Development and external validation of a predictive model for pathological complete response of rectal cancer patients including sequential PET-CT imaging

Radiotherapy and Oncology 98 (2011) 126–133

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

PET in radiotherapy

Development and external validation of a predictive model for pathological complete response of rectal cancer patients including sequential PET-CT imaging

Ruud G.P.M. van Stiphout [a,b,*], Guido Lammering [a,1], Jeroen Buijsen [a,1], Marco H.M. Janssen [a], Maria Antonietta Gambacorta [c], Pieter Slagmolen [d,e], Maarten Lambrecht [d], Domenico Rubello [f], Marcello Gava [g], Alessandro Giordano [h], Eric O. Postma [i], Karin Haustermans [d,1], Carlo Capirci [j,1], Vincenzo Valentini [c,1], Philippe Lambin [a,1]

[a] Department of Radiation Oncology (MAASTRO), Maastricht University Medical Centre, The Netherlands; [b] Department of Knowledge Engineering, Maastricht University, The Netherlands; [c] Radiotherapy Department, Università Cattolica S.Cuore, Rome, Italy; [d] Department of Radiation Oncology, University Hospital Leuven, Belgium; [e] Medical Image Computing, Departments of ESAT and Radiology, Catholic University Leuven, Belgium; [f] Department of Nuclear Medicine; and [g] Department of Physics, S. Maria della Misericordia State Hospital, Rovigo, Italy; [h] Nuclear Medicine Department, Università Cattolica S.Cuore, Rome, Italy; [i] Tilburg Centre for Cognition and Communication, Tilburg University, The Netherlands; [j] Division of Radiotherapy, S. Maria della Misericordia Hospital, Rovigo, Italy

ARTICLE INFO

Article history:
Received 3 May 2010
Received in revised form 23 November 2010
Accepted 5 December 2010

Keywords:
Response prediction
PET imaging
Machine learning
Rectal cancer
External validation

ABSTRACT

Purpose: To develop and validate an accurate predictive model and a nomogram for pathologic complete response (pCR) after chemoradiotherapy (CRT) for rectal cancer based on clinical and sequential PET-CT data. Accurate prediction could enable more individualised surgical approaches, including less extensive resection or even a wait-and-see policy.

Methods and materials: Population based databases from 953 patients were collected from four different institutes and divided into three groups: clinical factors (training: 677 patients, validation: 85 patients), pre-CRT PET-CT (training: 114 patients, validation: 37 patients) and post-CRT PET-CT (training: 107 patients, validation: 55 patients). A pCR was defined as ypT0N0 reported by pathology after surgery. The data were analysed using a linear multivariate classification model (support vector machine), and the model's performance was evaluated using the area under the curve (AUC) of the receiver operating characteristic (ROC) curve.

Results: The occurrence rate of pCR in the datasets was between 15% and 31%. The model based on clinical variables ($AUC_{train} = 0.61 \pm 0.03$, $AUC_{validation} = 0.69 \pm 0.08$) resulted in the following predictors: cT- and cN-stage and tumour length. Addition of pre-CRT PET data did not result in a significantly higher performance ($AUC_{train} = 0.68 \pm 0.08$, $AUC_{validation} = 0.68 \pm 0.10$) and revealed maximal radioactive isotope uptake (SUV_{max}) and tumour location as extra predictors. The best model achieved was based on the addition of post-CRT PET-data ($AUC_{train} = 0.83 \pm 0.05$, $AUC_{validation} = 0.86 \pm 0.05$) and included the following predictors: tumour length, post-CRT SUV_{max} and relative change of SUV_{max}. This model performed significantly better than the clinical model ($p_{train} < 0.001$, $p_{validation} = 0.056$).

Conclusions: The model and the nomogram developed based on clinical and sequential PET-CT data can accurately predict pCR, and can be used as a decision support tool for surgery after prospective validation.

Over the past decades, treatment outcomes for rectal cancer have changed dramatically. A better surgical technique, total mesorectal excision (TME), and the introduction of neoadjuvant treatments in locally advanced rectal cancer (LARC) have significantly decreased the risk of locoregional relapse [1,2]. In the last nine years at least seven published phase III trials have evaluated the role of adjuvant radiotherapy in rectal cancer [3]. These have provided an evidence base demonstrating the efficacy of both preoperative radiotherapy and preoperative concurrent chemotherapy (CRT). CRT has been reported to induce significant tumour downsizing and downstaging, [4–6] with a pathologic complete response (pCR) after CRT observed in 10%–30% of patients [2,4–8]. Although some studies showed no correlation, [9] many others reported that patients showing a pCR following preoperative CRT have improved long-term outcomes including excellent local control rates and disease-free survival, regardless of their initial clinical T- and N-stages [10–13].

However, despite the often phenomenal downsizing and sometimes even complete pathological responses after CRT, these patients are still operated with a standard extended surgical procedure due to the lack of reliable accurate preoperative

* Corresponding author. Address: Department of Radiation Oncology, 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).

[1] These authors equally contributed to this manuscript.

0167-8140/$ - see front matter © 2010 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.radonc.2010.12.002

diagnostic tools. However, it may be questioned whether a standard resection is still necessary, considering the good outcome of these patients reported with less invasive treatments [14,15]. If accurately selected, patients with a complete response (no residual tumour) may undergo a less extensive resection or even a so called 'wait-and-see' policy. Compared to standard surgery, the benefits of these treatments are reduced morbidity and mortality (e.g., anastomotic leakage, relaparotomy, wound and pelvic infection, abscess, colostomy, chronic wound healing disturbances, faecal or urinary incontinence and sexual dysfunction), improved quality of life and reduced treatment costs.

