European Journal of Nuclear Medicine and Molecular Imaging (2019) 46:864–877
https://doi.org/10.1007/s00259-018-4231-9

ORIGINAL ARTICLE

External validation of a combined PET and MRI radiomics model for prediction of recurrence in cervical cancer patients treated with chemoradiotherapy

François Lucia[1,2] · Dimitris Visvikis[2] · Martin Vallières[2] · Marie-Charlotte Desseroit[2] · Omar Miranda[1] · Philippe Robin[3] · Pietro Andrea Bonaffini[4] · Joanne Alfieri[5] · Ingrid Masson[6] · Augustin Mervoyer[6] · Caroline Reinhold[4] · Olivier Pradier[1,2] · Mathieu Hatt[2] · Ulrike Schick[1,2]

Received: 29 August 2018 / Accepted: 27 November 2018 / Published online: 7 December 2018

Abstract

Purpose The aim of this study was to validate previously developed radiomics models relying on just two radiomics features from ^{18}F-fluorodeoxyglucose positron emission tomography (PET) and magnetic resonance imaging (MRI) images for prediction of disease free survival (DFS) and locoregional control (LRC) in locally advanced cervical cancer (LACC).

Methods Patients with LACC receiving chemoradiotherapy were enrolled in two French and one Canadian center. Pre-treatment imaging was performed for each patient. Multicentric harmonization of the two radiomics features was performed with the ComBat method. The models for DFS (using the feature from apparent diffusion coefficient (ADC) MRI) and LRC (adding one PET feature to the DFS model) were tuned using one of the French cohorts ($n = 112$) and applied to the other French ($n = 50$) and the Canadian ($n = 28$) external validation cohorts.

Results The DFS model reached an accuracy of 90% (95% CI [79–98%]) (sensitivity 92–93%, specificity 87–89%) in both the French and the Canadian cohorts. The LRC model reached an accuracy of 98% (95% CI [90–99%]) (sensitivity 86%, specificity 100%) in the French cohort and 96% (95% CI [80–99%]) (sensitivity 83%, specificity 100%) in the Canadian cohort. Accuracy was significantly lower without ComBat harmonization (82–85% and 71–86% for DFS and LRC, respectively). The best prediction using standard clinical variables was 56–60% only.

Conclusions The previously developed PET/MRI radiomics predictive models were successfully validated in two independent external cohorts. A proposed flowchart for improved management of patients based on these models should now be confirmed in future larger prospective studies.

Keywords Radiomics · Prediction · Chemoradiotherapy · Cervical cancer · External validation

Mathieu Hatt and Ulrike Schick contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00259-018-4231-9) contains supplementary material, which is available to authorized users.

✉ François Lucia
francois.lucia@chu-brest.fr

[1] Radiation Oncology Department, University Hospital, Brest, France

[2] LaTIM, INSERM, UMR 1101, University Brest, Brest, France

[3] Nuclear Medicine Department, University Hospital, Brest, France

[4] Department of Radiology, McGill University Health Centre (MUHC), Montreal, Canada

[5] Department of Radiation Oncology, McGill University Health Centre (MUHC), Montreal, Canada

[6] Department of Radiation Oncology, Institut de Cancérologie de l'Ouest, Nantes, France

Introduction

Cervical cancer (CC) was the fourth most common cancer in women, and the seventh overall, with an estimated 528,000 new cases in 2012. There were an estimated 266,000 deaths worldwide, accounting for 7.5% of all female cancer deaths [1]. Most patients present with locally advanced (LACC; i.e., stage IB1 to IVA) disease at diagnosis. Chemoradiotherapy (CRT) followed by brachytherapy (BT) remains the standard of care in these patients, with expected cure rates of 30–90% depending on prognostic factors including International Federation of Gynecology and Obstetrics (FIGO) stage, histology, and lymph node (LN) metastases. Currently, the treatment is based on these clinical parameters, namely FIGO stage and N staging, although patients with the same clinicopathological characteristics have very different clinical outcomes, highlighting the need of novel tools to better identify patients at high risk of relapse [2].

