Radiotherapy and Oncology 183 (2023) 109628

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original Article

Validation of prediction models for radiation-induced late rectal bleeding: Evidence from a large pooled population of prostate cancer patients ☆

Alessandro Cicchetti [a,b,*], Claudio Fiorino [c], Martin A. Ebert [d,e,f], Jacopo Iacovacci [a,b], Angel Kennedy [e], David J. Joseph [d,f,g], James W. Denham [h], Vittorio Vavassori [i], Gianni Fellin [j], Cesare Cozzarini [k], Claudio Degli Esposti [l], Pietro Gabriele [m], Fernando Munoz [n], Barbara Avuzzi [o], Riccardo Valdagni [a,o,p,1], Tiziana Rancati [a,b,1]

[a] Prostate Cancer Program, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy; [b] Data Science Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy; [c] Medical Physics, San Raffaele Scientific Institute, Milan, Italy; [d] University of Western Australia, Perth, Western, Australia; [e] Radiation Oncology, Sir Charles Gairdner Hospital, Perth, Western, Australia; [f] 5D Clinics, Claremont, Western, Australia; [g] GenesisCare, Perth, Western, Australia; [h] School of Medicine and Public Health, University of Newcastle, New South Wales, Australia; [i] Radiation Oncology, Cliniche Humanitas-Gavazzeni, Bergamo, Italy; [j] Radiation Oncology, Ospedale Santa Chiara, Trento, Italy; [k] Radiation Oncology, San Raffaele Scientific Institute, Milan, Italy; [l] Radiation Oncology, Ospedale Bellaria, Bologna, Italy; [m] Radiation Oncology, Istituto di Candiolo- Fondazione del Piemonte per l'Oncologia IRCCS, Torino, Italy; [n] Radiation Oncology, Azienda Ospedaliera di Aosta, Aosta, Italy; [o] Radiation Oncology, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy; [p] Oncology and Hemato-Oncology, Università degli Studi, Milano, Italy

ARTICLE INFO

Article history:
Received 25 January 2021
Received in revised form 3 February 2023
Accepted 10 March 2023
Available online 18 March 2023

Keywords:
Model validation
NTCP models
Late rectal bleeding
Multivariable models
Prostate cancer

ABSTRACT

Purpose: To validate published models for the risk estimate of grade ≥ 1 (G1+), grade ≥ 2 (G2+) and grade = 3 (G3) late rectal bleeding (LRB) after radical radiotherapy for prostate cancer in a large pooled population from three prospective trials.

Materials and methods: The external validation population included patients from Europe, and Oceanian centres enrolled between 2003 and 2014. Patients received 3DCRT or IMRT at doses between 66–80 Gy. IMRT was administered with conventional or hypofractionated schemes (2.35–2.65 Gy/fr). LRB was prospectively scored using patient-reported questionnaires (LENT/SOMA scale) with a 3-year follow-up.

All Normal Tissue Complication Probability (NTCP) models published until 2021 based on the Equivalent Uniform Dose (EUD) from the rectal Dose Volume Histogram (DVH) were considered for validation.

Model performance in validation was evaluated through calibration and discrimination.

Results: Sixteen NTCP models were tested on data from 1633 patients. G1+ LRB was scored in 465 patients (28.5%), G2+ in 255 patients (15.6%) and G3 in 112 patients (6.8%). The best performances for G2+ and G3 LRB highlighted the importance of the medium–high doses to the rectum (volume parameters n = 0.24 and n = 0.18, respectively). Good performance was seen for models of severe LRB. Moreover, a multivariate model with two clinical factors found the best calibration slope.

Conclusion: Five published NTCP models developed on non-contemporary cohorts were able to predict a relative increase in the toxicity response in a more recent validation population. Compared to QUANTEC findings, dosimetric results pointed toward mid-high doses of rectal DVH. The external validation cohort confirmed abdominal surgery and cardiovascular diseases as risk factors.

 Radiotherapy and Oncology 183 (2023) 109628

Over the last two decades, quantitative information derived by dose and volume relations for rectal bleeding after prostate irradiation has been collected and analyzed. Many prospective trials have investigated the association of patient- and treatment-related parameters with acute and late side effects to optimize radiation therapy planning [1–3]. Based on data coming from these large cohorts, Normal Tissue Complication Probability (NTCP) models have been proposed for mild/moderate/severe late rectal bleeding (LRB) [3–10]. Some of these models also combined dosimetric and clinical information, with the latter acting as a dose–response modifier [3,5,6].

Peer review under responsibility of Faculty of Engineering, AlexandriaUniversity.

* Corresponding author at: Prostate Cancer Program, Fondazione IRCCS Istituto Nazionale dei Tumori, Via Venezian 1, 20133 Milano, Italy.
E-mail address: alessandro.cicchetti@istitutotumori.mi.it (A. Cicchetti).

[1] Equally contributed.

https://doi.org/10.1016/j.radonc.2023.109628
0167-8140/

The accurate application of treatment constraints and the possibility of tuning these values according to the characteristics of each patient are currently considered to be a tangible way to limit radio-induced symptoms. Nevertheless, although the availability of the above-mentioned published NTCP models, validation studies (TRIPOD type 3 and 4 [11]) to establish their applicability and generalizability in populations other than those used for model development are pretty rare and lacking [12,13]. It is worth noting that the purpose of an NTCP model is to provide valid outcome predictions for new patients. In principle, the dataset used to develop a model is not of interest other than to learn for the future. Hence, validation is a crucial aspect of the process that makes predictive models useful for the community [11]. External validation provides a measure of the "generalizability" and "transportability" of the prediction model to populations that are "plausibly related". "Plausibly related" populations can be defined as cohorts that could be slightly different from the one used for model development, e.g. treated at various hospitals, at different dose levels, with further radiotherapy (RT) techniques, in other countries or at different periods. Generalizability and transportability are desired properties from both a scientific and practical perspective. Quantifying the confidence and predictive accuracy of the model's performances provides the decision-maker with the information necessary for making high-consequence decisions. The more often a model is externally validated, and the more diverse these settings are, the more confidence we can gain in the use of the model for prospective decision-making and its possible use in interventional trials.

In this study, we aimed at multiple fully independent validations (other investigators, TRIPOD type 4), including geographic validations (other places), spectrum transportability (wide range of prescription doses), and treatment technique validation (models developed on three dimensional conformal RT, 3DCRT, while validation includes IMRT). The validation dataset consisted of a pooled population from three large prospective trials [14–16], including 1633 patients with 3-year minimum follow-up.

Material and methods

Patient population

Patients enrolled in three high-quality multicentre prospective trials on RT for prostate cancer were considered (time window 2002–2014):

1. TROG 03.04 RADAR (RADAR): a prospective multicentre randomized trial designed (details in [15,17–21]). A secondary endpoint of the trial was the prospective score of radiation-induced toxicity.
2. Airopros0102: a multicentre observational trial designed to prospectively assess the association between clinical and dosimetric features with acute/late rectal toxicity in a population of Italian prostate cancer patients treated with radical radiotherapy (details in [14,22–24]).
3. DUE01: a prospective multicenter observational trial focused on urinary toxicity and erectile dysfunction after radical high-dose RT for prostate cancer [16,25,26]. As a secondary endpoint, a rectal toxicity questionnaire was compiled by patients aiming to validate the models based on data from Airpros0102.