Thus, an accurate prediction of pCR can help in the selection of patients for more optimised treatment, sphincter-preserving surgery, less extensive resection, more intense radiation treatment, or even delayed surgery with a wait-and-see policy [2,3,16]. These considerations led to the overall goal of this study: to develop an accurate, data-driven model to predict pathologic complete response for rectal cancer patients as decision support for more individualised treatment approaches in the future.

The clinical variables associated with a better response to preoperative CRT include circumferential tumour extent, tumour differentiation, preoperative classification, carcinoembryonic antigen (CEA) level, distance from anal verge, and time to surgery [6,17,18]. Recently, it has also been suggested that PET imaging might be correlated with tumour response after CRT in locally advanced rectal cancer. However, the studies involved used only a small number of patients, which meant that contradictory results were found. Further, only semi-quantitative PET measurements were used and analysed with univariate statistics [4,5,7,19–26]. Multivariate analysis was performed in only one study, whose results lacked statistical significance [27]. Notably, no studies verified and validated their results with external datasets, despite the fact that this represents an important prerequisite for the generalizability of prediction models for other institutes.

In the current study, population based data from four different institutes were collected and used to train and validate predictive models for pCR. We hypothesised that the addition of PET imaging data to clinical variables significantly increases the performance of prediction models for pCR after CRT as compared to models based on clinical data alone.

The study was performed within the framework of a decision support system based on centralised datasets. The increasing amount of available patient information requires automatic methods for model building and analysis. Machine learning methods can be used to update the models continuously by feeding them with information of new patients. The increasing complexity of prediction models, too, means that the representation and interpretation of the results also become more important. Tools to enhance interpretation for the clinic include visualisation techniques such as nomograms and graphical networks. Nomograms are statistical tools that enable users to calculate the overall probability of a specific clinical outcome for an individual patient [28]. In this study, the nomogram with the highest accuracy for the prediction of pCR is provided.

Methods and materials

Study population

Six population based datasets were collected from four institutes: Maastro Clinic (GROW, MUMC, Maastricht, the Netherlands), Università Cattolica del S.Cuore (Rome, Italy), S. Maria della Misericordia Hospital (Rovigo, Italy) and University Hospital Gasthuisberg (Leuven, Belgium). In total, 953 patients met the criteria for inclusion: long-course RT with neoadjuvant chemotherapy and the availability of pathological outcome for pCR. Of these, 276 patients underwent a pre-CRT PET scan (one week before the start of CRT), and 169 patients had both pre- and post-CRT PET scans (one week before surgery, and six to eight weeks after the end of CRT). The sequential PET data from Rovigo have already been published as a prospective study [20], the Leuven data were collected prospectively for the BioCare project (LSHC-CT-2204-505785) and the rest of the data were gathered for a population-based study registered in the Dutch Trial Register (NTR2166). All compositions of the cohorts were approved by the local IRB committees. The patient characteristics are reported in Table 1. The datasets were divided into three groups, based on PET data availability: (1) clinical variables only, (2) clinical variables with pre-CRT PET variables (PET-pre), (3) clinical variables with both pre- and post-CRT PET variables (PET-post). For each group, a training set and an external validation set were defined. The training sets were used to identify the pCR predictors, while the validation sets were used to test the performance of the models in other centres. Datasets from a single centre with the highest number of patients were used for training. A dataset was deemed not useful for external validation if it originated from the same centre as the corresponding training set. The definition of the different combined training and validation sets is explained in Table 2, based on the datasets in Table 1.

The available clinical variables were age, gender (0: female, 1: male), clinical tumour (cT) and nodal (cN) stage, and two variables based on MRI (or endoscopy if MRI was unavailable): tumour location categorised in three levels (1: low, 0–5 cm from anal verge; 2: mid, 5–10 cm from anal verge; 3: high, >10 cm from anal verge) and tumour length (cm). For the patients who had PET-CT scans, the tumours were semi-automatically contoured at Maastro Clinic using dedicated software (TrueD, Siemens Medical, Erlangen, Germany). Standardised uptake-value (SUV) thresholding was based on the tumour-to-background signal ratio, with the gluteus muscle as reference background [29,30]. From the resulting tumour contour, maximal tumour diameter (MaxD), gross tumour volume (GTV), and maximal and mean SUV values within the GTV were calculated. If the post-CRT PET-CT scan was available, the same variables were scored, and a response index (RI) for each variable was calculated. For variable X, the response index is the relative percent difference between the value of the post-CRT and pre-CRT and it was defined as $RI = (X_{pre} - X_{post})/X_{pre} \times 100\%$. Thus, six variables were evaluated for the clinical dataset, 10 for the PET-pre dataset and 18 for the PET-post dataset. From these sets, the models selected subgroups of variables with significant predictive value for pCR.

All patients underwent surgery. 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, making the pCR a binary outcome (0/1). 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).

Statistical analysis

Missing values in the dataset were substituted by the mean [31]. This method performed similar to other, more complex substitution methods for small percentages of missing values (e.g., expectation–maximisation imputation, regression estimation). No variables in the datasets exceeded 5% of missing values. Patients who missed tumour location and length in the clinical datasets (Roma: $n = 132$ and Maastricht: $n = 29$) were excluded because of too large amounts of missing data for these variables. All patient numbers stated in this paper were extracted after the missing value procedure. To compare the weights of significance assigned to

Table 1
Patient characteristics for six datasets from four different institutes. Clinical, PET-pre and PET-post groups are defined. Percentages of the total patient numbers are given for binary or ordinal variables. Mean and standard deviation (SD) are given for continuous variables. *x* denotes missing values. RT = Radiotherapy, PF = per fraction.

