Radiotherapy and Oncology 113 (2014) 215–222

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

PET/CT in rectal cancer radiotherapy

Nomogram predicting response after chemoradiotherapy in rectal cancer using sequential PETCT imaging: A multicentric prospective study with external validation

Ruud G.P.M. van Stiphout [a,*], Vincenzo Valentini [b], Jeroen Buijsen [a], Guido Lammering [a,c], Elisa Meldolesi [b], Johan van Soest [a], Lucia Leccisotti [d], Alessandro Giordano [d], Maria A. Gambacorta [e], Andre Dekker [a], Philippe Lambin [a]

[a] *Department of Radiation Oncology (MAASTRO), Maastricht University Medical Centre, The Netherlands;* [b] *Radiotherapy Department, Università Cattolica S. Cuore, Rome, Italy;* [c] *Department of Radiotherapy, MediClin Robert Janker Klinik, Bonn, Germany;* [d] *Department of Nuclear Medicine, Università Cattolica S. Cuore, Rome; and* [e] *Bioimmagini e Scienze Radiologiche, Università Cattolica S. Cuore, Rome, Italy*

ARTICLE INFO

Article history:
Received 29 October 2013
Received in revised form 31 October 2014
Accepted 1 November 2014
Available online 20 November 2014

Keywords:
Tumour response
^{18}F-FDG PET imaging
Outcome prediction
Rectal cancer
Prospective study
External validation

ABSTRACT

Purpose: To develop and externally validate a predictive model for pathologic complete response (pCR) for locally advanced rectal cancer (LARC) based on clinical features and early sequential ^{18}F-FDG PETCT imaging.

Materials and methods: Prospective data (i.a. THUNDER trial) were used to train (*N* = 112, MAASTRO Clinic) and validate (*N* = 78, Università Cattolica del S. Cuore) the model for pCR (ypT0N0). All patients received long-course chemoradiotherapy (CRT) and surgery. Clinical parameters were age, gender, clinical tumour (cT) stage and clinical nodal (cN) stage. PET parameters were SUV_{max}, SUV_{mean}, metabolic tumour volume (MTV) and maximal tumour diameter, for which response indices between pre-treatment and intermediate scan were calculated. Using multivariate logistic regression, three probability groups for pCR were defined.

Results: The pCR rates were 21.4% (training) and 23.1% (validation). The selected predictive features for pCR were cT-stage, cN-stage, response index of SUV_{mean} and maximal tumour diameter during treatment. The models' performances (AUC) were 0.78 (training) and 0.70 (validation). The high probability group for pCR resulted in 100% correct predictions for training and 67% for validation. The model is available on the website www.predictcancer.org.

Conclusions: The developed predictive model for pCR is accurate and externally validated. This model may assist in treatment decisions during CRT to select complete responders for a wait-and-see policy, good responders for extra RT boost and bad responders for additional chemotherapy.

Early prediction of pathologic complete response (pCR) for locally advanced rectal cancer (LARC) patients is valuable because it allows for individualized treatment reorientation [15,28]. The standard treatment for LARC patients is preoperative chemoradiotherapy (CRT) followed by surgery. The neo-adjuvant treatment, intended to control pelvic disease and improve the chance of sphincter preservation, results in a pathological complete response (pCR) in 15–30% of the patients [9,21]. For these complete responders a wait-and-see policy after CRT is a possibility in order to reduce treatment-related morbidity and mortality, for which excellent results are reported [20]. This decision requires however very accurate predictions and assessment of complete response. Other treatment options under consideration are a radiotherapy boost after CRT for good responding patients to achieve more pCRs [8] and additional chemotherapy administration after CRT for the worst responding patients [3]. Both of these options require an early assessment of response even during CRT. Currently, the leading candidate predictive marker for histopathological response prediction in LARC is ^{18}F-fluorodeoxyglucose (FDG) positron emission tomography (PET) imaging. A meta-analysis from 2012 confirmed the added value of PET imaging, especially for intermediate PET imaging (during CRT) [31]. However, most studies evaluated pre-CRT versus post-CRT PET imaging. Besides that early prediction is preferred for treatment reorientation, later pre-

* Corresponding author at: MAASTRO Clinic, Dr. Tanslaan 12, PO Box 1588, 6201 BN Maastricht, The Netherlands.

E-mail address: ruud.vanstiphout@maastro.nl (R.G.P.M. van Stiphout).

http://dx.doi.org/10.1016/j.radonc.2014.11.002
0167-8140/© 2014 The Authors. Published by Elsevier Ireland Ltd.

dictions may also be affected by CRT-induced inflammatory tissue, which presents tumour equivalent signal on FDG-PET scans [26]. This recognition resulted in more early response assessment studies in the last few years (Table 1). The limitations of these studies were their small sample sizes ($N = 20–42$), the main focus on good versus bad responders (not pCR), the univariate setting in which analyses were performed and the lack of validation. To increase the clinical applicability of these decision making tools, they need to be based on more evidence (i.e. larger number of patients and external validation), be trained on several data sources [29] and they require focus on outcomes that are more relevant in terms of decisions, like pCR for a possible wait-and-see policy. We hypothesize that models with these requirements are the most suitable for decision making in clinical practice. The aim of this study is therefore to develop an externally validated multivariate predictive model for pCR combining clinical, pre-treatment and intermediate FDG-PETCT imaging parameters based on a prospective study. After development of a nomogram and the evaluation of its accuracy, risk group definition based on these predictions may provide decision support to clinicians for LARC patients (Fig. 1).