The RADAR and Airopros patients were treated with 3DCRT with prescribed doses ranging between 66 and 80 Gy (median dose 73.2 Gy, Inter-quartile range 70–75 Gy, a daily dose between 1.8–2 Gy/day). Conversely, the DUE01 patients underwent IMRT with conventional (2 Gy/fr, dose range 74–80 Gy) or moderate hypofractionated schedules (2.35–2.70 Gy/fr, physical prescription doses between 70–74 Gy).

Comorbidities (the presence of haemorrhoids, cardiovascular disease, hypertension, diabetes), concomitant/previous loco-regional diseases, use of drugs, previous pelvic/abdominal surgery and type and duration of hormonal therapy were prospectively recorded before treatment.

Details on the radiotherapy volume definitions, planning, treatment modalities, and the distribution of the clinical parameters in three cohorts have been previously reported [14,16,17,27]. The dose-volume histogram (DVH) of the solid ano-rectum follows the same anatomical definition (from the sigmoid junction to the anal verge), as reported in Foppiano et al. [28].

The study here presented was approved by local ethics committees (INT 202/14).

Toxicity endpoint definition and evaluation

All patients underwent a clinical examination at least every 6 months in the first 3 years, other than at the start/end of treatment.

As an added value of our analysis, the procedure required to harmonize the identification of rectal symptoms across studies was straightforward. Indeed, intestinal toxicity was assessed through a patient-reported questionnaire according to the LENT/SOMA (Late Effects of Normal Tissues/Subjective, Objective, Management and Analytic, [29]) scoring systems for late radiation morbidity in all three trials.

Mild, moderate, and severe LRB were considered in the analysis. We grouped the symptoms as follows:

a) grade 1 (G1): bleeding up to twice a week (and the baseline questionnaire indicating no bleeding);
b) grade 2 (G2): bleeding > 2 times/week (patients without bleeding at the baseline questionnaire);
c) grade 3 (G3): daily bleeding was experienced OR need for blood transfusions or laser coagulation procedures (patients without bleeding or G1 bleeding at the baseline questionnaire).

Patients experiencing the event at any time longer than 5 months after RT completion until the 3-year follow-up (even if they recovered) were considered bleeders. This definition was adopted to better compare the results with the "actuarial" definition of bleeding used in the published NTCP models.

Validation of previously published models

We considered 16 NTCP models for the prediction of grade ≥ 1 (G1+), grade ≥ 2 (G2+), and grade = 3 (G3) LRB with and without the inclusion of dose-modifying factors. These were all the models that specifically included the Equivalent Uniform Dose (EUD) from the rectal DVH as a dosimetric descriptor, published in the literature until 2021.

A cohesive definition of the rectum was found that was applicable across the model development studies and to the validation cohort. The majority of the analysis computed the EUD based on the DVH of the solid anorectum. The only exception was for the model developed by Defraene and colleagues, which considered the rectal wall DVH. To this purpose, a data harmonization of the dosimetric parameters from the Dutch cohort was performed based on a study conducted by our group and presented in Supplementary Materials. The aim was to identify converting factors for rectal EUD when computed under different circumstances, i.e. organ definition or contouring (solid vs wall). In SM, we included

Table 1
Details of normal tissue complication probability models considered in the present work for validation in the pooled population: 1a) models without the inclusion of dose-modifying factors, 1b) models with the inclusion dose-modifying factors.

1a)

Reference	N° pts (endpoint rate, %)	Prescribed dose (Gy) RT technique	NTCP Model & OaR	D50 (Gy) best fit (68%CI)	m or k best fit (68%CI)	n best fit (68%CI)
Endpoint: grade 1–2-3 late rectal bleeding						
Gulliford 2012 Fig. 1a [4]	361 (44%)	64–74 Gy 3DCRT	LKB Solid anorectum DVH	59.2 (−9.3,+8.8)	0.29 (−0.29,+0.30)	0.17 (−0.17,+0.30)
Brand 2021 Fig. 1b [9]	2008 (32.6%)	74 Gy IMRT	LKB Solid anorectum DVH	58.8 (−4.6,+7.2)	0.33 (−0.1,+0.14)	0.21 (−0.13,+0.13)
Endpoint: grade 2–3 late rectal bleeding						
Rancati 2004 Fig. 1c [5]	321 (7%)	64–70 Gy 3DCRT	LKB Solid anorectum DVH	75.7 (−1.5,+1.5)	0.14 (−0.01,+0.01)	0.24 (−0.05,+0.05)
Rancati 2011 Fig. 1d [6]	669 (8%)	70–78 Gy 3DCRT	Logit-EUD Solid anorectum DVH	88.9 (−1.3,+1.4)	10.1 (−0.6,+0.6)	0.03 (−0.03,+0.11)
Michalsky 2010 Fig. 1e [3]	Meta-analysis of 4 studies 1503 pts (13.5%)	60–79.2 Gy 3DCRT	LKB Solid anorectum DVH	76.9 (−1.6,+1.6)	0.13 (−0.02,+0.02)	0.09 (−0.03,+0.03)
Tucker 2010 Fig. 1f [7]	1010 (15%)	68.4–79.2 Gy 3DCRT	LKB Solid anorectum DVH	79.1 (−1.9,+2.6)	0.15 (−0.03,+0.04)	0.077 (−0.02,+0.04)
Gulliford 2012 Fig. 1g [4]	361 (15%)	64–74 Gy 3DCRT	LKB Solid anorectum DVH	68.9 (−2.1,+2.1)	0.16 (−0.03,+0.03)	0.18 (−0.07,+0.07)
Brand 2021 Fig. 1h [9]	2008 (14.6%)	74 Gy IMRT	LKB Solid anorectum DVH	75.8 (−7.6,+12.8)	0.27 (−0.13,+0.17)	0.16 (−0.15,+0.18)
Endpoint: grade 3 late rectal bleeding						
Rancati 2004 Fig. 1i [5]	547 (2%) (including 226 post-prostatectomy pts)	64–70 Gy 3DCRT	LKB Solid anorectum DVH	78.6 (−3.7,+3.7)	0.06 (−0.005, +0.005)	0.06 (−0.01,+0.01)
Rancati 2011 Fig. 1j [6]	669 (5%)	70–78 Gy 3DCRT	Logit-EUD Solid anorectum DVH	93.1 (−2.0,+2.0)	9.4 (−0.8,+1.4)	0.05 (−0.04,+0.05)
Defraene 2012 Fig. 1k [8]	512 (6%)	68–78 Gy 3DCRT	LKB Anorectal wall DVH	79.0 (−5.0,+7.5)	0.15 (−0.03,+0.05)	0.18 (−0.09,+0.15)

1b)