Center	Maastricht		Rome		Rovigo	Leuven
Dataset	M1	M2	R1	R2	C1	L1
Period	2004–2006	2004–2006	1984–2008	2007–2008	2003–2007	2005–2007
# Patients	114	21	677	18	107	16
Clinical	Validation	–	Training	–	–	–
PET-pre	Training	–	–	Validation	–	Validation
PET-post	–	Validation	–	Validation	Training	Validation
Gender (%)						
Male	63	67	63	83	74	81
Female	37	33	37	17	26	19
Age						
Mean	65.6	66.1	61.3	60.4	66.3	58.6
SD	10.0	10.6	10.2	7.1	10.8	10.1
cT (%)						
1	0	0	0	0	0	0
2	1	0	3	11	0	0
3	68	81	86	56	90	94
4	30	14	11	33	10	6
x	1	5	0	0	0	0
cN (%)						
0	25	38	23	17	51	0
1	48	48	45	33	38	62
2	26	10	30	50	10	38
x	1	4	2	0	1	0
cM (%)						
0	73	71	100	94	100	100
1	25	19	0	6	0	0
x	2	10	0	0	0	0
ypT0N0 (%)						
No	85	81	80	78	76	69
Yes	15	19	20	22	24	31
RT dose						
Mean	50.4	50.4	49.0	52.7	55.7	45.7
SD	0	0	5.5	3.3	3.1	1.8
RT dose PF	1.8	1.8	1.8	1.8	2.2	1.8
# Chemo types	1	1	11	2	1	1

Table 2
Predictor selection and ROC analysis. Predictive variables are given with their corresponding assigned normalised weights from multivariate analysis (MVA). For each variable the p-value from univariate analysis (UVA) is given. Mean AUC and standard deviation (SD) are given for each variable set. RI = response index, SUV = standard uptake value, MaxD = maximal diameter (PET-CT).

Variable set	Type	Size	Predictors (MVA)	Weights (MVA)	*p*-value (UVA)	AUC	SD
Clinical	Training (R1)	677	Tumour length	−0.085	<0.001	0.61	0.03
			cT-stage	−0.074	0.001		
			cN-stage	−0.060	0.001		
	Validation (M1)	85	–	–	–	0.69	0.08
Clinical + PET-pre	Training (M1)	114	$MaxD_{pre}$	−0.12	0.003	0.68	0.08
			cN-stage	−0.12	0.001		
			Tumour location	0.094	0.84		
			$SUV_{max\text{-}pre}$	−0.087	0.29		
	Validation (R2, L1)	34	–	–	–	0.68	0.10
Clinical + PET-pre + PET-post	Training (C)	107	RI_{SUVmax}	0.20	<0.001	0.83	0.05
			Tumour length	−0.20	<0.001		
			$SUV_{max\text{-}post}$	−0.14	<0.001		
	Validation (M2, R2, L1)	55	–	–	–	0.86	0.05

the variables by the model, all variables were normalised by subtracting the mean, and then divided by the standard deviation.

To classify the complete responders and non-responders, a linear multivariate method suitable for binary classification from the machine learning field was used: the support vector machine (SVM) [32]. The SVM variant used (proximal SVM or pSVM) performs equally accurately but much faster than normal support vector machines [33]. The different datasets' performances in predicting pCR were evaluated by analysing the area under the curve (AUC) of the receiver operating characteristic (ROC) curve [34]. The maximum value of the AUC is 1.0, indicating a perfect prediction model; a value of 0.5 indicates a random chance of correct prediction.

To select the variables that contribute to pCR prediction, an exhaustive feature search was performed, with all possible variable combinations used as input for the pSVM model. The set of vari-

ables resulting in the highest AUC was selected as the final predictive set. To avoid over-fitting of the model through selection of the highest AUC, the variable sets resulting in AUCs that deviated less than 5% from the maximal AUC were compared to the final variable set. If conflicts occurred or if variables did not contribute significantly, selected variables were interchanged by considering their prevalence in the highly predictive sets, the factor analysis and the Spearman correlation coefficient (i.e., highly correlated and dependent variables are not present in the same predictive set). Furthermore, an extra univariate analysis was performed using the Wilcoxon rank sum test.

Classification methods normally require at least several hundred cases. Because of the relatively small number of available patients, two extra evaluation methods were used. The first was leave-one-out (LOO) cross-validation, used to calculate an AUC for the training set. In LOO cross-validation, a single patient is selected from the original training dataset and used as the validation dataset, while the data from the remaining patients are used to train the model. This is repeated until all patients have been selected once for validation. However, no LOO cross-validation was used for the external dataset. The second evaluation method was bootstrapping, which results in a more accurate approximation of the real dataset distribution [35]. This means that 1000 datasets are generated from the original dataset containing n patients by selecting these n patients, but with resampling (i.e., patients can be present in the dataset more than once). For every bootstrapped dataset, an AUC was calculated. The mean AUC with the corresponding standard deviation was then calculated with size 1000. This non-parametric method allows comparison of the confidence intervals of the AUCs of different datasets without making assumptions about the AUC distributions [36]. The distribution of the difference in mean AUC (ΔAUC) between the datasets was tested by calculating the two-sided p-value, i.e., the fraction of ΔAUC samples smaller or larger than zero (depending on the dominant sign of ΔAUC).

