity parameters via machine learning. Med Phys. 2019; 46(9): 3823–3832, doi: 10.1002/mp.13669, indexed in Pubmed: 31222758.

14. WMA 2013. WMA Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. World Medical Association. https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/ (February, 01 2024.).

15. Du W, Cho SH, Zhang X, et al. Quantification of beam complexity in intensity-modulated radiation therapy treatment plans. Med Phys. 2014; 41(2): 021716, doi: 10.1118/1.4861821, indexed in Pubmed: 24506607.

16. Mohan R, Arnfield M, Tong S, et al. The impact of fluctuations in intensity patterns on the number of monitor units and the quality and accuracy of intensity modulated radiotherapy. Med Phys. 2000; 27(6): 1226–1237, doi: 10.1118/1.599000, indexed in Pubmed: 10902551.

17. Shen L, Chen S, Zhu X, 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(2): 207–215, doi: 10.1093/jrr/rrx100, indexed in Pubmed: 29415196.

18. Nicolini G, Clivio A, Cozzi L, et al. On the impact of dose rate variation upon RapidArc implementation of volumetric modulated are therapy. Med Phys. 2011; 38(1): 264–271, doi: 10.1118/1.3528214, indexed in Pubmed: 21361195.

19. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0 Contributors, SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020; 17(3): 261–272, doi: 10.1038/s41592-019-0686-2, indexed in Pubmed: 32015543.

20. Dhamnetiya D, Goel MK, Jha RP, et al. How to Perform Discriminant Analysis in Medical Research? Explained with Illustrations. J Lab Physicians. 2022; 14(4): 511–520, doi: 10.1055/s-0042-1747675, indexed in Pubmed: 36531553.

21. Breiman L. Random forests. Mach Learning. 2001; 45(1): 5–32, doi: 10.1023/A:1010933404324..

22. Li J, Wang Le, Zhang X, et al. Machine Learning for Patient-Specific Quality Assurance of VMAT: Prediction and Classification Accuracy. Int J Radiat Oncol Biol Phys. 2019; 105(4): 893–902, doi: 10.1016/j.ijrobp.2019.07.049, indexed in Pubmed: 31377162.

23. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33(1): 159–174, indexed in Pubmed: 843571.

24. Jurado-Bruggeman D, Hernández V, Sáez J, et al. Multi-centre audit of VMAT planning and pre-treatment verification. Radiother Oncol. 2017; 124(2): 302–310, doi: 10.1016/j.radonc.2017.05.019, indexed in Pubmed: 28687395.

25. Craft D, Süss P, Bortfeld T. The tradeoff between treatment plan quality and required number of monitor units in intensity-modulated radiotherapy. Int J Radiat Oncol Biol Phys. 2007; 67(5): 1596–1605, doi: 10.1016/j.ijrobp.2006.11.034, indexed in Pubmed: 17394954.

26. Younge KC, Matuszak MM, Moran JM, et al. Penalization of aperture complexity in inversely planned volumetric modulated arc therapy. Med Phys. 2012; 39(11): 7160–7170, doi: 10.1118/1.4762566, indexed in Pubmed: 23127107.

27. McNiven AL, Sharpe MB, Purdie TG. A new metric for assessing IMRT modulation complexity and plan deliverability. Med Phys. 2010; 37(2): 505–515, doi: 10.1118/1.3276775, indexed in Pubmed: 20229859.

28. Matuszak MM, Larsen EW, Fraass BA. Reduction of IMRT beam complexity through the use of beam modulation penalties in the objective function. Med Phys. 2007; 34(2): 507–520, doi: 10.1118/1.2409749, indexed in Pubmed: 17388168.

29. Hernandez V, Saez J, Pasler M, et al. Comparison of complexity metrics for multi-institutional evaluations of treatment plans in radiotherapy. Phys Imaging Radiat Oncol. 2018; 5: 37–43, doi: 10.1016/j.phro.2018.02.002, indexed in Pubmed: 33458367.

30. Antoine M, Ralite F, Soustiel C, et al. Use of metrics to quantify IMRT and VMAT treatment plan complexity: A systematic review and perspectives. Phys Med. 2019; 64: 98–108, doi: 10.1016/j.ejmp.2019.05.024, indexed in Pubmed: 31515041.