Reference	Prescribed dose (Gy) RT technique	NTCP Model & OaR	D50 (Gy) No risk factor best fit (68%CI)	Dose-modifying factor	m/k best fit (68%CI)	n best fit (68%CI)
Endpoint: grade 2–3 late rectal bleeding						
Rancati 2011 Fig. 1g [6]	70-78Gy 3DCRT	Logit-EUD Solid anorectum DVH	88.4 (−1.3,+1.5)	0.93 Abdominal surgery	10.7 (−0.7,+1.0)	0.03 (−0.01,+0.02)
Tucker 2010 Fig. 1h [7]	68.4–79.2Gy 3DCRT	LKB Solid anorectum DVH	79.1 (−1.9,+2.6)	0.95 Cardiovascular diseases	0.15 (−0.03,+0.04)	0.077 (−0.02,+0.04)
Endpoint: grade 3 late rectal bleeding						
Rancati 2011 Fig. 1l [6]	70-78Gy 3DCRT	LKB Solid anorectum DVH	91.7 (−2.3,+2.5)	0.90 Abdominal surgery	10.3 (−0.8,+1.2)	0.05 (−0.02,+0.03)
Defraene 2012 Fig. 1m [8]	68-78Gy 3DCRT	LKB Anorectal wall DVH	82.4 (5.9,+10.0)	0.91 Abdominal surgery	0.15 (−0.3,+0.5)	0.18 (−0.11,+0.14)
Defraene 2012 Fig. 1n [8]	68-78Gy 3DCRT	LKB Anorectal wall DVH	82.9 (5.9,+10.0)	0.91 Abdominal surgery 0.92 Cardiovascular diseases	0.15 (−0.3,+0.5)	0.18 (−0.11,+0.14)

RT = Radiation Oncology; NTCP = Normal Tissue Complication Probability; OaR = Organ at Risk; pts = patients; 3DCRT = three dimensional conformal Radiation Oncology; DVH = dose-volume histogram; LKB = Lyman-Kutcher-Burman (Lyman model coupled to Kutcher-Burman DVH reduction method); Logit + EUD = Logit equation coupled to DVH reduction to Equivalent Uniform Dose; CI = confidence interval; D50 = dose at which toxicity in 50% patients is expected if the rectum is uniformly irradiated to that dose; m/k = parameters expressing the slope of the NTCP curve, m for LKB model and k for Logit-EUD model; n = volume parameter, used in DVH reduction procedures. Relationship between steepness coefficients is m = 1.6/k.

RT = Radiation Oncology; NTCP = Normal Tissue Complication Probability; OaR = Organ at Risk; pts = patients; 3DCRT = three dimensional conformal Radiation Oncology; DVH = dose-volume histogram; LKB = Lyman-Kutcher-Burman (Lyman model coupled to Kutcher-Burman DVH reduction method); Logit + EUD = Logit equation coupled to DVH reduction to Equivalent Uniform Dose; CI = confidence interval; D50 = dose at which toxicity in 50% patients is expected if the rectum is uniformly irradiated to that dose; m/k = parameters expressing the slope of the NTCP curve, m for LKB model and k for Logit-EUD model; n = volume parameter, used in DVH reduction procedures.

the analysis for this last aspect, with discrimination for DVH of patients treated with or without lymph-nodes irradiation.

Some of the models also included the presence of covariates used as dose-modifying factors. Indeed, the presence of previous abdominal surgery and cardiovascular disease were found to be risk factors in models for LRB G2+ and G3. Details of the selected NTCP models are reported in Table 1a (NTCP models without dose-modifying factors) and Table 1b (NTCP models with dose-modifying factors). Fig. S1 in Supplementary Materials summarises the volume parameters and toxicity rate information for all the considered studies: prevalently serial-like behaviour of the anorectum for LRB was demonstrated. In contrast, different rates of LRB are reported. As both conventional and hypofractionated schedules were included in the study population, all doses were corrected for fraction size using the linear-quadratic model with $\alpha/\beta = 3$ Gy [30].

The performance of the NTCP models was evaluated through predictive accuracy (Hosmer-Lemeshow test, HL), calibration plot and the Area Under the Receiver Operating Characteristics Curve (AUC). Specifically, for what concerns the calibration plots and the agreement between observed endpoints and predictions, we considered:

(i). the calibration slope, i.e., the linear coefficient of the fit, that describes the increase of the observed toxicity rate with the model predicted probability (a slope coefficient equal to 1 represents the best result);
(ii). the calibration-in-the-large, i.e. the offset, that compares the mean of all predicted risks with the mean observed risk and is used to evaluate whether the predictions are systematically too low or too high (offset equal to 0 represents the best result).

For a correct prediction, these two parameters should be as close as possible to their reference values.

We decided to group the patients into 4 iso-probability classes (low-, medium/low, medium/high and high-risk class) by dividing the overall range of probability values predicted by the NTCP models. Therefore, the risk grade does not refer to the absolute risk but to the relative risk within the cohort. For instance, if the range of predicted probabilities was between 0.05 and 0.45, patients were divided into 4 equally-spaced groups, i.e. predicted probabilities in the range: (i) 0.05–0.14, (ii) 0.15–0.24, (iii) 0.25–0.34 and (iv) 0.35–0.45. However, while determining the calibration line, we decided to follow the approach suggested by Miller et al. [31], consisting in computing the calibration slope and the calibration-in-

Table 2
Rates of rectal bleeding in each trial and in the pooled population.

Population	G1-2–3(%)	G2-3(%)	G3(%)
RADAR	38.2%	24.5%	8.8%
DUE01	18.5%	7.8%	5.0%*
Airopros0102	27.2%	11.4%	6.2%*
Pooled Population	28.5%	15.6%	6.8%

*p-value of two-proportion z-test between population was <0.05 except for DUE01 VS Airopros0102 in G3 LRB.

Table 3
Results of the performance evaluation of published normal tissue complication probability models on the merged population. Values of calibration slope highlighted in bold are associated with models that were considered for the computation of the Equivalent Uniform Dose constraints for moderate and severe Late Rectal Bleeding.