Nomograms can reduce statistical predictive models to a single numerical estimate of the probability of an event, and visualise the effect of each selected variable on this probability [37]. The model output of the pSVM models consists of assigned weights for each variable and an offset. The probability of a patient having a pCR

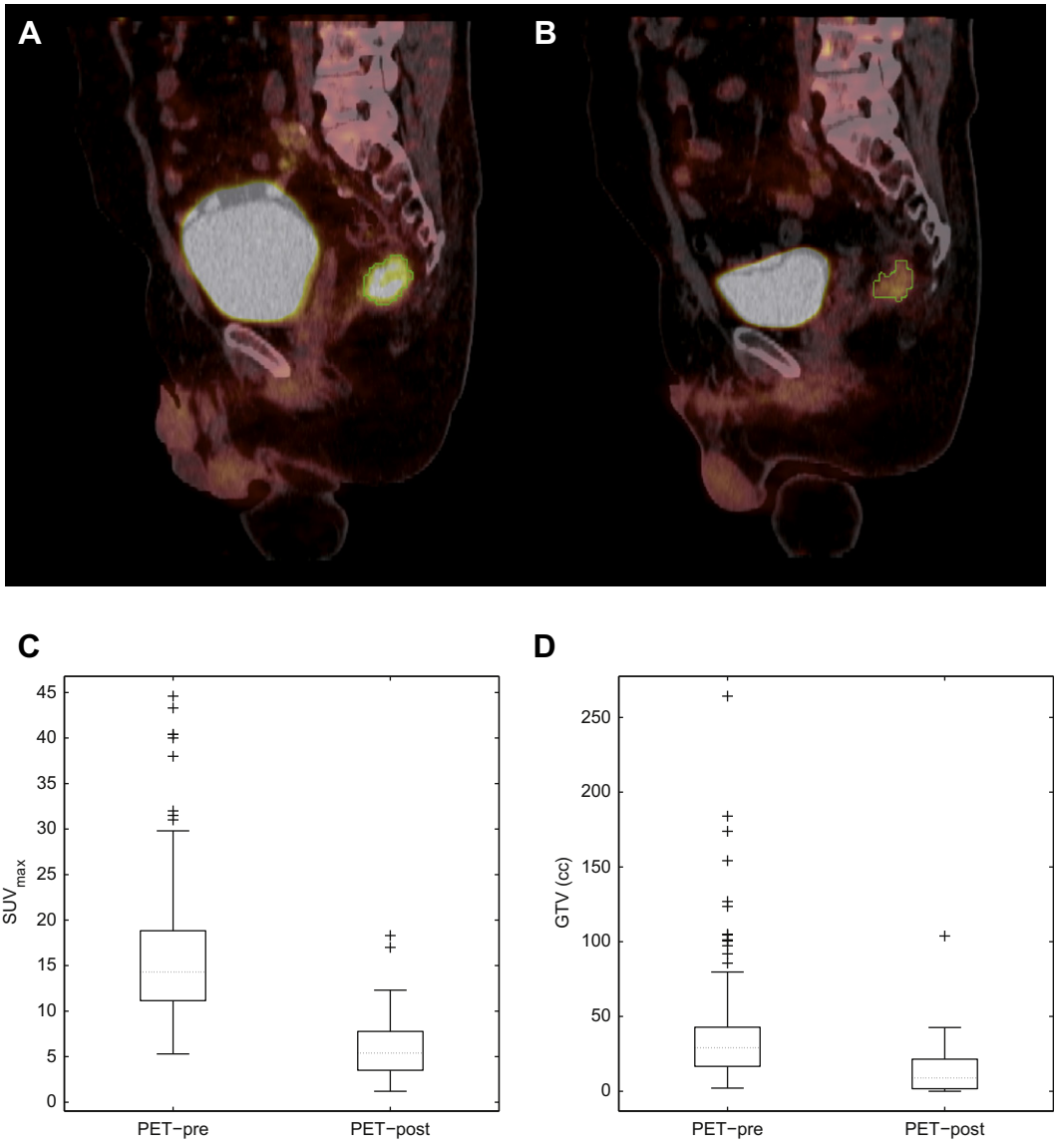

Fig. 1. (A) Tumour contour in a fused FDG-PET-CT made pre-CRT. (B) Corresponding post-CRT FDG-PET-CT scan with tumour contour. (C) Boxplot of SUV_{max} on PET-scans made pre-CRT and post-CRT (significant decrease: $p < 0.001$). (D) Boxplot of the GTV for the case of pre-CRT and post-CRT (significant decrease: $p < 0.001$).