Materials and methods

Study population

All data were prospectively collected (with written informed consent) between January 2007 and March 2012 within two institutes: MAASTRO Clinic (GROW, MUMC, Maastricht, The Netherlands) and Università Cattolica del S. Cuore (Rome, Italy). The following prospective observational studies were involved: a study (2007–2009) involving 47 patients from Maastricht [12,14], a pilot study (2007–2009) with 19 patients from Rome and a multicentre study (2009–2012) involving one protocol for both institutes (MAASTRO: 65 patients, Rome: 59 patients) with acronym THUNDER (THeragnostic Utilities for Neoplastic DisEases of the Rectum, NCT00969657). All patients from Maastricht were pooled and used to train a prediction model for pCR ($N = 112$). The pooled datasets from Rome were used for external validation of the model ($N = 78$). The study inclusion criteria were: histological proven rectal cancer (primary tumours), UICC stage I–III, no recurrences, only concurrent chemoradiotherapy treatment, minimal age of 18 years, and no previous radiotherapy to the pelvis. The available clinical variables used as candidate prognostic and predictive factors were age, gender, clinical tumour (cT) and nodal (cN) stage. The criteria followed to consider tumour nodal involvement at MRI were related to border contour (sharply demarcated or irregular border) and signal intensity characteristics (homogeneous or inhomogeneous) or size >8 mm [2,7]. All patients from Maastricht were treated preoperatively with radiotherapy (28 fractions of 1.8 Gy, 5 fractions/week) and concomitant chemotherapy (capecitabine, 825 mg/m^2, twice daily), followed by a total mesorectal excision 6–8 weeks after the end of CRT. A minority of the thunder patients ($N = 11$) with a clinical complete response (assessed using post-CRT MRI and endoscopy) were enrolled in a parallel study where a surgical wait-and-see policy was applied [20]. Some patients from Rome were also treated with 50.4 Gy schedule, but 78.2% of the patients were treated with 25 × 1.8 Gy schedule and a RT boost of 10 Gy. The majority of the Rome patients ($N = 62$) received a combination of capecitabine (1300 mg/m^2 daily) and oxaliplatin (60 mg/m^2 once a week for 5 weeks with 55.0 Gy RT or 130 mg/m^2 at 3 time points with 50.4 Gy RT), and the others capecitabine only (1650 mg/m^2 daily with 50.4 Gy RT or 1300 mg/m^2 daily with 55 Gy RT, $N = 14$) or raltitrexed (3 mg/m^2 at 3 time points, $N = 2$).

PETCT imaging

All patients underwent a pre-CRT PET scan (one week before the start of CRT) and an intermediate PET scan (two weeks after the start of CRT). All Maastricht PET-CT scans were performed by use of a dedicated Siemens Biograph 40 TruePoint PET-CT scanner (Siemens Medical, Erlangen, Germany). Rome scans were performed with a 3D GEMINI GXL PET-CT scanner with 16 slice CT (Philips Healthcare, Cleveland, OH). The PET acquisition settings were reported before and were calibrated for both institutes [14]. PET-based semi-automatic tumour contours were made by one observer using a dedicated software (TrueD, Siemens Medical, Erlangen, Germany). Contours were defined by a threshold for the standardized uptake-value (SUV) based on the tumour-to-background signal ratio, with the gluteus muscle as reference background [6,23]. From the resulting tumour contour, maximal tumour diameter (MaxDiam), metabolic tumour volume (MTV), and maximal and mean SUV values within the MTV were calculated. The same variables were scored for the intermediate CRT PET-CT scan and

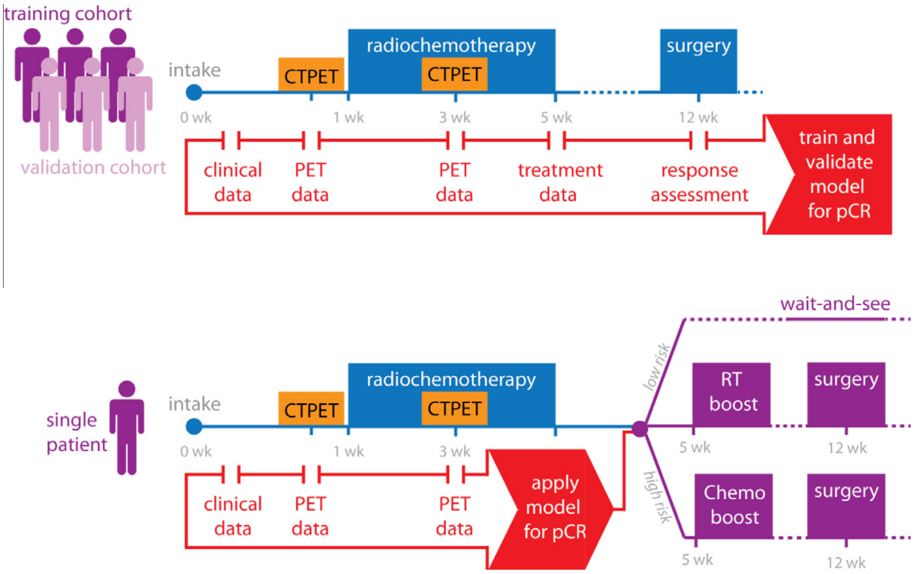

Fig. 1. Schematic overview of prediction model development (top) and the proposed application of the model in clinical practice after it has been tested in a clinical trial with a control arm (bottom).