Reference	Calibration slope (optimal = 1)	Calibration-in-the-large (optimal = 0)	AUC validation *vs* AUC development
Grade 1–2-3 late rectal bleeding			
Gulliford 2012 [4]	0.24	0.19	0.54 *vs* AUC development not available
DVH only			*Probability range: 0.07–0.72
Brand 2021 [9]	0.25	0.18	0.54 *vs* AUC development not available
DVH only			*Probability range: 0.07–0.73
Grade 2–3 late rectal bleeding			
Rancati 2004 [5]	**0.73**	**0.14**	0.56 *vs* AUC development not available
DVH only			*Probability range: 0–0.26
Rancati 2011 [6]	−0.10	0.11	0.49 *vs* 0.63
DVH only			*Probability range: 0 – 0.23
Michalski 2010 [3]	0.46	0.12	0.55 *vs* AUC development not available
DVH only			*Probability range: 0 – 0.37
Tucker 2010 [7]	0.41	0.12	0.54 *vs* AUC development not available
DVH only			*Probability range: 0 – 0.33
Gulliford 2012 [4]	0.33	0.12	0.57 *vs* AUC development not available
DVH only			*Probability range: 0 – 0.56
Rancati 2011 [6]	−0.11	0.11	0.48 *vs* 0.64
DVH + abdominal surgery			*Probability range: 0 – 0.39
Tucker 2010 [7]	0.58	0.13	0.56 *vs* AUC development not available
DVH + cardiovascular diseases			*Probability range: 0 – 0.39
Brand 2021 [9]	0.53	0.08	0.57 *vs* AUC development not available
DVH only			*Probability range: 0 – 0.38
Grade 3 late rectal bleeding			
Rancati 2004 [5]	**0.70**	**0.06**	0.59 *vs* AUC development not available
DVH only			*Probability range: 0 – 0.23
Rancati 2011 [6]	0.41	0.05	0.55 *vs* 0.63
DVH only			*Probability range: 0 – 0.17
Defraene 2012 [8]	**0.88**	**0.04**	0.58 *vs* 0.70
DVH only			*Probability range: 0 – 0.39
Rancati 2011 [6]	0.36	0.05	0.62 *vs* 0.67
DVH + abdominal surgery			*Probability range: 0 – 0.23
Defraene 2012 [8]	**1.36**	**0.04**	0.61 *vs* 0.74
DVH + abdominal surgery			*Probability range: 0 – 0.37
Defraene 2012 [8]	**0.94**	**0.05**	0.61 *vs* 0.77
DVH + abdominal surgery + cardiovascular diseases			*Probability range: 0 – 0.57

DVH = dose-volume histogram; GI = gastro-intestinal; Probability range = Range of model predicted probabilities in the external validation population.

the-large by fitting through a linear regression model the binary outcome variable (yes/no scoring of toxicity) as a function of the model predicted probability. In this way, we avoided having any dependence on the fit from the number of the considered experimental points (grouping of patients with respect to iso-predicted probabilities).

Statistical analyses were performed using R (https://www.r-project.org) and the KNIME software (KNIME GmbH, Germany).

Results

The merged dataset consisted of 1633 patients (654 from RADAR, 707 from Airopros0102 and 272 from DUE01). We had complete clinical and dosimetric information for all of these patients and at least 3 years of follow-up. The Supplementary Material (Table S1) shows the patient characteristics table. G1+ LRB was scored in 465 (28.5%) patients. G2+ and G3 LRB were reported by 255 (15.6%) and 112 (6.8%) patients, respectively. LRB rates stratified for every single trial are shown in Table 2.

Fig. S2 in the Supplementary Material shows the average rectum DVH for patients without toxicity and patients with G1+/G2+/G3 LRB. The average DVHs were significantly different between patients without toxicity and patients with any grade of LRB (results for t-test on DVH cutpoints at 5 Gy steps are also reported in the Supplementary Material). Regions of high dose discriminated between average DVH for G1+ and G2+/G3 LRB. Still, no difference between G2+ and G3 was observed in the high dose range. Regions of the low-medium dose also separated G2+ vs G3 LRB.

Table 3 presents a summary of the performance of the NTCP models we tested by reporting for each model the values of the calibration slope (optimal value equal to 1), of the calibration-in-the-large (optimal value equal to 0), its AUC in development (where available from original publication) and the AUC we obtained from our validation analysis. Fig. 1 depicts the calibration plots and the fit of the calibration line for each model for G1+ (Fig. 1a) and G2+ (Fig. 1b to Fig. 1h) LRB. Calibration plots of models for G3 LRB are reported in Fig. 1i-m. The P-value for the HL test was < 0.001 in all cases, indicating that the agreement between absolute predicted probabilities and absolute observed toxicity rates is poor. Most of the models were far from a correct calibration (i.e. calibration slope = 1 and calibration-in-the-large = 0; the perfect calibration line is described by the dashed line in each calibration plot). However, an increase in the toxicity risk accompanied by an increase in the predicted probability can be observed for specific predictive models across the spectrum of toxicity grades.

In particular, a clear trend can be identified in 5 models: (iii) Rancati et al. 2004 [5] for G2+ LRB with the inclusion of DVH (Fig. 1b) and G3 LRB with the inclusion of DVH for (Fig. 1i), (iii + iv + v) Defraene et al. [8] for G3 LRB with the inclusion of DVH (Fig. 1k) and G3 LRB with the inclusion of DVH and clinical risk factors (Fig. 1m and 1n).

Discussion

Predictive models published over the last ten years have shown that a large number of features are involved as risk/protective factors in the development of radiation-induced side effects. The approach used in these studies, i.e., the detailed recording of a large number of clinical observations followed by statistical analysis, classifies them as a phenomenological description of a complex situation [2,3]. These models are not based on radiobiological theories and experiments; therefore, they cannot be considered as absolute truth. On the other hand, their performance, validity, and also generalizability requires to be tested on external populations and not only in their development cohort [30–32].

The possibility of running external validation can be considered for existing independent prospectively-followed populations. In doing that, it is of paramount importance that the considered endpoint is scored using the same criteria in both the original model-development population and in the independent cohort used for the model validation).

In this work, we chose to join three large populations accrued in prospective trials to assess the generalizability of all NTCP models published in the literature until 2021 based on EUD from rectal DVH, intending to reach TRIPOD type 4 models possibly. These three populations were somewhat different in Radiation Oncology practice (i.e. RT schedule, irradiation of pelvic lymph nodes, irradiation of seminal vesicles, and use of hormone therapy), rectal DVHs and the prevalence of clinical risk factors. This heterogeneity of the cohorts allowed us to test the selected models in terms of generalizability to a great extent.