^{18}F-fluorodeoxyglucose (FDG) positron emission tomography/computed tomography (PET/CT) and magnetic resonance imaging (MRI) are essential in the initial staging, therapeutic strategy [3] and treatment response assessment [4] in CC. In the past few years, there has been a growing interest in systematically extracting more information from medical images beyond conventional metrics (such as anatomical tumor size or standardized uptake values (SUV_{max}), an approach known as radiomics [5]. Radiomics features are engineered statistical or model-based metrics used to quantify tumor characteristics such as intensity, shape, and heterogeneity, some of which can provide clinically relevant and complementary information beyond usual clinical variables in several pathologies [6], including CC [7–11]. In CC, these relevant features have been extracted from pretreatment PET/CT [7, 9, 11], diffusion-weighted MRI (DW-MRI) [10], or dynamic contrast enhancement MRI (DCE-MRI) [8]. In a previous work, we trained and internally validated in a monocentric context two simple radiomics-based models relying on two textural features: one from the ^{18}F-FDG PET ($GLNU_{GLRLM}$) and one from the apparent diffusion coefficient (ADC) map derived from DW-MRI ($Entropy_{GLCM}$). These models were highly accurate to predict disease-free survival (DFS) and locoregional control (LRC) in LACC patients undergoing CRT, with significantly higher prognostic power than conventional factors [12].

It has been highlighted recently that most published radiomics models are never properly validated in an external multicentric setting [13], which is one of the current main limitations of the clinical transfer of radiomics approaches [14]. Our goal was therefore to validate our previously developed radiomics models in two external cohorts.

Materials and methods

Patients

Patients with histologically proven LACC, staged IB1-IVA (FIGO 2009 definition) and treated with definitive curative CRT and subsequent BT from August 2010 to July 2017 (to ensure a minimum follow-up of 1 year) at three institutions were included in this retrospective study (Table 1). Patients with stage IB1 and IIA1 were only considered for inclusion if they had positive LN.

All patients were required to have clinical pelvic examination, PET/CT imaging, and pelvic MRI, both at diagnosis, a surgical staging with para-aortic lymph node dissection if PET/CT was negative in these area and at least 1 year of follow-up. Exclusion criteria were history of previous chemotherapy or RT and/or distant metastatic disease.

Collected data included age and date of diagnosis, histology, FIGO stage, presence of positive LN on PET/CT, tumor size as measured on MRI, body mass index (BMI), complete blood counts (CBC) before treatment, external beam radiotherapy (EBRT) and BT doses, date and status (i.e., alive, deceased, recurrence) at last follow-up. Date and site of recurrence were also collected. Recurrences were considered as local (vaginal and/or cervical), regional (pelvic/para-aortic), or distant (upper abdominal and/or extra-abdominal) (https://www.nccn.org/professionals/physician_gls/pdf/cervical.pdf).

A total of 112 patients were recruited at the University Hospital of Brest (France) and were used to re-train the original models. In our previous publication, only 102 patients were exploited due to insufficient follow-up for some of them: the model was trained in the first 69 patients and internally validated in the next 33 patients [12]. Two other cohorts of patients treated at the University Hospitals of Nantes (France) ($n = 50$) and McGill (Canada) ($n = 28$) were considered as external testing cohorts. All patients provided signed permission for the use of their clinical data for scientific purposes and informed consent for the anonymous publication of data. Institutional Review Boards approved this study (29BCR18.0015).

Imaging and reconstruction protocols

Full details regarding the PET/CT and MRI acquisition and reconstruction settings are provided in the supplemental table 1.

PET/CT

In Brest, the Philips Gemini (Philips Medical Systems, Cleveland, OH) was used for the first six patients and the Siemens Biograph (SIEMENS Healthineers Medical Solutions, Knoxville, TN, USA) for the next 106 patients after