Table 1
All studies for response prediction in locally advanced rectal cancer using early sequential PET imaging.

Study	Year	Accrual time	N	RT dose (Gy)	Chemo	2nd PET (days)	Surg (weeks)	Outcome	PA	Predictors	AUC
Avallone [1]	2012	NR	42	45	5-FU LFU Oxa Ral	12	8	Resp	TRG	$\text{SUVi}_{\text{mean}}$ SUVi_{max} $\Delta\text{SUV}_{\text{mean}}$ $\Delta\text{SUV}_{\text{max}}$	NR
Goldberg [10]	2012	2008.2–2009.7	20	50.4–55	5-FU Cap UFT	8	6–10	pCR Resp	TRG	pCR: $\Delta\text{SUV}_{\text{max}}$ resp: none	NR
Leibold [17]	2011	2001–2005	27	50.4	5-FU	8–14	5–8	Resp	95%[a]	VRS	NR
Lambrecht [16]	2010	2005.5–2009.8	22	50.4	5-FU	14–16	6–8	pCR	TNM	$\Delta\text{SUV}_{\text{max}}$	NR
Janssen [12][b]	2009	2007.4–2009.3	30	50.4	Cap	8 + 15[c]	6–8	Resp	TRG	$\Delta\text{SUV}_{\text{max}}$ $\Delta\text{SUV}_{\text{mean}}$	0.87
Guerra [11]	2009	2006.5–2009.3	28	50.4	5-FU	21	12	Resp	TRG	$\text{SUVi}_{\text{mean}}$	0.83
Rosenberg [25]	2008	2006.3–2007.1	30	45	5-FU	14	5	Resp	TRG[d]	None[e]	0.70
Cascini [5]	2006	NR	33	45	5-FU LFU Oxa Ral	12	8	Resp	TRG	SUVi_{max} $\text{SUVi}_{\text{mean}}$ $\Delta\text{SUV}_{\text{mean}}$ $\Delta\text{SUV}_{\text{max}}$	NR

2nd PET: time between start of chemoradiotherapy and the intermediate PET-scan; 5-FU: fluorouracil; Cap: capecitabine; ΔSUV: (relative) difference between pretreatment and intermediate standard uptake value of the PET contour; LFA: levofolinic acid; N: number of patients in study; NR: not reported; Oxa: oxaliplatin; PA: pathology; pCR: pathologic complete response; Ral: raltitrexed; Resp: tumour response; RT: radiotherapy; Surg: time between end of chemoradiotherapy and surgery; SUVi: intermediate standard uptake value of the PET contour; TRG: tumour regression grade; VRS: visual response score; UFT: tegafur–uracil.

[a] Pathologic response >95%.

[b] The separate validation study is not reported [14].

[c] Results reported for day 15 (optimal time point in this study).

[d] According to Becker standard, other studies used Mandard standard.

[e] Trend for $\Delta\text{SUV}_{\text{mean}}$ ($p = .085$).

for each variable a response index (RI) was calculated. The RI is the relative difference between the value of the intermediate scan and pre-CRT scan and defined as $(X_{\text{pre}} - X_{\text{intermediate}})/X_{\text{pre}} * 100\%$. SUV measures were corrected for blood glucose level [13].

Pathological assessment

Pathological complete response was defined as ypT0N0, extracted from the pathologic reports of surgical specimens. All other cases (ypT+ and/or ypN+) were considered non-responders. The specimens were not re-evaluated centrally but the pathology protocols were very similar between institutes (3–5 mm slices of rectum tumour, intensified evaluation on several blocks of tissue at the tumour site, evaluation on 2–3 sublevels when no tumour tissue was found in initial block). For the 11 wait-and-see patients, a pCR was assigned if the patient was locally recurrence free after 12 months of follow-up, a time point motivated by the extremely low recurrence rate in the wait-and-see study [20].

Statistical analysis

All statistical analyses were implemented and performed in MATLAB (version 7.1, MathWorks Inc., Natick, MA). Any missing values (Maastricht: 1.1%, Rome: 1.9%) in the datasets were substituted using the Expectation–Maximization method [18]. Datasets were pooled per institute on an individual patient level. Wilcoxon rank-sum tests were performed to test for association between a single variable and pCR. In the multivariate setting, logistic regression was applied to classify complete responders and non-responders, using the significant predictors from the univariate analysis as inputs. In the case of two very highly correlated input variables, only one was selected (using Spearman's correlation matrix, $p < .05$). The model's accuracy was evaluated with the area under the curve (AUC) of the receiver operating characteristic (ROC) curve [24]. The maximum value of the AUC is 1.0, indicating a perfect prediction model while a value of 0.5 indicates a random chance of correct prediction. Predictors were only selected if their addition resulted in a sufficient AUC change (>.01). For the final accuracy assessment, a bootstrapping scheme was applied (sampling with replacement, $N = 1000$), resulting in 95% confidence intervals for AUC. A nomogram was generated to represent a visualization of the final predictive model in which three risk groups were defined by applying two thresholds for the estimated probability for pCR. To define the low probability group for pCR, first the (weighted) average rate of non-responders (TRG3–4, tumour regression grade) was calculated from literature and thereafter a threshold was selected that resulted in this percentage of non-responders [1,5,10,11,14,17]. The threshold for the high estimated probability of pCR was calculated using decision curve analysis [30]. This method optimizes the threshold by calculating the net benefit of applying such a model and comparing it to the situations in which none or all patients are treated with a wait-and-see policy. These two thresholds define the three probability groups for pCR (high, mid, low), which correspond respectively to complete, good and bad responders with respect to pCR.