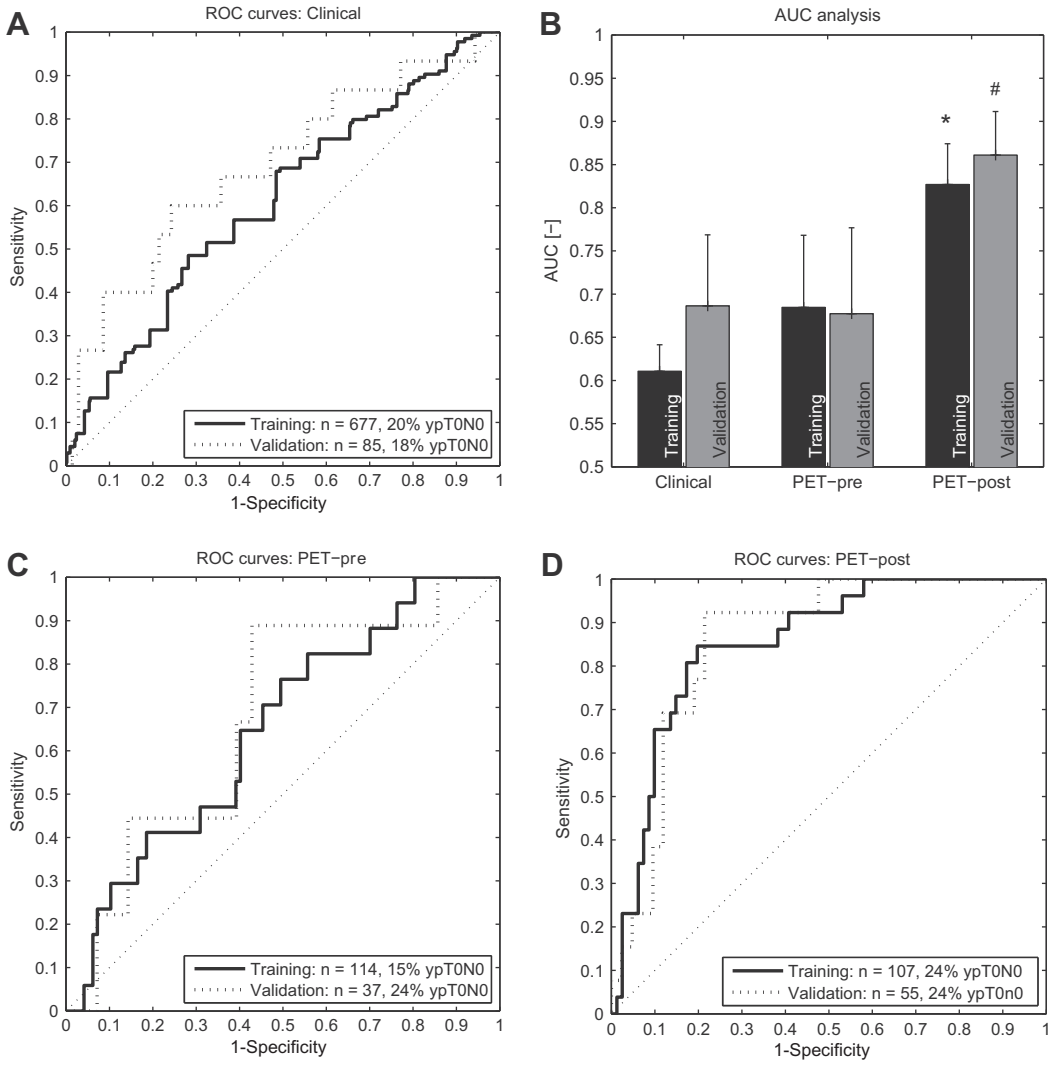

Fig. 2. ROC curves of training and validation datasets for the clinical set (A), the PET pre-CRT set (C) and the PET post-CRT set (D). The straight dashed line represents a random prediction model. The bar plot (B) shows the corresponding mean AUC for each dataset and the standard deviation (error bars). There was a significant difference with clinical datasets of (∗) $p < 0.05$ and (#) $p < 0.06$.

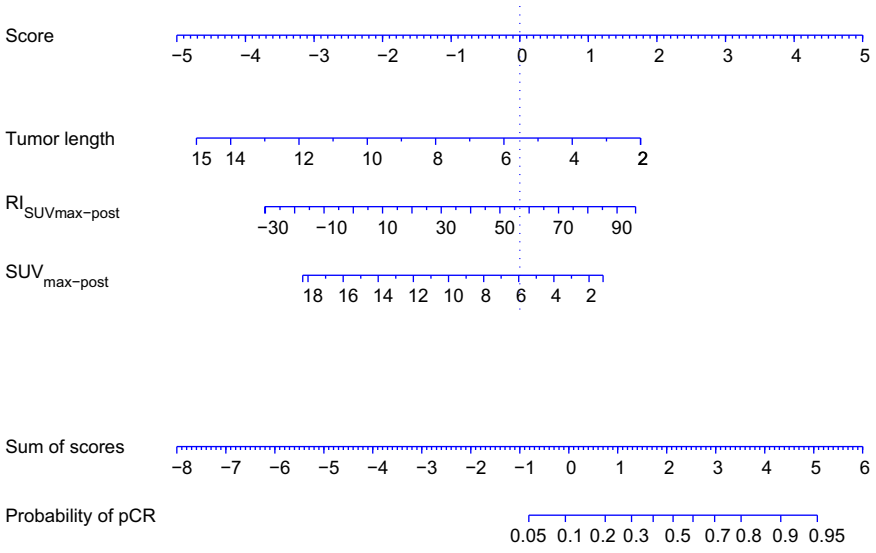

Fig. 3. Nomogram for PET post-CRT dataset. A score for each predictor can be read out at the top scale (Score). All summed scores (Sum of scores scale) can be converted directly to the probability of responding with a pCR (ypT0N0). The probability scale is the only logarithmic scale.

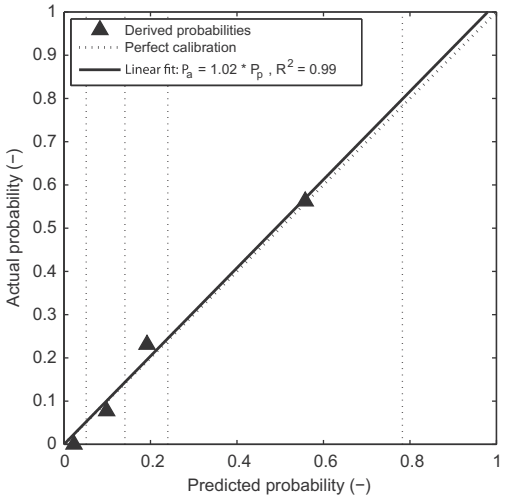

Fig. 4. Calibration of the nomogram for the validation data. For the four equally numbered subgroups (vertical lined intervals in figure), the predicted probability of a pCR and the actual fraction in the population were evaluated. The dashed line represents perfect calibration and the solid line is the linear fit of the calibration data.