Table 1 Patients' characteristics

	Brest		Nantes		McGill		Difference (*p* value)
	n = 112	%	*n* = 50	%	*n* = 28	%	
Age median (range)	56 (29–90)		51 (23–73)		52 (26–86)		0.48
FIGO stage							
IB1	3	3	1	2	0	0	0.68
IB2	11	10	4	8	4	14	
IIA	7	6	3	6	0	0	
IIB	61	54	27	54	14	50	
IIIA	2	2	1	2	1	4	
IIIB	15	13	6	12	7	25	
IVA	13	12	6	12	2	7	
Histology							
Squamous	91	81	40	80	24	86	0.84
Adenocarcinoma	15	13	7	14	1	3	
Adenosquamous carcinoma	1	1	0	0	0	0	
Clear cell carcinoma	5	5	3	6	3	11	
Lymph node involvement							
Uninvolved	55	49	23	46	7	25	0.59
Involved	57	51	27	54	21	75	
Pelvic	40	70	19	70	16	76	
Pelvic and para-aortic	17	30	8	30	5	24	
CBC median (range)							
White blood cells	8.1 10^3/ml (4.6–25.6)		8.4 10^3/ml (4.5–26.1)		X		0.56
Hemoglobin	127 g/dl (71–151)		123 g/dl (67–147)		X		0.44
Platelets	257.3 10^3/ml (171–819)		273.9 10^3/ml (154–767)		X		0.34
Body-mass index median (range)	23.1 (14–42)		22.7 (13–39)		X		0.53
Treatment							
3D-CRT	59	53	2	4	4	14	< 0.0001
IMRT	53	47	48	96	24	86	
EBRT dose median (range)	45 (45–54)		45 (45–54)		45 (45–54)		1.00
BT dose median (range)	24 (21–28)		15 (15–15)		24 (24–24)		1.00
Overall treatment time (range)	49 (47–53)		48 (46–52)		50 (48–54)		0.46

FIGO International Federation of Gynecology and Obstetrics, *CBC* complete blood counts, *3D-CRT* three-dimensional conformal radiotherapy, *IMRT* intensity-modulated photon radiotherapy, *EBRT* external beam radiotherapy, *BT* brachytherapy

a scanner replacement. Patients fasted for 4 h before acquisition, and the blood glucose level had to be less than 7 mmol/l before injection of 5 MBq/kg of ^{18}F-FDG. PET acquisitions were carried out approximately 60 min after injection. The CT consisted of a 64-slice multidetector-row spiral scanner with a transverse field of view of 700 mm. Standard CT parameters were used: a collimation of 16 × 1.2 mm^2, pitch 1, tube voltage of 120 kV, and effective tube current of 80 mA. Routine clinical image reconstruction protocols were used: for the Philips GEMINI, data were reconstructed using the RAMLA 3D (two iterations, relaxation parameter 0.05) whereas for the Siemens Biograph, images were reconstructed with Fourier rebinning (FORE) followed by OSEM (two iterations, eight subsets). In both cases, images were corrected for attenuation using the corresponding low-dose CT, reconstructed with a 2 × 2 × 2 mm^3 voxels grid and post-filtered with a 5-mm FWHM 3D Gaussian.

In Nantes, the Siemens Biograph (SIEMENS Healthineers Medical Solutions, Knoxville, TN, USA) was used for all patients with the same parameters as in Brest.

In McGill, the Discovery ST (GE Healthcare, Waukesha, WI, USA) was used for all patients with some acquisition and reconstruction parameters differing from those of Brest and Nantes. Notably the OSEM algorithm (two iterations, eight subsets) was used to reconstruct PET images on a 3.65 × 3.65 × 3.27 mm^3 voxels grid.

MRI

In Brest, all MRI studies were performed on a 1.5-T unit (Siemens Medical Solutions, Magnetom Aera, Erlangen, Germany or General Electric, Milwaukee, WI, USA) using a phased-array body coil, 2 weeks before the start of CRT with set image protocols. MRI was performed at least 10 days after cone biopsy to avoid false-positive findings due to post-biopsy inflammation. No patient had an absolute contraindication to the MRI examination. The MRI protocol included high-resolution turbo T2-weighted sequences in the sagittal, axial, and axial oblique (perpendicular to the long axis of the cervix) planes. T1-weighted and T2-weighted axial images were obtained through the pelvis and up to the level of the renal hilum to assess nodal status. The MRI protocol also included axial TSE T2-weighted fat-suppressed and axial oblique and sagittal diffusion-weighted images (DWI) (b values of 0, 400, and 1000) without slices gap. All except two allergic patients (training set) received a 0.1 mmol/kg injection of gadobenate dimeglumine (Multihance; Bracco Diagnostics, Milan, Italy). After 3 min, a T1-weighted fat-suppressed sequence (CE-MRI) in the axial and sagittal plane was acquired.

In Nantes, all MRI studies were performed with a 1.5-T unit (Philips Medical Systems, Best, Netherlands or Optima MR450w GE Healthcare, Little Chalfont, UK) in the same conditions but with some differences in protocols of sequences, as described in the supplemental table 1.