Results

Dataset characterization

Pooling the clinical and PETCT imaging data per institute resulted in similar cohort characteristics (Table 2). The validation dataset from Rome had in general: less males (64.1% vs 74.1%, $p = .187$), higher clinical tumour stages (cT4: 30.8% vs 8.0%, $p < .001$), and more nodal involvement (94.9% vs 86.6%, $p = .10$). Almost all Maastricht patients received 50.4 Gy of RT (95.5%) while the majority of the Rome patients received 55.0 Gy (78.2%). There was also a small but non-significant difference in the number of pathologic complete responders (23.1% vs 21.4%, $p = .927$). The average time from last RT fraction to surgery was equal (73.6 ± 18.8 vs 72.9 ± 13.2 days, mean ± SD). Despite harmonization of the PET protocols in the THUNDER study, the time between the PET scans was on average lower for the Maastricht dataset (21.9 ± 2.5 vs 28.5 ± 10.5 days). Also the time between tracer injection and PET acquisition was lower for the intermediate PET scan in Maastricht (69.5 ± 15.2 vs 81.9 ± 23.0 min).

Table 2
Patient characteristics of the training data set (Maastricht) and the validation data set (Rome).

		Maastricht		Rome	
		N	[%]	*N*	[%]
Clinical	Sex				
	Female	29	[25.9]	28	[35.9]
	Male	83	[74.1]	50	[64.1]
	Age (years)				
	Median	65.0		66.3	
	Range	44.0–81.1		27.0–82.7	
	Clinical tumour stage				
	2	17	[15.2]	5	[6.4]
	3	86	[76.8]	49	[62.8]
	4	9	[8.0]	24	[30.8]
	Clinical nodal stage				
	0	15	[13.4]	4	[5.1]
	+	97	[86.6]	74	[94.9]
PET imaging	Time between PET scans (days)				
	Mean	21.9		28.5	
	Standard deviation	±2.5		±10.5	
	Time 1st PET injection to acquisition (minutes)				
	Mean	82.6		80.2	
	Standard deviation	±18.1		±21.3	
	Time 2nd PET injection to acquisition (minutes)				
	Mean	69.2		81.9	
	Standard deviation	±15.2		±23.0	
Treatment	Total radiotherapy dose (Gy)				
	<50.4	5	[4.5]	2	[2.6]
	50.4	107	[95.5]	15	[19.2]
	55.0	0	[0.0]	61	[78.2]
	Time last RT fraction to surgery (days)				
	Mean	73.6		72.9	
	Standard deviation	±18.8		±13.2	
Outcome	Pathologic complete response				
	Yes	24	[21.4]	18	[23.1]
	No	88	[78.6]	60	[76.9]

Predictor selection

Univariate analyses (Table 3) showed that age, pre-treatment SUV measures, intermediate SUV$_{\text{mean}}$, and response index for maximal diameter have no significant predictive value for pCR ($\alpha = .05$). Negatively correlated significant predictors (i.e. increasing value results in lower pCR rate) were cT-stage, cN-stage, pre-treatment and intermediate MTV and maximal diameter, and intermediate SUV$_{\text{max}}$. Positively correlated significant predictors (i.e. increasing value results in higher pCR rate) were the response indexes for SUV$_{\text{mean}}$, SUV$_{\text{max}}$ and MTV. Female gender was also found to be significantly associated with high pCR rate.

In multivariate logistic modelling only cT-stage was found significant in the total group of input predictors ($p = .027^*$). However, highly correlated input variables increase *p*-values in a multivariate setting (Fig. 2). The following decisions were made based on the analyses to select the final set of predictors:

- Gender was excluded: non-significant *p*-value and no other correlations with inputs

Table 3
Prediction results. For each variable, the distributions are compared between training and validation dataset and a univariate analysis is performed (training set only). Multivariate analysis including feature selection are included.

	Univariate		Multivariate			Predictor selection		
	p	pCR ↑	OR	[95% CI]	*p*	OR	[95% CI]	*p*
Gender	.048*	Female	0.65	[0.19–2.23]	.495			
Age	.523	–						
cT-stage	.002*	↓	0.20	[0.05–0.83]	.027*	0.19	[0.06–0.64]	.007*
cN-stage	.001*	↓	0.21	[0.04–1.04]	.056	0.23	[0.06–0.88]	.032*
SUV$_{\text{mean}}$0	.747	–						
SUV$_{\text{max}}$0	.617	–						
MTV0	.002*	↓	1.08	[1.00–1.17]	.061			
MaxDiam0	.004*	↓	0.90	[0.42–1.90]	.778			
SUV$_{\text{mean}}$15	.067	–						
SUV$_{\text{max}}$15	.030*	↓	0.96	[0.83–1.10]	.545			
MTV15	.000*	↓	0.82	[0.67–1.02]	.072			
MaxDiam15	.005*	↓	0.95	[0.58–1.56]	.849	0.74	[0.53–1.03]	.078
RI_SUV$_{\text{mean}}$	.022*	↑	1.01	[0.90–1.13]	.893	1.04	[1.00–1.07]	.025*
RI_SUV$_{\text{max}}$	.030*	↑	1.01	[0.91–1.13]	.839			
RI_MTV	.017*	↑	0.99	[0.96–1.03]	.757			
RI_MaxDiam	.544	–						

Significant *p*-values are indicated by an asterisk (*).