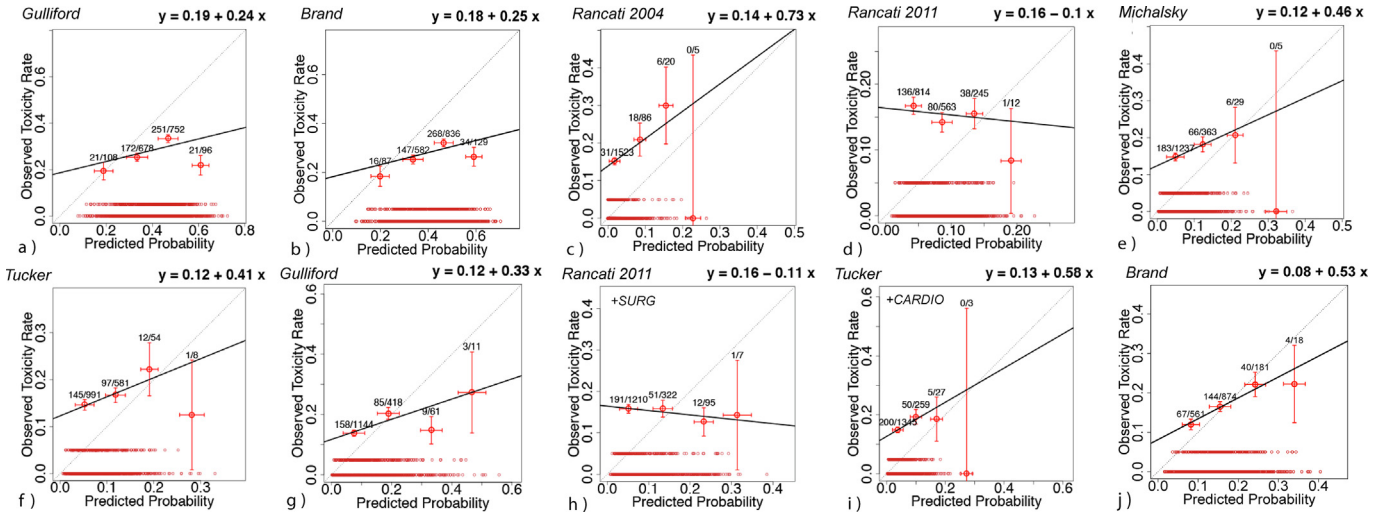

Fig. 1. Calibration plot and fit equation for grade 1+ (a,b), grade 2+ (c-j), grade 3 (k-p) models. for each plot, the predicted probability was computed using the ntcp model included in the reference on the top of the graph. plot (h) and (i) depict multivariate models for grade 2+ with the inclusion of previous abdominal surgery and presence of cardiovascular disease, respectively. Plot (n) and (o) depict multivariate models for Grade 3 with the inclusion of previous abdominal surgery, plot (p) depicts a multivariate model including previous abdominal surgery and presence of cardiovascular disease. Dotted line represents a perfect calibrated model, red circles show the binomial distribution of toxicity values (0 and 1 lined with 0 and 0.05 value of the Observed Toxicity Rate, respectively) with Predicted Probability estimate.

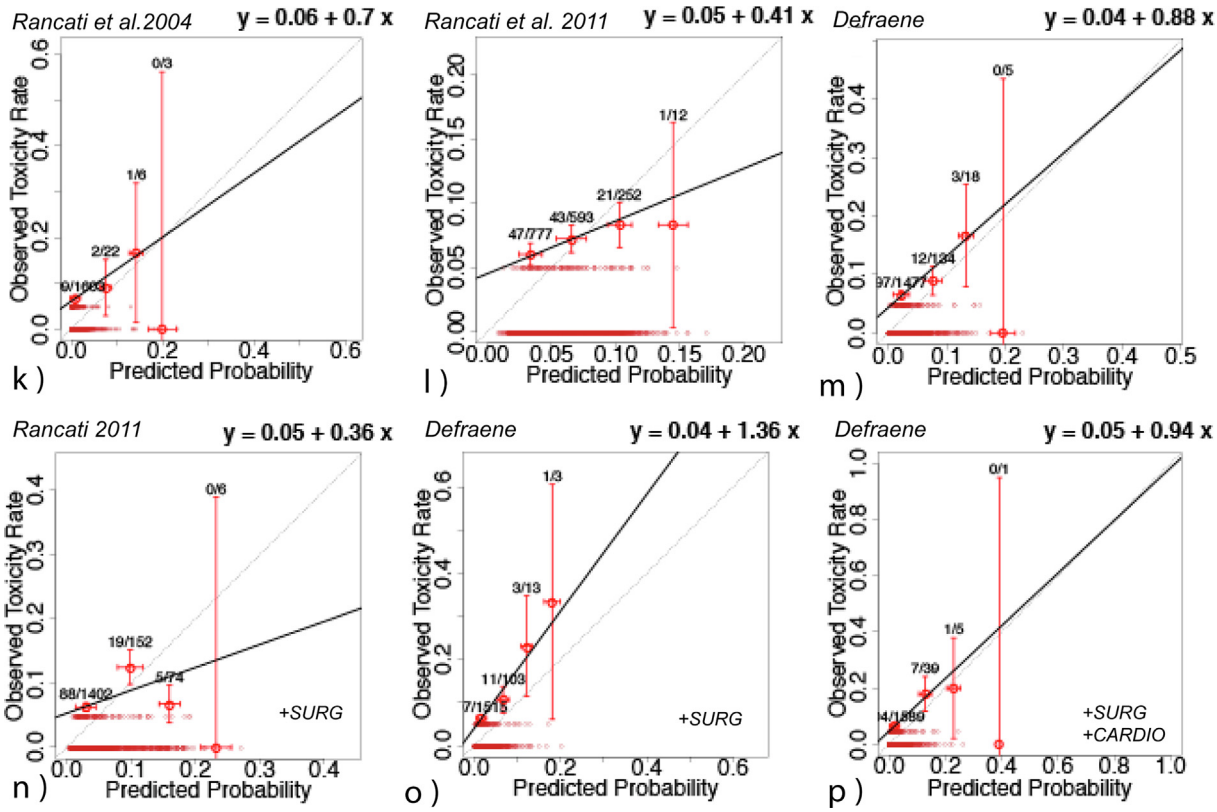

Fig. 1 (*continued*)

The first important step was the harmonization of clinical/dosimetric variables and toxicity endpoints. Validation was considered for all the bleeding endpoints (severity) studied in the literature: G1+, G2+, and G3. These endpoints were scored in most trials following the SOMA/LENT scale. Relevant patient-related features (comorbidities, use of drugs, previous surgeries) were explicitly recorded in all trials through a questionnaire compiled by the patients before radiotherapy. The three studies defined the anorectal volume in the same way (solid organ, length from the sigmoid junction to the anus), thereby allowing the direct comparison of DVHs.

The validation process explicitly included the calculation of absolute toxicity probabilities following the different published NTCP models (without any adjustment of the model parameters) and the assessment of the model's performance through the evaluation of calibration, goodness-of-fit, and AUC.

Of note, there are several aspects to take into account for the validation of NCPT models:

(i) Confirming the risk or protective behaviour of the features (Odds Ratio above vs below 1);
(ii) Confirming the effect size of the features (absolute values of the Odds Ratios in the validation set);
(iii) Confirming an increased risk of toxicity with an increase of toxicity probability prediction (slope of calibration plot as much as possible close to 1;
(iv) Confirming the offset and so the absolute toxicity risk across the whole cohort if the previous point is satisfied.

Five of the 16 considered NTCP models exhibited a good performance on the merged population considered in this study with calibration slope in the range of 0.7–1.36 (i.e. they satisfied points i, ii, and iii above), thus confirming the pivotal role of the EUD in dose–response of the rectum and, as a consequence, in treatment planning, optimization constraints for an incidence of 5–10% of LRB following these models are reported in Table 4.

Table 4
Equivalent Uniform Dose corresponding to 5% and 10% risk of developing grade ≥ 2 (G2+) (Rancati et al. [5] used as a validated model) and grade = 3 (G3) (Rancati et al. [5 and] Defraene et al. [8] used as validated model) late rectal bleeding.