In McGill, all MRI studies were performed with a 1.5-T unit (Signa Excite; GE Healthcare, Waukesha, WI, USA, or Magnetom Avanto, Siemens, Erlangen, Germany) in the same conditions but with some differences in the sequences protocols (supplemental table 1).

Treatment

Consortium guidelines were applied to outline the clinical target volume (CTV), the planning target volume (PTV), and organs-at-risk [15]. Treatment consisted of three-dimensional conformal radiotherapy (3D-CRT) ($n = 59$, $n = 2$, and $n = 4$ in Brest, Nantes, and McGill, respectively) or intensity-modulated radiotherapy (IMRT) ($n = 53$, $n = 48$, and $n = 24$ in Brest, Nantes, and McGill, respectively) delivered using a linear accelerator.

All patients received pelvic EBRT or extended-field RT to the para-aortic area depending on their staging work-up, using high energy photons (18 MV), at a dose of 45–50.4 Gy with standard fractionation. In patients with positive pelvic or para-aortic LN, an image-guided targeted boost was delivered to a dose of 50.4–54 Gy to the involved nodes (17, eight, and five patients in Brest, Nantes, and McGill, respectively). The week following EBRT, patients received 3–4 fractions of MRI-guided high-dose-rate (HDR) intracavitary BT every 4 days in Brest and McGill, and 30 to 50 pulsed dose rate BT (PDR) in Nantes. The prescribed HDR and PDR doses to the high-risk CTV were 7 Gy and 15 Gy (30 to 50 cGy/h), respectively. No patient experienced RT breaks secondary to acute toxicity (median RT duration, 49 days; range, 47–53 days). All patients received 4–6 cycles of concomitant chemotherapy with weekly cisplatin (40 mg/m^2) or carboplatin (AUC 2) in case of renal contraindication.

Follow-up

PET/CT was performed with the same scanner as for the diagnosis in each of the three centers, 2 to 3 months (2.6 ± 0.5) after treatment completion in order to assess therapeutic response using the PERCIST criteria: patients were classified as having complete metabolic response (CMR), partial metabolic response (PMR), stable metabolic disease (SMD), or progressive metabolic disease (PMD) [16]. Clinical follow-up consisted of physical examination every third month until 2 years after diagnosis, every sixth month up to 5 years, annually thereafter, and was done alternatively by the radiation oncologist and the gynecologist. Follow-up imaging studies consisted of MRI 3 months after treatment completion and annually until 2 years after treatment, CT every 6 months until 2 years after treatment completion and if clinically indicated after, and/or PET/CT if clinically indicated.

Original model development

We summarize in this section how the models we aim to validate were originally developed. More details are provided in our previous publication [12].

Based on the segmented tumors in each of the images (see section *Segmentation* below), 92 radiomics features (19 shape, 11 first-order, and 62 higher-order metrics) were extracted, blinded to the outcomes. Textures were computed in a single matrix considering all 13 orientations simultaneously (merging strategy) [17]. For these, three different grey-level discretization methods [18] were considered: fixed number (64) of bins (Q_L), histogram equalization (64 bins also) (Q_E) or fixed-width bins of 0.5 SUV for PET, 10 units for T2- and CE- MRI and 10 mm^2/s for ADC map (Q_F). All features were implemented according to the current version of image biomarkers standardization initiative (IBSI) guidelines and values were checked against the benchmark [17]. A total of 864 image features (216 per modality), eight clinical and histopathological parameters (age, FIGO stage, N stage, BMI, histology, white blood cells, platelets, hemoglobin) and three treatment parameters (radiotherapy dose, brachytherapy dose, and overall treatment time) were included in the statistical analysis. In the original training set ($n = 69$), all parameters including usual confounding factors (e.g., volume, clinical variables, etc.) were tested using univariate Cox proportional hazards