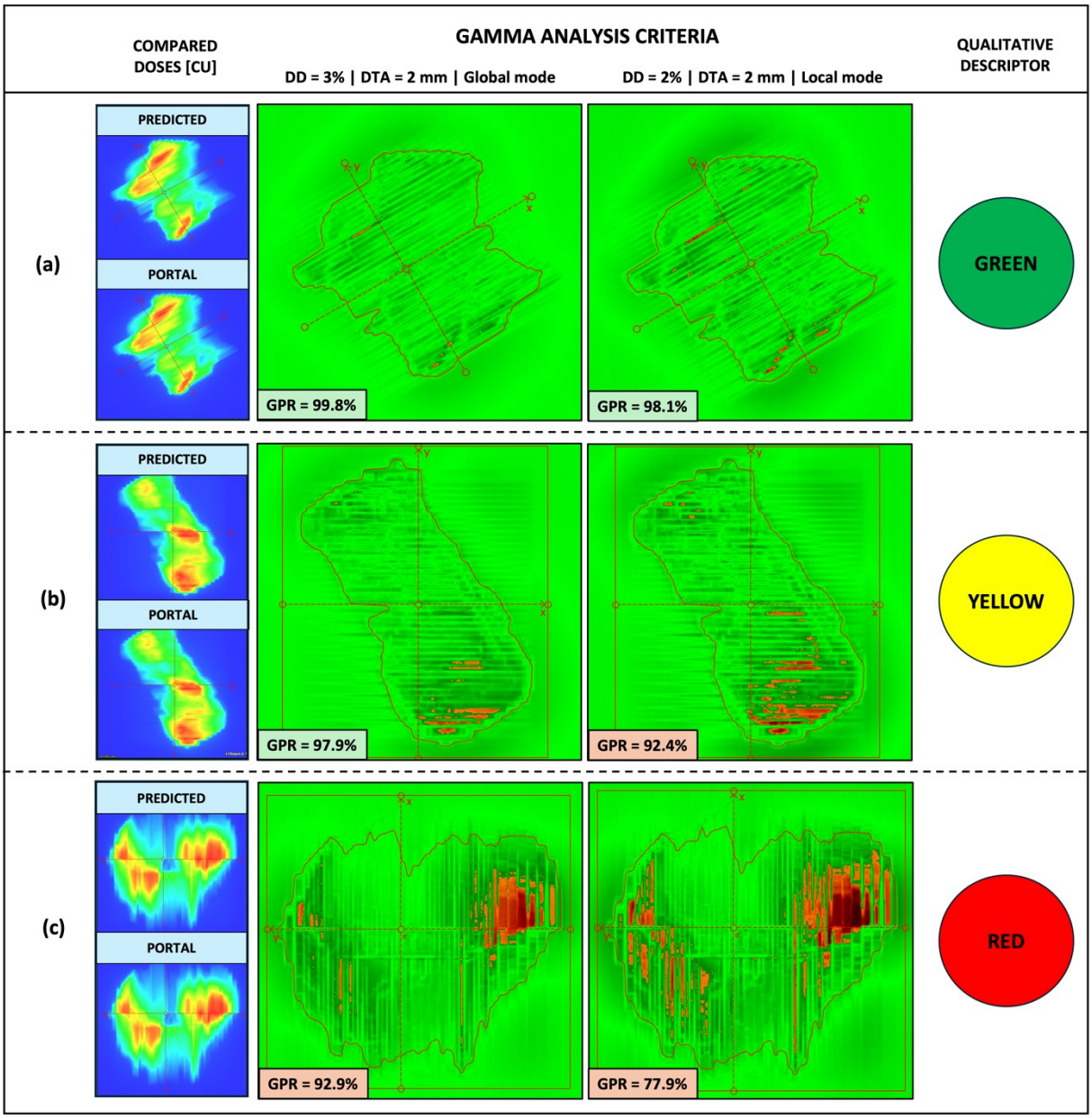

Figure 1. The examples of the comparisons for which, as a result of gamma analysis, three different QDs were granted, i.e. green (**A**), yellow (**B**) and red (**C**). GPR — gamma passing rate; DD — dose difference; DTA — distance-to-agreement