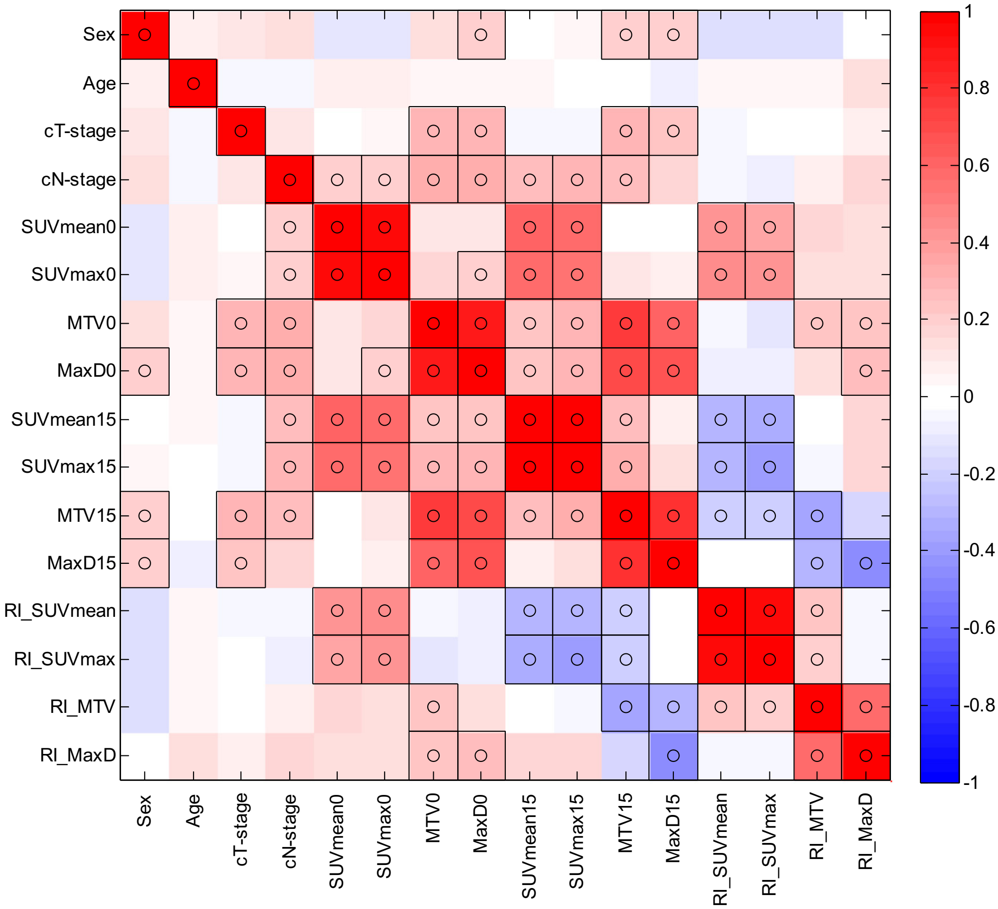

Fig. 2. Spearman correlation matrix to identify significant (boxed + inner circle) correlations between model input variables.

- cN-stage was included: significance near decision boundary ($p = .056$)
- The MTV measures were excluded: many outliers were detected (pre-treatment: $N = 11$ with MTV differences up to 4 times the average volume, intermediate: $N = 12$ with MTV differences up to 10 times the average volume). These measures also did not have an added predictive value to the final selection.
- RI of SUV$_{mean}$ was included at the expense of RI of SUV$_{max}$: literature reported sufficient evidence for both of them (Table 1). However, these measures are highly correlated (Fig. 2). Univariately, RI of SUV$_{mean}$ had highest discriminative ability ($p = .022^*$ vs $p = .030^*$) and was therefore selected.
- From the other predictors, pre-treatment and intermediate maximal diameter and intermediate SUV$_{max}$, only intermediate maximal diameter was selected because it showed an AUC increase >0.02 when added to the final set (and significance $\alpha < 0.1$).

Hence, the final selected predictors in the multivariate model were cT-stage ($p = .007^*$), cN-stage ($p = .032^*$), intermediate maximal diameter ($p = .078$) and RI of SUV$_{mean}$ ($p = .025^*$).

Validation of the nomogram

The multivariate model with the four selected predictors to estimate the probability of a pCR was visually represented by a nomogram (Fig. 3A). Bootstrapped AUCs were 0.78 (95% CI: 0.65–0.89) for the training dataset and 0.70 (95% CI: 0.55–0.84) for the validation dataset. With the aim of estimating three probability groups for pCR, two probability thresholds were defined to separate these groups. The 12.8% threshold (<12.8% low probability on pCR, >12.8% medium probability of pCR) was defined based on literature in which on average 49.2% of the patients are non-responders (weighted for number of patients). The other threshold was calculated at 53% based on decision curve analysis where the optimal net benefit of applying a wait-and-see policy was maximized (Fig. S1). These three probability groups (low, medium, high) resulted in significantly increasing pCR rates (training data: respectively 7.3%, 21.3%, 100% pCR; validation data: 13.0%, 30.8% and 66.7% pCR). The highest probability groups contained 8.9% (training) and 7.7% (validation) of the total number of patients, while the lowest probability groups contained respectively 49.1% and 59.0% (Fig. 3B).

Discussion

In this study a multivariate nomogram with clinical parameters and early sequential PETCT imaging markers predicting pCR in LARC was developed based on a large prospective study and validated prospectively within another institute. Good performances were measured for both training and validation dataset. After risk group identification a subgroup of just below 10% of the patients with high estimated probability for complete response was identified.