modeling, for which statistical significance was corrected for multiple testing with the Bonferroni method to reduce the false-positive discovery rate [19]. Corrected p values below α/K (K = 875 and $\alpha = 0.05$, i.e., $p < 0.000057$) were therefore considered statistically significant. Cox-regression models with the stepwise method were used for multivariate analysis including only the uncorrelated parameters (Spearman rank correlation below 0.5) found significant in the univariate analysis. In addition, correlations between the parameters identified in the multivariate analysis and standard metrics (e.g., volume, FIGO stage, etc.) were checked to avoid building models that would end up being surrogates of usual variables. The receiver operating characteristic (ROC) curve was used to determine the best cut-off values of significant parameters according to the Youden index. Our study followed the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines [20]. The resulting models were based on either a single or two textural features: $\text{Entropy}_{\text{GLCM}}$ from the ADC map of the DWI MRI using a cut-off of 12.64 to predict DFS, and the same parameter (and same cut-off value), combined with the $\text{GLNU}_{\text{GLRM}}$ of the FDG PET using a cut-off of 103.71 for prediction of LRC. Entropy from the grey-level co-occurrence matrix (GLCM) calculated in the ADC map is considered a measurement of heterogeneity at a local scale (variations of intensity between each voxel and its immediate surrounding in 3D). The parameter measured in the FDG PET image on the other hand is grey-level non uniformity (GLNU), calculated in a different type of texture matrix (GLRLM, grey level run length matrix) and is considered a measurement of heterogeneity at a larger scale (i.e., groups of voxels). $\text{Entropy}_{\text{GLCM}}$ has been shown to be quite robust to variations in most image properties [21], whereas $\text{GLNU}_{\text{GLRLM}}$ has been described as more sensitive, especially to voxel size [22]. Although $\text{Entropy}_{\text{GLCM}}$ calculated in the ADC map from MRI was sufficient alone to predict DFS with high accuracy, $\text{GLNU}_{\text{GLRLM}}$ calculated in the FDG PET provided the complementary information required to predict LRC with a similar level of accuracy. Using another threshold of ADC $\text{Entropy}_{\text{GLCM}}$ (equal to 12.7) to predict LRC led to an AUC of 0.84 only, compared to 0.94 when combining both features.

Original models re-training and external validation

The following sections describe the analysis performed in the present study.

Segmentation and extraction of radiomics features

All images in the two testing cohorts (Nantes and McGill) were processed following the same protocol as the training cohort (Brest). The ten additional patients of the Brest training cohort were also processed in the same way.

Only primary tumors, not pathological lymph nodes, were analyzed. The PET and MRI images were processed independently by a single expert radiation oncologist (F. Lucia). To reduce user-dependency of this step, robust (semi)automated methods were exploited. The metabolically active volumes on PET images were automatically delineated with the fuzzy locally adaptive Bayesian (FLAB) algorithm [23, 24]. The anatomic volumes were also delineated on (i) the ADC map derived from DWI-MRI, (ii) CE-MRI, and (iii) T2. Each sequence was segmented independently because of anatomical changes between each sequence acquisition, using a previously validated semi-automatic approach exploiting 3D Slicer™ and the Growcut algorithm [25]. This approach only requires painting strokes on the apparent foreground and background as input. Visual examples are provided in [12]. For the purpose of the present analysis, although all radiomics features were extracted from all images in an automated pipeline in exactly the same way as in our previous publication [14], only the two features required for the models were used.

Harmonization method

To pool radiomics features extracted from the three centers relying on different PET and/or MRI protocols, we used an a posteriori harmonization statistical method named ComBat, initially proposed for genomic studies to correct the so-called batch effect, and previously applied to image features from MRI [26] and PET [27]. The ComBat model [28] assumes that the value of each feature y measured in VOI j and scanner i can be written as:

$$y_{ij} = \alpha + X_{ij}\beta + \gamma_i + \delta_i + \varepsilon_{ij}$$

where α is the average value for feature y, X is a design matrix for the covariates of interest, β is the vector of regression coefficients corresponding to each covariate, γ_i is the additive effect of scanner i on features supposed to follow a normal distribution, δ_i describes the multiplicative scanner effect supposed to follow an inverse gamma distribution, and ε_{ij} is an error term (normally distributed with a zero mean) [29]. ComBat harmonization consists in estimating γ_i and δ_i using empirical Bayes estimates (noted γ_i^* and δ_i^*) [28]. The normalized value of feature y for VOI j and scanner i is then obtained as

$$y_{ij}^{ComBat} = \frac{y_{ij} - \hat{\alpha} - X_{ij}\hat{\beta} - \gamma_i^*}{\delta_i^*} + \hat{\alpha} + X_{ij} + \hat{\beta}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are estimators of parameters α and β, respectively. The harmonization determines a transformation for each feature separately based on the batch (here Brest, Nantes, and McGill) effect observed on feature values. We used ComBat without accounting for any biological covariate

(i.e., X = 0) because there was no difference between cohorts in terms of clinical or histopathological parameters. We used the "combat" R function provided at https://github.com/Jfortin1/ComBatHarmonization.