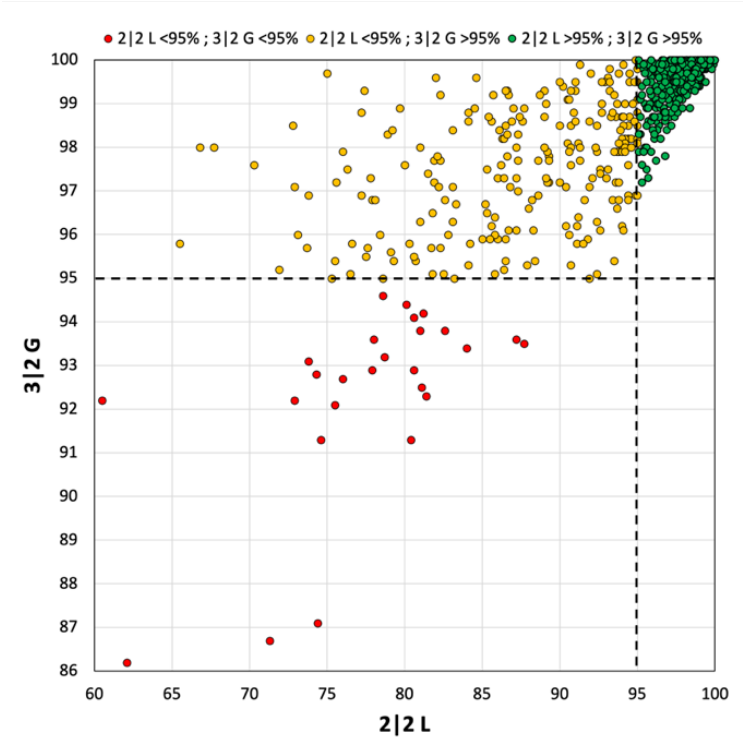

Figure 2. Relation between the gamma passing ratios (GPRs) obtained through gamma analyses based on two different criteria for the DD and DTA and realised in two different modes (global and local). 3|2 G — the results of GPRs for the gamma analysis based on criteria DD = 3% and DTA = 2 mm and realised in a global mode; 2|2 L — the results of GPRs for the gamma analysis based on criteria DD = 2% and DTA = 2 mm and realised in a local mode

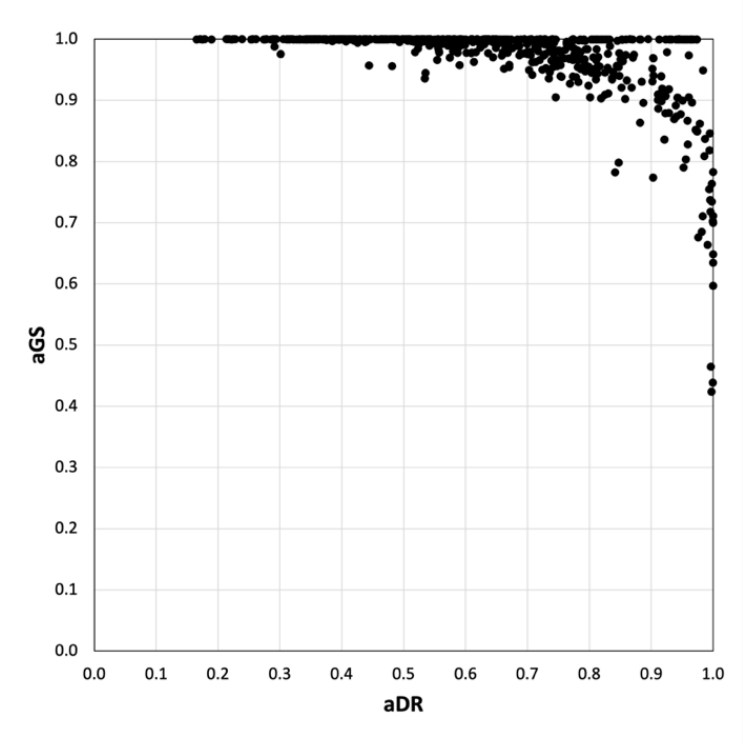

Figure 3. Relation between the average normalised dose rate (aDR) and the average normalised speed of the gantry movement (aGS) during the arc irradiation