can be calculated using logistic regression on the pSVM output [38]. The complete procedure to convert SVM output to a nomogram is described in detail elsewhere [39]. Developing a nomogram requires threshold selection in the ROC curve. For response prediction specificity is most important, because it is not preferred to predict non-responders as responders, which would result in under-treatment. Therefore, the threshold was selected in such a way that at least 90% of non-responders were correctly predicted. Partial ROC curve optimisation [40] has been tested but it had no gain for specificity compared to overall AUC maximisation. Calibration of the nomogram, i.e., the agreement between predicted probability of complete response and true probability in the population, was performed by an assessment of the overall agreement and the Hosmer–Lemeshow statistic in four subgroups of patients in the validation data. The nomogram algorithm was implemented in MATLAB (version 7.1, MathWorks Inc., Natick, MA), as were all algorithms described in this section.

Results

The occurrence of pCR in the patient population varied between 15% and 31% (mean: 21.8%, SD: 5.4%) depending on the dataset (Table 1). A first evaluation of CRT's effect on the tumour demonstrated significant downsizing of the tumour in the PET-CT, and a significant decrease in metabolic activity within the tumour (Fig. 1). Both gross tumour volume and maximal SUV decreased significantly between the pre- and post-CRT PET-CT scans ($p < 0.001$).

Table 2 shows the predictor selection results and the ROC curve analysis. For the clinical dataset, the univariate analysis reveals three variables significantly associated with pCR (95% confidence interval): tumour length ($p < 0.001$), cN-stage ($p = 0.001$), and cT-stage ($p = 0.001$). These variables were also selected in the multivariate analysis. The normalised weights assigned to them by the pSVM model are tumour length (−0.085), cT-stage (−0.074), and cN-stage (−0.060). The selected variables were ranked in importance (i.e., weights). The sign of the weights can be interpreted by the effect on the probability of a pCR. For a negative sign, this probability decreases when the variable increases. For the clinical dataset, this means that the probability of a pCR increases for small tumour lengths and low cT- and cN-stages. The predictive performance of the clinical dataset for pCR, expressed by the AUC of the ROC curve, is 0.61 ± 0.03 (mean ± SD) for the training set and 0.69 ± 0.08 for the external validation set.

For the dataset with pre-CRT PET data, the multivariate analysis selected these variables (ranked by weight): maximal diameter (−0.12), cN-stage (−0.12), tumour location (0.094), and SUV_{max} (−0.087). This resulted in a high probability of pCR for patients with small maximal tumour diameters, low cN-stage, high tumour locations, and small maximal metabolic activity. Maximal diameter ($p = 0.003$) and cN-stage ($p = 0.001$) were selected by univariate analysis, while the other two variables were not. The AUCs for the training and validation sets were both 0.68, but the SD differed (0.08 and 0.10, respectively).

The dataset including the post-CRT PET data resulted in the highest performance: $AUC_{train} = 0.83 \pm 0.05$ and $AUC_{validation} = 0.86 \pm 0.05$. The response index for SUV_{max} (0.20), the tumour length (−0.20), and the post-CRT SUV_{max} were found to be predictive for pCR and significantly associated with pCR in the univariate analysis ($p < 0.001$).

In evaluating the predictive value of the additional PET data to the clinical data, only the AUCs of the post-CRT PET data differed significantly from the clinical dataset AUC (Fig. 2). The p-value for the AUC difference for the training set was <0.001, while that for the validation sets was 0.056 (just outside the 95% confidence interval). When only post-CRT PET data were used for the models (i.e., no clinical variables), the significant difference between the AUCs and the clinical dataset was no longer observed (training: $p = 0.47$, validation: $p = 0.58$). This indicated that a combination of both clinical and PET data was required to reach a significantly higher performance when using PET as a predictive imaging modality.

The assigned weights for all the predictors formed the basis for the construction of the nomogram. The nomogram based on the post-CRT dataset is provided in Fig. 3. The nomogram performs with a sensitivity of 0.62 and a specificity of 0.88 for the validation data. In the training phase these were respectively 0.65 and 0.90. The calibration of the nomogram (Fig. 4) with the validation data reveals that the overall predicted and the actual probability are equal (23.6%, OR = 1.0). If the validation data are divided into four equally numbered groups, the Hosmer–Lemeshow test results in a p-value of 0.78, which means a good calibration in this test ($p > 0.05$). The linear fit through these probabilities results in a slope of 1.02 with R^2 of 0.99, confirming a good balance between calibration and discrimination.

Discussion

We have developed predictive models based on clinical and PET-based data for pathologic complete response in patients diagnosed with rectal cancer. The performance of these models was externally validated using patient cohorts from different institutes treated with long-course preoperative chemoradiotherapy. The models showed that the accuracy of the predictions increased over time, i.e., when more information became available. Information from PET-CT scans significantly improved the performance of the models.

The significant difference in AUCs that we reported between the performance of the clinical model and the post-CRT PET data model reflects what others have found in their post-treatment PET analyses; like us, some have reported (significant) indications that the response index and post-treatment SUV_{max} are predictive for response, while the pre-treatment PET data do not provide enough predictive power [12,19,27]. However, our PET-based models also contain clinical variables, which appeared to be necessary to obtain the high performance provided in Table 2. The most important

clinical variables were tumour length and maximal diameter, which were selected in the models and are significantly correlated (spearman $\rho = 0.55$, $p < 0.001$). Overall, this means that the dominant tumour dimension in combination with (differences in) the maximal metabolic activity inside the tumour is the most predictive variable set for pCR, which was confirmed in the external datasets.