Model predictors

The selected predictive factors for pCR were cT-stage, cN-stage, ΔSUV$_{mean}$ and intermediate maximal tumour diameter. A recent large analysis of 3105 patients found that cT-stage was predictive for pCR ($p < .0001$) but that cN-stage shows only a trend ($p = .10$) [21]. This study contained however also old cases where CT was

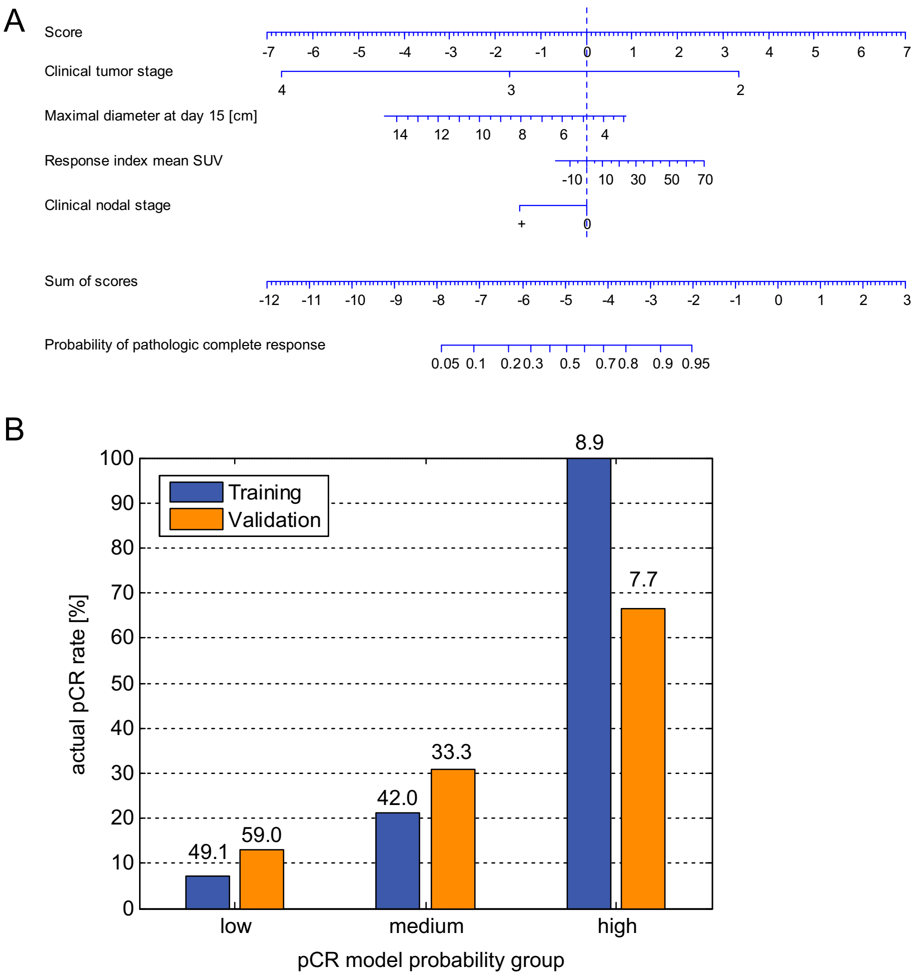

Fig. 3. (A) Nomogram of pathologic complete response based on the multivariate analysis. Three probability groups are defined and pCR rates of those groups are plotted. (B) Actual pCR rates in the training and validation set plotted for three probability groups from the model, which are defined by low: $p \leqslant 12.8\%$, medium: $12.8\% < p < 53\%$, high: $p \geqslant 53\%$. Relative number of patients in the group (%) is shown above the bars. The model is available on www.predictcancer.org.

used for cN-stage scoring. Another analysis of 677 patients associated both cT-stage ($p < .001$) and cN-stage ($p < .001$) with pCR [29]. The same study also found that pre-treatment (metabolic) tumour size was predictive ($p = .003$), but others show that only changes in metabolic volume in the pre-post treatment setting were significant for pCR and not the intermediate case ($p = .010$) [27]. In the presented study's univariate analysis, it is clear that tumour size is important for pCR. The change in SUV_{mean} at the intermediate time point was found predictive in earlier studies for responders versus non-responders [1,5,14], but SUV_{max} is stronger when predictions for pCR changes are made [10,16]. In our study the response index of SUV_{mean} was a stronger predictor than the RI of SUV_{max}. These two measures are also highly correlated, especially since our PET contouring was semi-automatic and calibrated for both institutes, resulting in less variation in SUV_{mean} due to contouring [4].

A variable selection scheme was chosen based on univariate analysis, correlation between input variables and contribution to multivariate prediction performance. Other strategies such as penalized feature selection are also common when dealing with highly correlated variables, but in general this results in interpretability issues for the (shrunken) coefficients [19].

Model performance

The performance of the nomogram measured by AUC of 0.78 (95% CI: 0.65–0.89) for the training dataset and 0.70 (95% CI: 0.55–0.84) for the validation dataset are lower than the ones reported by single PET parameters in literature (0.70 [25], 0.83 [11], 0.87 [14]). However, these studies predict response (TRG1–2) instead of pCR. Response prediction is in practice more accurate because the number of events for good response (45–55%) is much higher than that of pCR (15–30%). Identifying the complete responders in an early phase is useful to avoid additional treatments and related toxicity for these patients. Another possible reason for a lower overall performance can be the noisy pathology outcome (noncentralized), but this is compensated by the high number of patients. The current studies reported in literature with low number of samples are sensitive to positive (or negative) findings by mere chance, and therefore it is reasoned that this large study's performance is expected to reflect reality better.