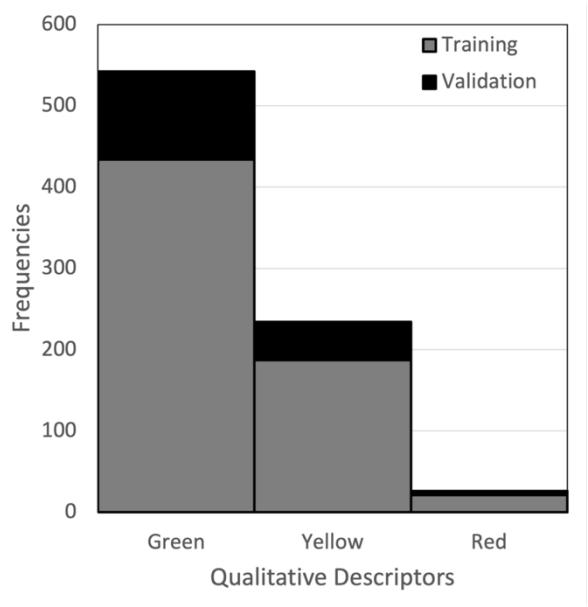

Figure 4. The number of observations grouped by qualitative descriptors and related to the training set (grey) and validation set (black)

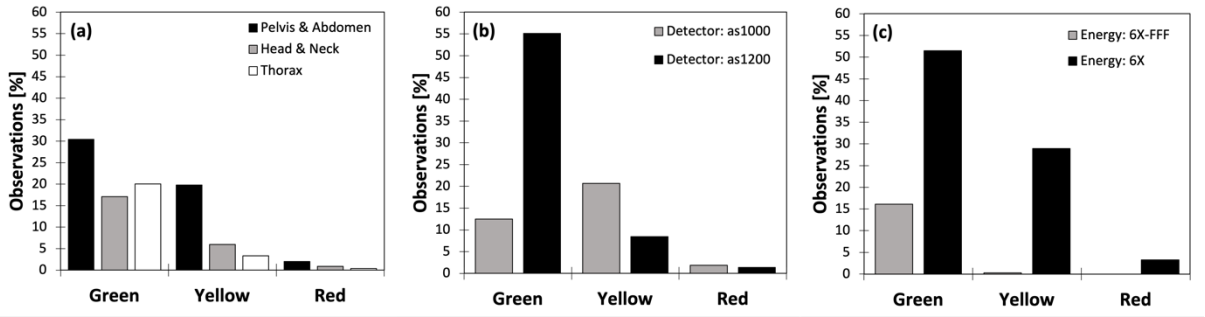

Figure 5. The percentage of observations grouped by qualitative descriptors (green, yellow, and red) and related to area of the irradiation(**A**), detector type (**B**), and energy used (**C**)

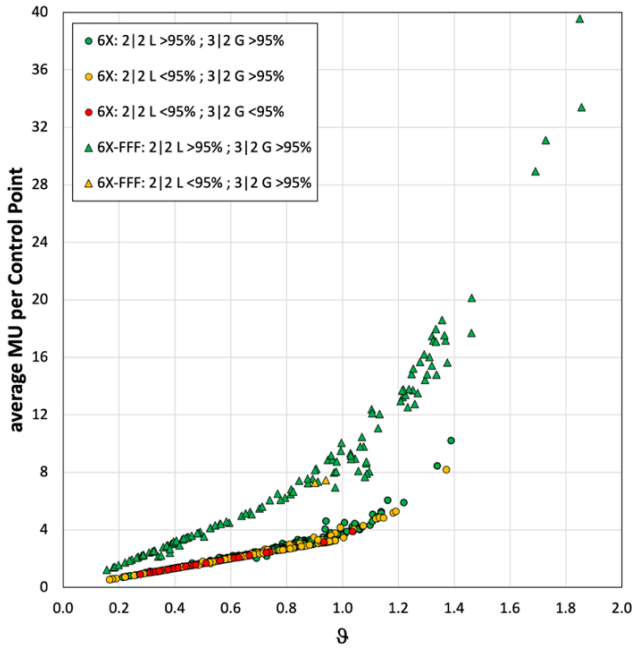

Figure 6. Relations between average monitor units (MU) per control point and the join function (ϑ). The data relating to 6X-FFF presented as a triangle and 6X as a circle. The qualitative descriptor of dosimetry verification is determined by the colours green, yellow and red