Statistical analysis

The updated cut-off values, before or after harmonization with ComBat, for both ADC MRI $Entropy_{GLCM}$ and FDG PET $GLNU_{GLRLM}$ were recomputed on the entire training cohort ($N = 112$) using ROC curves and the Youden index.

The models for DFS (using only ADC MRI $Entropy_{GLCM}$) and LRC (using both ADC MRI $Entropy_{GLCM}$ and FDG PET $GLNU_{GLRLM}$) prediction were then applied to the two testing cohorts using the radiomics features values before or after harmonization with ComBat. The two testing cohorts were also pooled as a single cohort to further assess prediction performance of the models. Accuracy, sensitivity, and specificity values for identifying patients with recurrence and lack of locoregional control were calculated. The corresponding Kaplan–Meier curves were generated and distributions of survival times were compared using the log-rank test. Adjusted hazard ratios (HRs) and the corresponding 95% confidence intervals (CI) were calculated. All statistical analyses were performed using MedCalc Statistical Software version 15.8 (MedCalc Software bvba, Ostend, Belgium; https://www.medcalc.org; 2015).

Results

Patient and tumor characteristics

The training and the two testing cohorts had similar clinical, treatment, and histopathological characteristics, except for EBRT modalities (more IMRT in the testing cohorts, $p < 0.0001$), which however have no impact on the effectiveness of treatment but only on its toxicity [30].

Outcome

Training cohort (Brest, with updated information with respect to [12] taking into account the ten additional patients)

After a median follow-up of 24.3 months (range, 6–83 months), 20 patients (18%) had died. Progression or disease recurrence occurred in 39 patients (35%): 20 patients (18%) had an isolated pelvic recurrence and 19 (17%) a distant recurrence (12 with isolated distant recurrence and seven with both regional and distant recurrence). Post-CRT PET/CT identified 70 CMR, 36 PMR, one SMD, and five PMD. Out of 20 patients without LRC, nine had a CMR and 11 a PMR. Out of 19 patients with distant metastases at staging, nine had a CMR, five a PMR, and five a PMD. Overall 2-year DFS and LRC were 68% and 82%, respectively.

Testing cohorts (Nantes and McGill)

Nantes After a median follow-up of 33 months (range, 14–50 months), five patients (10%) had died. Progression or disease recurrence occurred in 15 patients (30%): seven patients (14%) had an isolated pelvic recurrence and eight (16%) a distant recurrence (five with isolated distant recurrence and three with both regional and distant recurrence). Post-CRT PET/CT identified 28 CMR, 18 PMR, one SMD, and three PMD. Out of seven patients without LRC, four had a CMR and three a PMR. Out of eight patients with distant metastases at staging, four had a CMR, three a PMR, and one a PMD. Overall 2-year DFS and LRC were 73%, and 87%, respectively.

McGill After a median follow-up of 26 months (range, 7–86 months), four patients (14%) had died. Progression or disease recurrence occurred in 13 patients (46%): six patients (21%) had an isolated pelvic recurrence and seven (25%) a distant recurrence (four with isolated distant recurrence and three with both regional and distant recurrence). Post-CRT PET/CT identified 18 CMR, eight PMR, 0 SMD, and two PMD. Out of six patients without LRC, four had a CMR and two a PMR. Out of seven patients with distant metastases at staging, four had a CMR, two a PMR, and one a PMD. Overall 2-year DFS and LRC were 65%, and 79%, respectively.

Radiomics features cut-off values: updates in the training cohort

DFS

An AUC of 0.94 was obtained for ADC $Entropy_{GLCM}$ (Fig. 1). Optimal cut-off values were 12.64 and 14.27 before and after ComBat harmonization. The estimated 3-year DFS rates between patients with low versus high ADC $Entropy_{GLCM}$ were 92% and 9%, respectively, with a HR of 26.6 (95% CI [12.8–55.0]) (supplemental figure 1).