Grade	NTCP Model	EUD @ 5% risk of bleeding (Gy)	EUD @ 10% risk of bleeding (Gy)
G2+	Rancati et al. (2004) [6]	58.5	62.5 Gy
G3	Defraene et al. [8]	59.9	64.3
		without risk factors	without risk factors
		57.3	61.5
		with one clinical risk factor	with one clinical risk factor
		51.0	55.2
		with two clinical risk factors	with two clinical risk factors
G3	Rancati et al. (2004) [5]	70.4 Gy	72.4 Gy

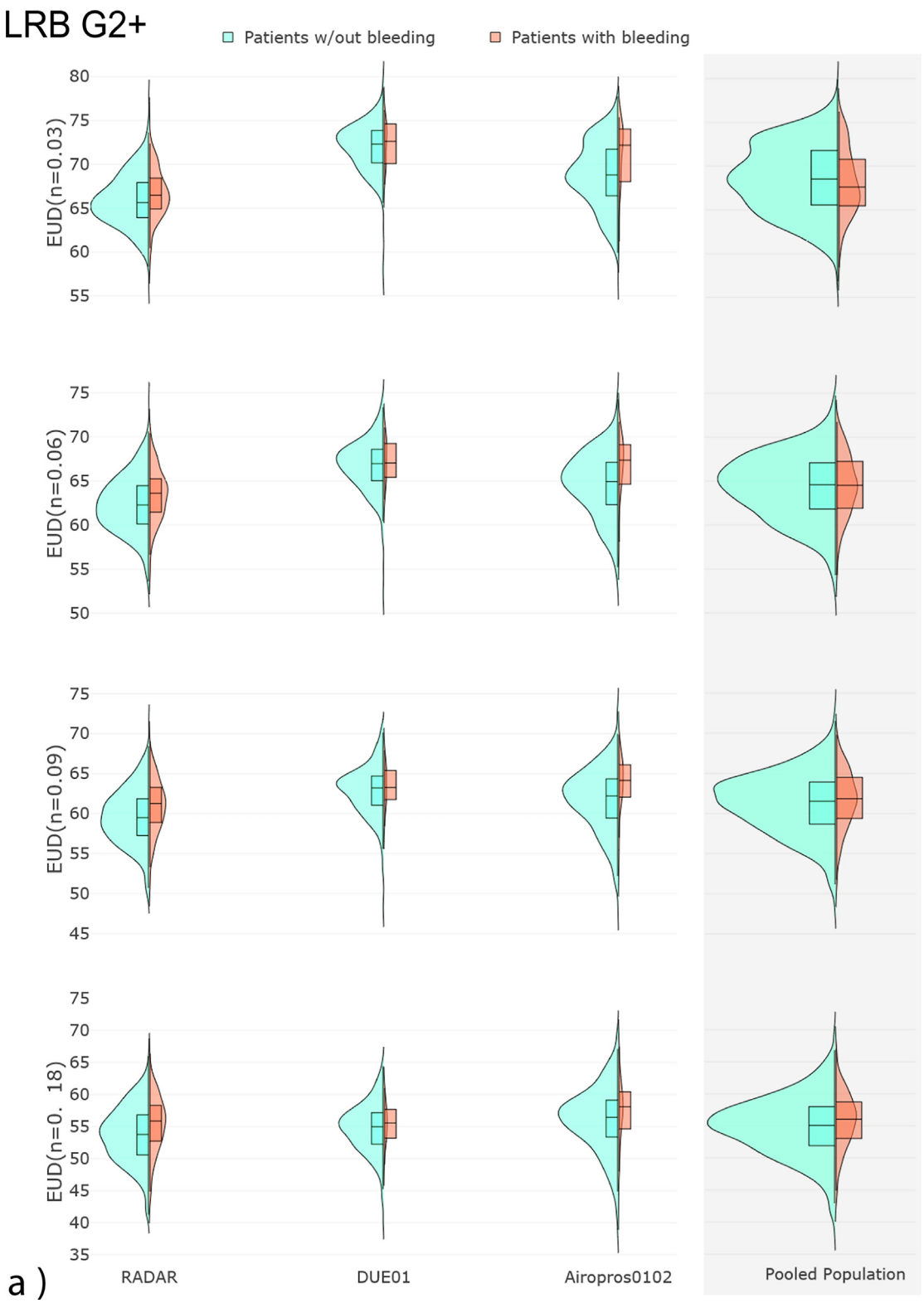

Fig. 2. A) violin plots and box plots for theEquivalent Uniform Dose (EUD) computed with 4 different volume parameter values (n = 0.03, 0.06, 0.09, 0.18) and used to discriminate between patients with and without grade $\geq$ 2 late rectal bleeding (LRB G2+). These plots are used to describe: (i) the effect of the volume parameter on the ranking of patients with toxicity and (ii) the impact of the same parameter on having a single homogeneous population (EUD computed with n = 0.03) instead of two separated distributions (EUD computed with n = 0.18). Plots were generated for every cohort and for the pooled population (last column in grey). A concise scheme for LRB grade $\geq$ 3 (G3) is shown in plot b). b) Violin plots and box plots for the EUD computed with four different volume parameter values (n = 0.03, 0.06, 0.09, 0.18) and used to discriminate between patients with and without grade 3 late rectal bleeding in the pooled population.

Notably, these five models had an acceptable calibration slope coupled to general offsets in absolute predictions (i.e. they did not satisfy point iv; therefore, they could not confirm the absolute toxicity risk across the whole cohort). This offset is very similar among the models investigating the same toxicity grade, suggesting a solid consistency among predictions. Calibration-in-the-large is often overlooked in validation studies [33], with the major focus usually on discrimination. Nonetheless, a satisfying agreement between predicted and observed toxicity rates at the absolute level is of great importance. When a model is used in clinical decision-making/interventional trials, the acceptable absolute risk is considered, and the reliability of such an absolute estimate is of paramount importance. A poor estimate of risk predictions can be due to causes connected to variables and characteristics which are not related to algorithm development [33]. Patient characteristics and toxicity incidence vary significantly between hospitals, countries, and times due to treatment and healthcare policy changes. Such heterogeneity between settings can affect estimated probabilities and calibration. The predictors in the algorithm (e.g., missing one crucial variable) may explain a part of the heterogeneity, but differences between predictors often do not explain all differences between settings. The second set of possible causes for poor calibration relates to overfitting, which is usually due to too complex modelling strategies for the amount of data available for training [33]. This should be of minor importance for the NTCP models considered in this analysis. Large-sized populations were considered, and reasonably straightforward modelling was used, with few possible predictors.

As we found poor calibration-in-the-large for the models considered in this analysis, when considering using these tools for decision-making/interventional trials, we suggest checking on their performance on historical patients from the same centre to understand if re-calibration could be recommended.

Models for G3 bleeding showed the best performance across the external validation process, also when clinical factors were included. Possibly, it could be a consequence of the fact that grade 3 LRB is a more objective endpoint with respect to grade 1 and grade 2 bleeding. Even in trials where there is an effort to have a prospective objective scoring, mild and moderate bleeding is still subjected to some confounding factors in scoring (e.g. the number of days in a week a patient is acknowledging a paper bleeding and the importance that is given to these events during a follow-up which is performed every 6 months). This is also somehow highlighted by the more homogeneous toxicity rates for grade 3 LRB among the cohorts. Of note, the DVHs are more effective in discriminating between patients with grade 3 bleeding with respect to patients with grade < 3 toxicity (see Fig. S2). Among NTCP models for G3 bleeding, the best performance was found in the models published by Defraene et al. [8], which were developed on a cohort of patients with a range of prescribed doses very similar to the one considered in the independent pooled population of the validation set. The introduction of the presence of previous abdominal surgery as a risk factor slightly improved the AUC but poured the calibration slope. When considering the addition of cardiovascular disease in a three-variable model, we found a better coherence between Predicted Probability and Observed Toxicity Rate reflected in a better calibration. This result suggests that such a model is the most indicated for predicting the absolute risk of severe bleeding.