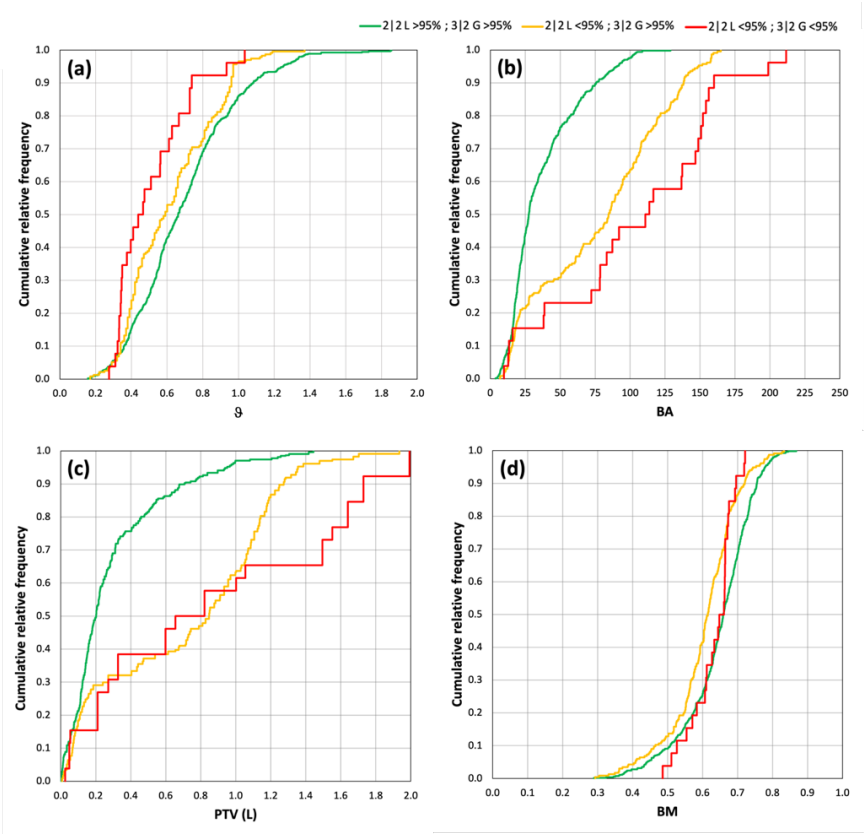

Figure 7. Relation between cumulative frequency of the qualitative descriptors of the dosimetry verification results and the values of join function (ϑ) (**A**), beam aperture (BA) (**B**), planning target volume (PTV) (**C**) and beam modulation (BM) parameter (**D**)

Table 1. Three-level qualitative descriptor based on the gamma passing rates results obtained for different criteria of gamma analysis.

Qualitative Descriptor	**Gamma Passing Rate for specified gamma analysis criteria**	
	3% | 2 mm | Global mode	**2% | 2 mm | Local mode**
Green	≥ 95%	≥ 95%
Yellow	≥ 95%	< 95%
Red	< 95%	< 95%

Table 2. The p-values obtained from the Dunn multiple pairwise comparisons performed during the Kruskal-Wallis analysis of qualitative descriptor differentiation by plan and complexity metrics. Analysis performed at 0.05 significance level.

Parameter	**Green vs. Yellow**	**Green vs. Red**	**Yellow vs. Red**
D_{arc}	0.655	0.966	0.899
PTV	< 0.001	< 0.001	0.537
BA	< 0.001	< 0.001	0.042
BI	0.011	0.084	0.474
BM	< 0.001	0.271	0.233
aMU/CP	< 0.001	< 0.001	0.027
sdMU/CP	< 0.001	< 0.001	0.213
MU/Gy	< 0.001	< 0.001	0.033
aDR	0.003	< 0.001	0.021
sdDR	0.058	0.141	0.477
aGS	< 0.001	< 0.001	0.484
sdGS	< 0.001	< 0.001	0.458
ϑ	0.001	< 0.001	0.022

D_{arc} — the part of the fraction dose delivered during the arc irradiation; PTV — planning target volume; BA — beam aperture; BI — beam intensity; BM — beam modulation; aMU/CP — the average number of monitor units in Gy per control point during the arc irradiation; sdMU/CP — the corresponding standard deviation; aDR — the average normalised dose rate during the arc irradiation; sdDR — the corresponding standard deviation; aGS — the average normalised speed of the gantry movement during the arc irradiation; sdGS — the corresponding standard deviation; ϑ — join function