LRC

AUCs of 0.83, 0.86 and 0.94 were obtained for ADC $Entropy_{GLCM}$, PET $GLNU_{GLRLM}$ and their combination respectively (Fig. 2). Optimal cut-off values were 103.71 and 92.27 for PET $GLNU_{GLRLM}$ before and after ComBat harmonization. The estimated 3-year LRC rates between patients with low versus high PET $GLNU_{GLRLM}$-Q_E were

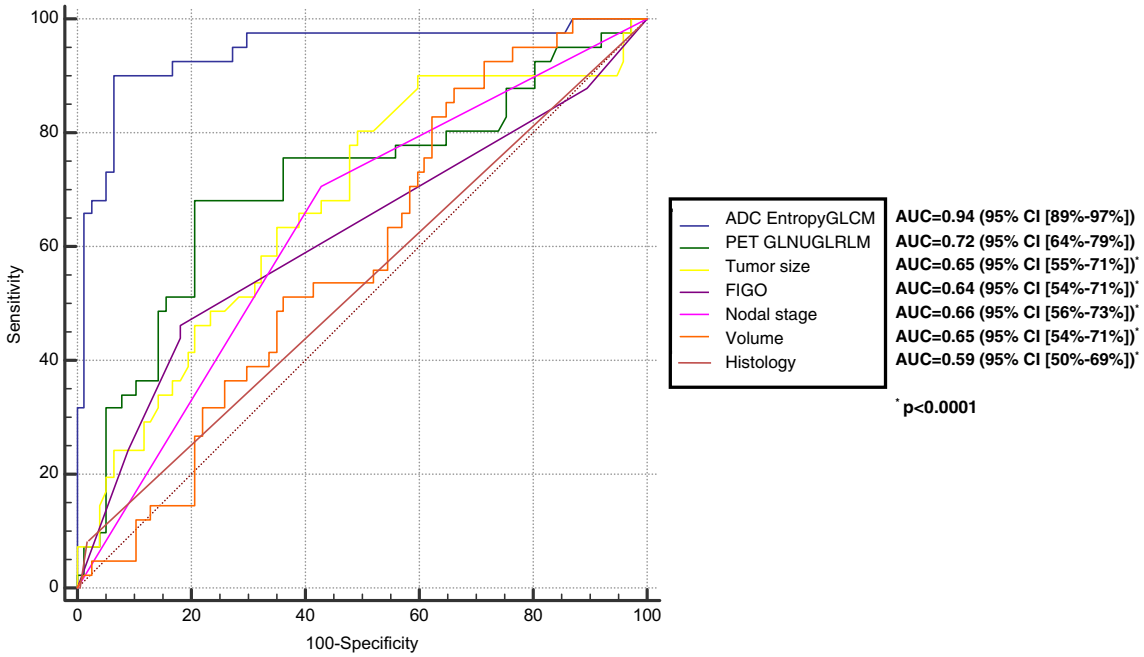

Fig. 1 DFS prediction in training cohort ($N = 112$). **a** Comparison of ROC curves for MRI ADC Entropy$_{GLCM}$ and PET GLNU$_{GLRLM}$ in comparison with clinico-pathological features

97% and 37% (HR = 28.9 (95% CI [10.3–81.0])). Using ADC Entropy$_{GLCM}$ for which the same cut-off values as for DFS were determined, 3-year LRC rates were 97% and 44% (HR = 22.1 (95% CI [8.4–58.1])). Finally, the combination of these 2 features (with the same cut-off values) provided an even better predictive model for LRC (97% vs. 7%, HR of 66.9 (95% CI [17.3–159.8])) (supplemental figure 2).

Radiomics models evaluation in the testing cohorts

Nantes

DFS Before ComBat harmonization, ADC Entropy$_{GLCM}$ (cut-off 12.64) reached an accuracy of 82% (sensitivity 93%, specificity 77%) to predict recurrence with a HR of 24.0 (95% CI [8.4–68.6], $p < 0.0001$). After Combat harmonization (cut-off