When stratifying the patients in three risk groups, the performance is acceptable: 100% accuracy for high probability for pCR group for the training dataset ($N = 10$) and 67% for the validation dataset ($N = 6$). The two misclassified patients in the validation set have TRG2, ypT1N0 or ypT2N0 status and a clinical response on the PETCT scan two months after the end of CRT, thus they are considered as good responders. The model is very conservative in selecting patients with high probability for pCR because the aim is to keep the number of false positives as low as possible to avoid undertreatment in a wait-and-see policy. This means that patients that present a pCR eventually still end up in the predicted good and bad responder group as indicated in Supplementary Table S2. These patients might get overtreatment in case extra RT is administered for good responders, but we believe the increased pCR rate benefit for this group outweighs the possible mild complications.

The difference in performance between training and validation dataset is likely to be based on differently distributed data with respect to the predictors. The validation set has lower model estimates for the probabilities of pCR due to higher pretreatment cT-stages ($p = .007$), and significantly lower RIs of SUV_{mean} ($p < .001$). The latter difference may be explained by the higher times between tracer injection and the intermediate PET acquisition in the validation institute in comparison to the training institute, resulting in higher intermediate SUV_{mean} values ($p < .001$), despite PET protocol harmonization. Although the external validation is not independent due to the harmonized protocols, the discussed differences in image acquisition and treatment characteristics guarantee generalizability of the model to a certain level.

Model applicability

As suggested, the developed model can be used to assist in decision making for LARC already during CRT (Fig. 1). However, three notes have to be made. First, this model is only useful for decisions made during or immediately after CRT, like RT boost or additional chemotherapy. The decision for a wait-and-see approach can better be made just before surgery by using both specialized prediction models [29] and careful assessment of imaging, endoscopies and biopsies [20]. The advantages of an earlier estimate of pCR are avoidance of overtreatment of complete responders, a possible increase of the number of complete responders with a RT boost for good responders and perhaps a change in treatment strategy for non-responding patients.

Secondly, any developed model requires prospective validation by means of a randomized trial, comparing an arm with standard treatment for all (CRT + surgery) to an arm receiving individualized treatment based on the prediction model. Such a study is currently being set up.

And last, other predictors from different sources might be considered to further improve accuracy. Diffusion-weighted magnetic resonance imaging (DW-MRI) at different time points is reported as a promising candidate which increases the prediction accuracy significantly in combination with PETCT imaging [16]. Here it is important that also cost effectiveness studies are carried out to ensure that the costs for making multiple PET or MRI scans are justified when compared to cost reductions due to personalizing treatment with better outcomes and fewer complications. Blood biomarkers also can have additional value as for example has been reported for serum carcinoembryonic antigen (CEA) [22]. For all these additional sources however, cost–benefit analyses are advised because saturation of the prediction accuracy can become an issue.

Conclusions

We have developed an externally validated and accurate prediction model for pathologic complete response in locally advanced rectal cancer based on large prospective studies. This nomogram can be used to distinguish three types of patients, i.e. complete responders, good responders and non-responders, for which respectively a wait-and-see policy, radiotherapy boost and additional chemotherapy can be administered. This personalized treatment approach is expected to promote more complete responders, reduce the number of surgeries and related complications, and to avoid unnecessary toxicity.

Conflict of interest

The authors are not aware of any actual or potential conflict of interest.

Acknowledgments

Authors acknowledge financial support from the QuIC-ConCePT project, which is partly funded by EFPIA companies and the Innovative Medicine Initiative Joint Undertaking (IMI JU) under Grant Agreement No. 115151. Authors also acknowledge financial support from the National Institutes of Health (NIH-USA U01 CA 143062-01), the CTMM framework (AIRFORCE project, grant 030-103), EU 7th framework program (METOXIA, EURECA), euroCAT (IVA Interreg – www.eurocat.info) and the Dutch Cancer Society (KWF UM 2011-5020, KWF UM 2009-4454).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.radonc.2014.11.002.

Reference

[1] Avallone A, Aloj L, Caraco C, et al. Early FDG PET response assessment of preoperative radiochemotherapy in locally advanced rectal cancer: correlation with long-term outcome. Eur J Nucl Med Mol Imaging 2012;39:1848–57.

[2] Barbaro B, Vitale R, Valentini V, et al. Diffusion-weighted magnetic resonance imaging in monitoring rectal cancer response to neoadjuvant chemoradiotherapy. Int J Radiat Oncol Biol Phys 2012;83:594–9.

[3] Braendengen M, Tveit KM, Berglund A, et al. Randomized phase III study comparing preoperative radiotherapy with chemoradiotherapy in nonresectable rectal cancer. J Clin Oncol 2008;26:3687–94.

[4] Buijsen J, van den Bogaard J, van der Weide H, et al. FDG-PET-CT reduces the interobserver variability in rectal tumor delineation. Radiother Oncol 2012;102:371–6.

[5] Cascini GL, Avallone A, Delrio P, et al. 18F-FDG PET is an early predictor of pathologic tumor response to preoperative radiochemotherapy in locally advanced rectal cancer. J Nucl Med 2006;47:1241–8.