Whether the corresponding AUC of 0.86 is accurate enough for clinical practice depends on the choice of the threshold in the ROC curve. A high specificity is preferred over a high sensitivity to avoid possible under-treatment (less surgery when surgery is required) rather than over-treatment (standard treatment when less surgery could have been considered). The provided nomogram focuses on specificity (training: 0.90, validation 0.88). Selecting higher specificities results in fast decreasing sensitivities. Careful follow-up is therefore necessary for the patients selected for a 'wait-and-see' policy to detect any possible local recurrences early on. To gain more specificity in the future, the addition of new variables and the other classification methods would have to be considered.

The nomogram performs well, i.e., the distribution of the probability of a pCR provided by the nomogram represents the true distribution in the data, confirmed by overall calibration, calibration of the slope and Hosmer–Lemeshow test (Fig. 4). Because of the number of events and the division of the patient cohorts into few probability intervals, the higher probabilities occur much less frequently and are thus the least accurate. Therefore, prospective validation of the model and the nomogram is required to ensure sufficient statistical power for clinical application of the models. Besides the number of patients to increase the models' accuracy, more predictors could be added to increase the models' performance, including biological variables such as gene signatures [41] and blood biomarkers, and also more imaging variables from (perfusion) CT and (diffusion) MRI. The first indications have also appeared that PET-CT data during CRT may be highly predictive for response [25,26,42]. This time point is more favourable than post-CRT because of the possibility of earlier treatment changes and the decreased presence of inflammatory rectum cases, potentially causing impaired evaluation of fused PET-CT scans. After prospective validation of the model, an intervention trial with less surgery for patients with a high probability for pCR will be performed.

The population based collected datasets date back five years, except for the clinical Roma database, which was collected from 1984 onward. Therefore, this dataset shows a higher variety in treatment schemes than the other datasets. This could explain the discrepancy of the higher prediction performance of the clinical validation set. On the other hand, the validation set is much smaller, implying that the distribution of data could not be representative of the true distribution. The consequence of population based data collection is that treatment protocols are not well tuned. This results in, for example, small differences in irradiation schemes and deviations in the evaluation of pathology outcome. Ideally, pathology is reviewed centrally to reduce the intra- and inter-observer variabilities for the outcome measure. However, in this study the quality of pathology is acceptable because of the prospective nature of most datasets and because the outcome was limited to only complete response evaluation. Also, glucose correction for SUV values was not applied to all datasets. However, minor variation in treatment schemes can be seen as an advantage because it leads to higher generalizability for other centres. In other words, the model still performs well, despite the disparities mentioned here.

In conclusion, we have shown that sequential PET-CT data in combination with clinical variables significantly increase the performance of prediction models for pathologic complete response. So far, this is the largest study of its kind and the only one that used external datasets for validation. The dominant tumour dimension and the maximal uptake of radioactive isotopes in the tumour as well as its relative difference between PET scans were found to be the best predictors for pCR resulting in very good overall performance AUC's of 0.83 and 0.86 for training and validation, respectively. Including also biological and other imaging variables will probably further improve the performance. When prospectively validated, the model and the nomogram therefore provide a valuable decision support for more individualised treatment approaches in the future.

Note: The predictive models in this paper are published on the website http://www.predictcancer.org.

Conflict of interests

We are not aware of any actual or potential conflicts of interest.

References

[1] Adjuvant radiotherapy for rectal cancer: a systematic overview of 8507 patients from 22 randomised trials. Lancet 2001; 358: 1291–1304.
[2] Valentini V, Beets-Tan R, Borras JM, et al. Evidence and research in rectal cancer. Radiother Oncol 2008;87:449–74.
[3] Valentini V, Aristei C, Glimelius B, et al. Multidisciplinary rectal cancer management: 2nd European rectal cancer consensus conference (EURECA-CC2). Radiother Oncol 2009;92:148–63.
[4] Capirci C, Rampin L, Erba PA, et al. Sequential FDG-PET/CT reliably predicts response of locally advanced rectal cancer to neo-adjuvant chemo-radiation therapy. Eur J Nucl Med Mol Imaging 2007;34:1583–93.
[5] Capirci C, Rubello D, Chierichetti F, et al. Long-term prognostic value of 18F-FDG PET in patients with locally advanced rectal cancer previously treated with neoadjuvant radiochemotherapy. AJR Am J Roentgenol 2006;187:W202–8.
[6] Valentini V, Coco C, Cellini N, et al. Ten years of preoperative chemoradiation for extraperitoneal T3 rectal cancer: acute toxicity, tumor response, and sphincter preservation in three consecutive studies. Int J Radiat Oncol Biol Phys 2001;51:371–83.
[7] Vliegen RF, Beets-Tan RG, Vanhauten B, et al. Can an FDG-PET/CT predict tumor clearance of the mesorectal fascia after preoperative chemoradiation of locally advanced rectal cancer? Strahlenther Onkol 2008;184:457–64.
[8] O'Neil BH, Tepper JE. Current options for the management of rectal cancer. Curr Treat Options Oncol 2007;8:331–8.
[9] Pucciarelli S, Toppan P, Friso ML, et al. Complete pathologic response following preoperative chemoradiation therapy for middle to lower rectal cancer is not a prognostic factor for a better outcome. Dis Colon Rectum 2004;47:1798–807.
[10] Vecchio FM, Valentini V, Minsky BD, et al. The relationship of pathologic tumor regression grade (TRG) and outcomes after preoperative therapy in rectal cancer. Int J Radiat Oncol Biol Phys 2005;62:752–60.
[11] Habr-Gama A, Perez RO, Nadalin W, et al. Long-term results of preoperative chemoradiation for distal rectal cancer correlation between final stage and survival. J Gastrointest Surg 2005;9:90–9. discussion 99–101.
[12] Capirci C, Valentini V, Cionini L, et al. Prognostic value of pathologic complete response after neoadjuvant therapy in locally advanced rectal cancer: long-term analysis of 566 ypCR patients. Int J Radiat Oncol Biol Phys 2008;72:99–107.
[13] Rodel C, Martus P, Papadoupolos T, et al. Prognostic significance of tumor regression after preoperative chemoradiotherapy for rectal cancer. J Clin Oncol 2005;23:8688–96.
[14] Borschitz T, Wachtlin D, Mohler M, Schmidberger H, Junginger T. Neoadjuvant chemoradiation and local excision for T2–3 rectal cancer. Ann Surg Oncol 2008;15:712–20.
[15] Lezoche G, Baldarelli M, Guerrieri M, et al. A prospective randomized study with a 5-year minimum follow-up evaluation of transanal endoscopic microsurgery versus laparoscopic total mesorectal excision after neoadjuvant therapy. Surg Endosc 2008;22:352–8.
[16] Habr-Gama A, Perez RO, Nadalin W, et al. Operative versus nonoperative treatment for stage 0 distal rectal cancer following chemoradiation therapy: long-term results. Ann Surg 2004;240:711–7. discussion 717–718.
[17] Berger C, de Muret A, Garaud P, et al. Preoperative radiotherapy (RT) for rectal cancer: predictive factors of tumor downstaging and residual tumor cell density (RTCD): prognostic implications. Int J Radiat Oncol Biol Phys 1997;37:619–27.
[18] Das P, Skibber JM, Rodriguez-Bigas MA, et al. Predictors of tumor response and downstaging in patients who receive preoperative chemoradiation for rectal cancer. Cancer 2007;109:1750–5.
[19] Amthauer H, Denecke T, Rau B, et al. Response prediction by FDG-PET after neoadjuvant radiochemotherapy and combined regional hyperthermia of rectal cancer: correlation with endorectal ultrasound and histopathology. Eur J Nucl Med Mol Imaging 2004;31:811–9.