Finally, the dosimetric factors considered by the five selected models, i.e. EUD with volume parameter n = 0.24 for G2+, and n = 0.18 and n = 0.06 for G3, were found to be significant risk factors also on the merged cohort (odds ratios from the univariate analysis are available in the Supplementary Materials). These results might allow us to infer that symptoms, such as late rectal bleeding, are strongly associated with the radiotherapy dose. Furthermore, even if the serial behaviour of the rectum is broadly acclaimed in radiation oncology, it is crucial to point out that this validation study slightly moves the focus to the medium–high dose region, with a "softening" of the seriality of the rectum when compared, for example, to the QUANTEC statements (n = 0.09), with best models using n = 0.18 and n = 0.24. These n values allow the modelling description of the high rates of toxicity in the RADAR cohort, with patients treated with 3DCRT and DVHs of patients with/without toxicity showing a good separation in the region between 55 and 68 Gy. To better understand how the n value acts on the resulting distribution of EUD, we included in Fig. 2 a complete example for grade $\geq$ 2 late rectal bleeding (Fig. 2a). Violin and box plots show how patients with/without toxicity are distributed in every cohort according to EUD computed with different volume parameters n. The higher the value of n, the more significant the overlapping among the EUD distributions in the three populations. At the same time, the separation between the box plots of bleeders and not bleeders increases with increasing n, showing promising results for n = 0.18 (and in a similar way for n = 0.24). Indeed, EUD calculated with n = 0.03 is not a good dosimetric descriptor for the risk of bleeding G2+, entailing a decreasing risk of toxicity with the increase of the EUD (a short video describing this process is included in SM). In a complementary way, Fig. 2b depicts the trend in EUDs calculated with different n values vs grade 3 late rectal bleeding. In this case, EUDs giving more weight to the high dose tail of the DVH (low n values) are significantly associated with an increased risk of severe bleeding, but (as for G2 + bleeding, see Fig. S3), the related p-values decrease for increasing n value, i.e. for EUDs including a heightened weight of the medium doses to the rectum.

A possible limitation of the analysis performed is that two of the three cohorts pooled to construct the validation population (RADAR and Airopros0102) consider patients treated with 3DCRT. However, even if IMRT is currently widely recognized as the best available approach for external beam prostate RT, 3DCRT is still considered across Europe (~33% in the Prospective Observational Pros-IT CNR Study [34], 16% in the Prospective Observational REQUITE study [35]) and widely used in middle-and low-income countries where the IMRT is not covered by the public health system [36]. Additionally, because the majority of the NTCP models evaluated are contemporary to the validation 3DCRT cohorts, they were indirectly assessed against geographical changes. On the other hand, the contemporary DUE01 cohort, including patients treated with IMRT/VMAT/Tomotherapy and hypofractionation, allowed for an indirect assessment of the generalizability of models with respect to RT techniques and schedules.

Conclusions

The performance of 16 NTCP models for the estimate of late rectal bleeding risk following prostate cancer radiotherapy was assessed on a large validation population independent from the development population of the original models. Five models exhibited good performance, with 3 models for grade 3 bleeding resulting in more general applicability across cohorts and when clinical features were added as dose-modifying factors. EUD proved to be a valid dosimetric descriptor to be associated with rectal bleeding. Notably, this validation study moves the EUD emphasis to the medium–high dose region, with a "softening" of the serial behaviour of the rectum, with best models using n = 0.18 and n = 0.24.

Data sharing statement

The data that support the findings of this study are available from the corresponding author, [A. C.], upon reasonable request.

Conflict of interest statement

The authors have nothing to disclose.

Acknowledgments

AC is supported by the AIRC IG 21479 (P.I. TR). JI is supported by the grant FRRB 2721017. TR is supported by Fondazione Italo Monzino. The study was also supported by AIRC IG 14603 and 230150, NHMRC (300705, 455521, 1006447). We are also grateful for the assistance of the Trans-Tasman Radiation Oncology Group (TROG).

I want to thank Ludovica Angeletti and Francesca Venturoni for helping me with graphic design.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2023.109628.