[6] Daisne JF, Sibomana M, Bol A, Doumont T, Lonneux M, Gregoire V. Tri-dimensional automatic segmentation of PET volumes based on measured source-to-background ratios: influence of reconstruction algorithms. Radiother Oncol 2003;69:247–50.

[7] Engelen SM, Beets-Tan RG, Lahaye MJ, Kessels AG, Beets GL. Location of involved mesorectal and extramesorectal lymph nodes in patients with primary rectal cancer: preoperative assessment with MR imaging. Eur J Surg Oncol 2008;34:776–81.

[8] Gerard JP, Chapet O, Nemoz C, et al. Improved sphincter preservation in low rectal cancer with high-dose preoperative radiotherapy: the lyon R96–02 randomized trial. J Clin Oncol 2004;22:2404–9.

[9] Gerard JP, Conroy T, Bonnetain F, et al. Preoperative radiotherapy with or without concurrent fluorouracil and leucovorin in T3–4 rectal cancers: results of FFCD 9203. J Clin Oncol 2006;24:4620–5.

[10] Goldberg N, Kundel Y, Purim O, et al. Early prediction of histopathological response of rectal tumors after one week of preoperative radiochemotherapy using 18 F-FDG PET-CT imaging. A prospective clinical study. Radiat Oncol 2012;7:124.

[11] Guerra L, Niespolo R, Di Pisa G, et al. Change in glucose metabolism measured by 18F-FDG PET/CT as a predictor of histopathologic response to neoadjuvant treatment in rectal cancer. Abdom Imaging 2011;36:38–45.

[12] Janssen MH, Ollers MC, Riedl RG, et al. Accurate prediction of pathological rectal tumor response after 2 weeks of pre-operative radiochemotherapy using FDG-PET-CT imaging. Int J Radiat Oncol Biol Phys 2009.

[13] Janssen MH, Ollers MC, van Stiphout RG, et al. Blood glucose level normalization and accurate timing improves the accuracy of PET-based treatment response predictions in rectal cancer. Radiother Oncol 2010;95:203–8.

[14] Janssen MH, Ollers MC, van Stiphout RG, et al. PET-based treatment response evaluation in rectal cancer: prediction and validation. Int J Radiat Oncol Biol Phys 2012;82:871–6.

[15] Lambin P, van Stiphout RG, Starmans MH, et al. Predicting outcomes in radiation oncology – multifactorial decision support systems. Nat Rev Clin Oncol 2013;10:27–40.

[16] Lambrecht M, Deroose C, Roels S, et al. The use of FDG-PET/CT and diffusion-weighted magnetic resonance imaging for response prediction before, during and after preoperative chemoradiotherapy for rectal cancer. Acta Oncol 2010;49:956–63.

[17] Leibold T, Akhurst TJ, Chessin DB, et al. Evaluation of (1)(8)F-FDG-PET for early detection of suboptimal response of rectal cancer to preoperative chemoradiotherapy: a prospective analysis. Ann Surg Oncol 2011;18:2783–9.

[18] Ludbrook J. Outlying observations and missing values: how should they be handled? Clin Exp Pharmacol Physiol 2008;35:670–8.

[19] Ma S, Huang J. Penalized feature selection and classification in bioinformatics. Brief Bioinform 2008;9:392–403.

[20] Maas M, Beets-Tan RG, Lambregts DM, et al. Wait-and-see policy for clinical complete responders after chemoradiation for rectal cancer. J Clin Oncol 2011;29:4633–40.

[21] Maas M, Nelemans PJ, Valentini V, et al. Long-term outcome in patients with a pathological complete response after chemoradiation for rectal cancer: a pooled analysis of individual patient data. Lancet Oncol 2010;11:835–44.

[22] Moureau-Zabotto L, Farnault B, de Chaisemartin C, et al. Predictive factors of tumor response after neoadjuvant chemoradiation for locally advanced rectal cancer. Int J Radiat Oncol Biol Phys 2011;80:483–91.

[23] Ollers M, Bosmans G, van Baardwijk A, et al. The integration of PET-CT scans from different hospitals into radiotherapy treatment planning. Radiother Oncol 2008;87:142–6.

[24] Pepe MS. Receiver operating characteristic methodology. J Am Stat Assoc 2000;95.

[25] Rosenberg R, Herrmann K, Gertler R, et al. The predictive value of metabolic response to preoperative radiochemotherapy in locally advanced rectal cancer measured by PET/CT. Int J Colorectal Dis 2009;24:191–200.

[26] Strauss LG. Positron emission tomography: current role for diagnosis and therapy monitoring in oncology. Oncologist 1997;2:381–8.

[27] Sun W, Xu J, Hu W, Zhang Z, Shen W. The role of sequential 18(F)-FDG PET/CT in predicting tumour response after preoperative chemoradiation for rectal cancer. Colorectal Dis 2013;15:e231–238.

[28] Valentini V, Lambin P, Myerson RJ. Is it time for tailored treatment of rectal cancer? from prescribing by consensus to prescribing by numbers. Radiother Oncol 2012;102:1–3.

[29] van Stiphout RG, Lammering G, Buijsen J, et al. Development and external validation of a predictive model for pathological complete response of rectal cancer patients including sequential PET-CT imaging. Radiother Oncol 2011;98:126–33.

[30] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26:565–74.

[31] Zhang C, Tong J, Sun X, Liu J, Wang Y, Huang G. 18F-FDG-PET evaluation of treatment response to neo-adjuvant therapy in patients with locally advanced rectal cancer: a meta-analysis. Int J Cancer 2012;131:2604–11.