Table 3. Descriptive statistics for the models of discriminant analysis (DA), random decision forest (RDF) and hybrid model, and the values of sensitivity and specificity from the model related to specified qualitative descriptors (green, yellow, red)

	DA	**RDF**	**Hybrid**
General models statistics			
Accuracy	0.550	0.875	0.894
Correct class	88	140	143

Misclass	72	20	17
Sensitivity | Specificity of Qualitative Descriptors			
Green	0.352 | 0.981	0.944 | 0.808	0.944 | 0.808
Yellow	0.957 | 0.381	0.766 | 0.929	0.766 | 0.947
Red	1.000 | 0.994	0.400 | 0.987	1.000 | 0.994
Averaged	0.770 | 0.785	0.703 | 0.908	0.903 | 0.916

Supplementary File

Table S1. Treatment plans localisations

Region	Localization	Number of arcs
Abdomen and Pelvic	Bladder	12
	Gynaecology	59
	Prostate	228
	Rectum	64
	Stomach	2
	Adrenal	2
	Metastasis to bones	26
	Metastasis to soft tissues	26
	TOTAL	*419*
Thorax	Lung	136
	Oesophagus	29
	Metastasis to bones	26
	TOTAL	*191*
Head and Neck	Laryngopharynx	135
	Oropharynx	45
	Nasopharynx	2
	Metastasis to bones	4
	Brain	6
	TOTAL	*192*

Table S2. Spearman correlation coefficients between plan and complexity metrics. Results in bold are statistically significant at $\alpha = 0.05$

Variables	PTV (L)	D(fr,arc)	MU/Gy	BA	BI	BM	aMU/CP	sdMU/CP	aDR	sdDR	aGS	sdGS	ϑ
PTV (L)	1												
D(fr,arc)	-0.226	1											
MU/Gy	-0.216	-0.195	1										
BA	0.897	-0.195	-0.438	1									
BI	0.628	-0.337	0.146	0.552	1								
BM	0.164	-0.258	0.441	0.012	0.714	1							
aMU/CP	-0.439	0.690	0.292	-0.533	-0.457	-0.184	1						
sdMU/CP	-0.375	0.410	0.231	-0.442	-0.259	-0.010	0.563	1					
aDR	-0.364	0.737	0.199	-0.472	-0.400	-0.144	0.890	0.416	1				
sdDR	0.135	-0.206	-0.002	0.098	0.227	0.325	-0.285	0.332	-0.273	1			
aGS	0.452	-0.496	-0.325	0.572	0.283	-0.035	-0.705	-0.692	-0.731	0.049	1		
sdGS	-0.426	0.452	0.355	-0.554	-0.242	0.092	0.678	0.694	0.698	0.012	-0.983	1	
ϑ	-0.370	0.737	0.208	-0.481	-0.398	-0.135	0.893	0.438	0.999	-0.263	-0.749	0.719	1

Legend:	no correlation	small	fair	moderate	substantial	almost perfect

Table S3. Confusion matrices for the training sample for the DA model

from \ to	Green	Yellow	Red	Total	% correct
Green	160	273	1	434	36.87%
Yellow	7	175	5	187	93.58%
Red	0	0	21	21	100.00%
Total	167	448	27	642	55.45%

Table S4. Confusion matrices for the training sample for the RDF model

From\to	Green	Yellow	Red	Total	% correct
Green	412	21	1	434	94.9
Yellow	38	142	7	187	75.9
Red	4	8	9	21	42.9
Total	454	171	17	642	87.7

Table S5. Confusion matrices for the validation sample for the DA model

From\to	Green	Yellow	Red	Total	% correct
Green	38	70	0	108	35.19%
Yellow	1	45	1	47	95.74%
Red	0	0	5	5	100.00%
Total	39	115	6	160	55.00%

Table S6. Confusion matrices for the validation sample for the RDF model

from \ to	Green	Yellow	Red	Total	% correct
Green	102	6	0	108	94.4
Yellow	9	36	2	47	76.6
Red	1	2	2	5	40.0
Total	112	44	4	160	87.5

Table S7. Confusion matrices for the validation sample for the hybrid model.

From\to	Green	Yellow	Red	Total	% correct
Green	102	6	0	108	94.4
Yellow	10	36	1	47	76.6

Red	0	0	5	5	100.0
Total	112	42	6	160	89.4