References

[1] Fiorino C, Valdagni R, Rancati T, Sanguineti G. Dose-volume effects for normal tissues in external Radiation Oncology: pelvis. Radiother Oncol 2009;93:153–67. https://doi.org/10.1016/j.radonc.2009.08.004.
[2] Landoni V, Fiorino C, Cozzarini C, et al. Predicting toxicity in Radiation Oncology for prostate cancer. Phys Med 2015. https://doi.org/10.1016/j.ejmp.2016.03.003.
[3] Michalski JM, Gay H, Jackson A, et al. Radiation dose-volume effects in radiation-induced rectal injury. Int J Radiat Oncol Biol Phys 2010;76:S123–9. https://doi.org/10.1016/j.ijrobp.2009.03.078.
[4] Gulliford SL, Partridge M, Sydes MR, et al. Parameters for the Lyman Kutcher Burman (LKB) model of Normal Tissue Complication Probability (NTCP) for specific rectal complications observed in clinical practise. Radiother Oncol. 2012;102:347–51. https://doi.org/10.1016/j.radonc.2011.10.022. Epub 2011 Nov 25. PubMed PMID: 22119373.
[5] Rancati T, Fiorino C, Gagliardi G, et al. Fitting late rectal bleeding data using different NTCP models: results from an Italian multi-centric study (AIROPROS0101). Radiat Oncol Oncol 2004;73:21–32. https://doi.org/10.1016/j.radonc.2004.08.013.
[6] Rancati T, Fiorino C, Fellin G, et al. Inclusion of clinical risk factors into NTCP modelling of late rectal toxicity after high dose Radiation Oncology for prostate cancer. Radiat Oncol Oncol 2011;100:124–30. https://doi.org/10.1016/j.radonc.2011.06.032.
[7] Tucker SL, Dong L, Bosch WR, et al. Late rectal toxicity on RTOG 94–06: analysis using a mixture Lyman model. Int J Radiat Oncol Biol Phys 2010;78:1253–60. https://doi.org/10.1016/j.ijrobp.2010.01.069.
[8] Defraene G, Van den Bergh L, Al-Mamgani A, et al. The benefits of including clinical factors in rectal normal tissue complication probability modeling after Radiation Oncology for prostate cancer. Int J Radiat Oncol Biol Phys 2012;82:1233–42. https://doi.org/10.1016/j.ijrobp.2011.03.056.
[9] Brand DH, Brüningk SC, Wilkins A, et al. CHHiP Trial Management Group. Estimates of Alpha/Beta (α/β) Ratios for Individual Late Rectal Toxicity Endpoints: An Analysis of the CHHiP Trial. Int J Radiat Oncol Biol Phys. 2021;110:596–608. https://doi.org/10.1016/j.ijrobp.2020.12.041. Epub 2021 Jan 4. PMID: 33412260; PMCID: PMC8129972.
[10] Peeters ST, Lebesque JV, Heemsbergen WD, et al. Localized volume effects for late rectal and anal toxicity after Radiation Oncology for prostate cancer. Int J Radiat Oncol Biol Phys 2006;64:1151–61.
[11] Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med 2015 Jan 6;162:W1–W. https://doi.org/10.7326/M14-0698.
[12] Semenenko VA, Tarima SS, Devisetty K, et al. Validation of normal tissue complication probability predictions in individual patient: late rectal toxicity. Int J Radiat Oncol Biol Phys. 2013;85:1103–9. https://doi.org/10.1016/j.ijrobp.2012.07.2375. Epub 2012 Sep 25.
[13] Thor M, Deasy J, Iyer A, et al. Toward personalized dose-prescription in locally advanced non-small cell lung cancer: Validation of published normal tissue complication probability models. Radiother Oncol 2019;138:45–51. https://doi.org/10.1016/j.radonc.2019.05.011. Epub 2019 May 27.
[14] Fellin G, Fiorino C, Rancati T, et al. Clinical and dosimetric predictors of late rectal toxicity after conformal radiation for localized prostate cancer: results of a large multicentric observational study. Radiother Oncol 2009;93:197–202.
[15] TROG. TROG clinical trials summary. TROG 03.04, 2005 - A Randomised Trial Investigating the Effect on Biochemical (PSA) Control and Survival of Different Durations of Adjuvant Androgen Deprivation in Association With Definitive Radiation Treatment for Localised Carcinoma of the Prostate (RADAR).
[16] Palorini F, Rancati T, Cozzarini C, et al. Multivariable models of large International Prostate Symptom Score worsening at the end of therapy in prostate cancer Radiation Oncology. Radiother Oncol 2016;118:92–8.
[17] Denham JW, Wilcox C, Lamb DS, et al. Rectal and urinary dysfunction in the TROG 03.04 RADAR trial for locally advanced prostate cancer. Radiother Oncol 2012;105:184–92.
[18] Ebert MA, Foo K, Haworth A, et al. Gastrointestinal dose-histogram effects in the context of dose-volume-constrained prostate radiation therapy: analysis of data from the RADAR prostate radiation therapy trial. Int J Radiat Oncol Biol Phys 2015;91:595–603.
[19] Ebert MA, Harrison KM, Howlett SJ, et al. Dosimetric intercomparison for multicenter clinical trials using a patient-based anatomic pelvic phantom. Med Phys 2011;38:5167–75.
[20] Ebert MA, Haworth A, Kearvell R, et al. Detailed review and analysis of complex Radiation Oncology clinical trial planning data: Evaluation and initial experience with the swan software system. Radiother Oncol 2008;86:200–10.
[21] Kearvell R, Haworth A, Ebert MA, et al. Quality improvements in prostate Radiation Oncology: Outcomes and impact of comprehensive quality assurance during the TROG 03.04 'RADAR' trial JMIRO, 57 (2013), pp. 247–257 G. Fellin, T. Rancati, C. Fiorino et al. Long term rectal function after high-dose prostatecancer Radiation Oncology: results from a prospective cohort study. Radiother Oncol 2014;110:272–7.
[22] Fellin G, Rancati T, Fiorino V, et al. Long term rectal function after high-dose prostate cancer Radiation Oncology: results from a prospective cohort study. Radiother Oncol 2014;110:272–7.
[23] Valdagni R, Vavassori V, Rancati T, et al. Increasing the risk of late rectal bleeding after high-dose Radiation Oncology for prostate cancer: the case of previous abdominal surgery. Results from a prospective trial. Radiother Oncol 2012;103:252–5.
[24] Valdagni R, Kattan MW, Rancati T, et al. Is it time to tailor the prediction of radio-induced toxicity in prostate cancer patients? Building the first set of nomograms for late rectal syndrome. Int J Radiat Oncol Biol Phys 2011;82:19571966.
[25] Cozzarini C, Rancati T, Palorini F, et al. Patient-reported urinary incontinence after radiotherapy for prostate cancer: Quantifying the dose-effect. Radiother Oncol 2017;125:101–6. https://doi.org/10.1016/j.radonc.2017.07.029. Epub 2017 Aug 18.
[26] Carillo V, Cozzarini C, Rancati T, et al. Relationships between bladder dose–volume/surface histograms and acute urinary toxicity after Radiation Oncology for prostate cancer. Radiother Oncol 2014;111:100–5.
[27] Cicchetti A, Rancati T, Ebert M, et al. Modelling late stool frequency and rectal pain after radical radiotherapy in prostate cancer patients: Results from a large pooled population. Phys Med 2016;32:1690–7. https://doi.org/10.1016/j.ejmp.2016.09.018. Epub 2016 Oct 6.
[28] Foppiano F, Fiorino C, Frezza G, et al. The impact of contouring uncertainty on rectal 3D dose-volume data: results of a dummy run in a multicenter trial (AIROPROS01-02). Int J Radiat Oncol Biol Phys 2003;57:573–9. https://doi.org/10.1016/s0360-3016(03)00659-x.
[29] LENT SOMA scales for all anatomic sites. Int J Radiat Oncol Biol Phys. 1995;31:1049–91. https://doi.org/10.1016/0360-3016(95)90159-0. PMID: 7713776
[30] Bentzen SM, Dörr W, Gahbauer R, et al. Bioeffect modeling and equieffective dose concepts in radiation oncology – Terminology, quantities and units. Radioth Oncol 2012;105:266–8. https://doi.org/10.1016/j.radonc.2012.10.006.
[31] Miller ME, Langefeld CD, Tierney WM, Hui SL. C J McDonald validation of probabilistic predictions. Med Decis Making Jan-Mar 1993;13:49–58. https://doi.org/10.1177/0272989X9301300107.
[32] Van der Schaaf A, Langendijk JA, Fiorino C, Rancati T. Embracing phenomenological approaches to normal tissue complication probability modeling: a question of method. Int J Radiat Oncol Biol Phys 2015;91:468–71. https://doi.org/10.1016/j.ijrobp.2014.10.017.
[33] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med 2019;17:230. https://doi.org/10.1186/s12916-019-1466-7. PMID: 31842878; PMCID: PMC6912996.
[34] Noale M, Bruni A, Triggiani L, et al. Impact of gastrointestinal side effects on patients' reported quality of life trajectories after radiotherapy for prostate cancer: data from the prospective, Observational Pros-IT CNR Study. Cancers 2021;13:1479. https://doi.org/10.3390/cancers13061479.
[35] Seibold P, Webb A, Aguado-Barrera ME. REQUITE: A prospective multicentre cohort study of patients undergoing radiotherapy for breast, lung or prostate cancer. Radiother Oncol 2019;138:59–67. https://doi.org/10.1016/j.radonc.2019.04.034.
[36] Konski A. Cost effectiveness of prostate cancer radiotherapy. Transl Androl Urol 2018;7:371–7. https://doi.org/10.21037/tau.2017.12.38.