[20] Capirci C, Rubello D, Pasini F, et al. The role of dual-time combined 18-fluorideoxyglucose positron emission tomography and computed tomography in the staging and restaging workup of locally advanced rectal cancer, treated with preoperative chemoradiation therapy and radical surgery. Int J Radiat Oncol Biol Phys 2009.
[21] Denecke T, Rau B, Hoffmann KT, et al. Comparison of CT, MRI and FDG-PET in response prediction of patients with locally advanced rectal cancer after multimodal preoperative therapy: is there a benefit in using functional imaging? Eur Radiol 2005;15:1658–66.
[22] Guillem JG, Moore HG, Akhurst T, et al. Sequential preoperative fluorodeoxyglucose-positron emission tomography assessment of response to preoperative chemoradiation: a means for determining longterm outcomes of rectal cancer. J Am Coll Surg 2004;199:1–7.
[23] Kalff V, Duong C, Drummond EG, Matthews JP, Hicks RJ. Findings on 18F-FDG PET scans after neoadjuvant chemoradiation provides prognostic stratification in patients with locally advanced rectal carcinoma subsequently treated by radical surgery. J Nucl Med 2006;47:14–22.
[24] Melton GB, Lavely WC, Jacene HA, et al. Efficacy of preoperative combined 18-fluorodeoxyglucose positron emission tomography and computed tomography for assessing primary rectal cancer response to neoadjuvant therapy. J Gastrointest Surg 2007;11:961–9. discussion 969.
[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] 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.
[27] Konski A, Li T, Sigurdson E, et al. Use of molecular imaging to predict clinical outcome in patients with rectal cancer after preoperative chemotherapy and radiation. Int J Radiat Oncol Biol Phys 2009;74:55–9.
[28] Eastham JA, Kattan MW, Scardino PT. Nomograms as predictive models. Semin Urol Oncol 2002;20:108–15.
[29] 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.
[30] 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.
[31] Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods 2002;7:147–77.
[32] Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Inform 2007;2:59–77.
[33] Fung GM, Mangasarian OL. Proximal support vector machine classifiers. International Conference on Knowledge Discovery and Data Mining, San Francisco (California, USA): ACM. 2001; pp 77–86.
[34] Pepe MS. Receiver operating characteristic methodology. J Am Stat Assoc 2000:95.
[35] Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat Med 2000;19:1141–64.
[36] Niang, N, Saporta, G. Resampling ROC curves. Statistics for Data Mining, Learning and Knowledge Extraction, Aveiro (Portugal). 2007.
[37] Shariat SF, Capitanio U, Jeldres C, Karakiewicz PI. Can nomograms be superior to other prediction tools? BJU Int 2009;103:492–5. discussion 495–497.
[38] Iasonos A, Schrag D, Raj GV, Panageas KS. How to build and interpret a nomogram for cancer prognosis. J Clin Oncol 2008;26:1364–70.
[39] Jakulin A, Možina M, Demšar J, Bratko I, Zupan B. Nomograms for visualizing support vector machines. International Conference on Knowledge Discovery and Data Mining, Chicago (Illinois, USA): ACM. 2005; pp 108–117.
[40] Dodd LE, Pepe MS, Partial AUC. Estimation and regression. Biometrics 2003;59:614–23.
[41] Starmans MH, Krishnapuram B, Steck H, et al. Robust prognostic value of a knowledge-based proliferation signature across large patient microarray studies spanning different cancer types. Br J Cancer 2008;99:1884–90.
[42] 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.