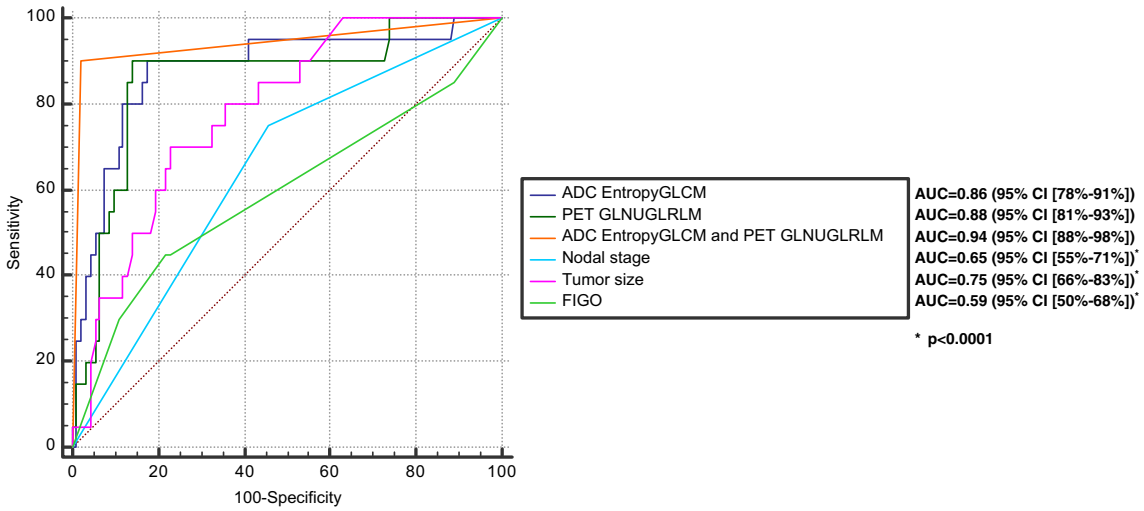

Fig. 2 LRC prediction in training cohort ($N = 112$). **a** Comparison of ROC curves for MRI ADC Entropy$_{GLCM}$, PET GLNU$_{GLRLM}$ and their combination, in comparison with clinico-pathological features

14.27), the accuracy improved to 90%, with the same sensitivity (93%) but with an improved specificity of 89%. This led to an HR of 39.8 (95% CI [12.6–126.1], $p < 0.0001$) (supplemental figure 3). By comparison, the best accuracy obtained using standard clinical factors was 61% (Fig. 3).

LRC Before ComBat harmonization, the model combining PET $GLNU_{GLRLM}$ (cut-off 103.71) and ADC $Entropy_{GLCM}$ (cut-off 12.64) reached an accuracy of 86% (sensitivity 86%, specificity 86%) to predict LRC with a HR of 29.2 (95% CI [4.1–109.5], $p < 0.0001$). After ComBat harmonization (cut-off 92.27 and 14.27), the accuracy improved to 98% with the same sensitivity of 86% but with an improved specificity of 100%. This led to an improved HR of 71.8 (95% CI [4.5–357.7], $p < 0.0001$) (supplemental figures 5 and 6). By comparison, the accuracy obtained using standard clinical factors was 57% at best (Fig. 4).

McGill

DFS Before ComBat harmonization, ADC $Entropy_{GLCM}$ reached an accuracy of 85% (sensitivity 100%, specificity 73%) to predict recurrence with an undefined HR ($p = 0.008$). After ComBat harmonization, the accuracy improved to 90% at the cost of a slightly lower sensitivity (92%) but with a higher specificity of 87%. This led to an HR of 16.3 (95% CI [5.4–48.8], $p < 0.0001$) (supplemental figure 3). The best accuracy obtained using standard clinical factors was 59% only (Fig. 3).

LRC Before ComBat harmonization, the model combining PET $GLNU_{GLRLM}$ and ADC $Entropy_{GLCM}$ reached a limited accuracy of 71% (sensitivity 83%, specificity 68%) to predict LRC with an HR of 10.2 (95% CI [1.9–56.2], $p = 0.007$). After ComBat harmonization, however, the model accuracy was improved to 96% with the same sensitivity of 83% but an improved specificity of 100%, which led to an improved HR of 27.2 (95% CI [3.0–146.0], $p < 0.0001$) (supplemental figures 5 and 6). By comparison, the best accuracy obtained using standard clinical factors was 58% (Fig. 4).

Nantes and McGill pooled

DFS Before ComBat harmonization, ADC $Entropy_{GLCM}$ reached an accuracy of 76% (sensitivity 100%, specificity

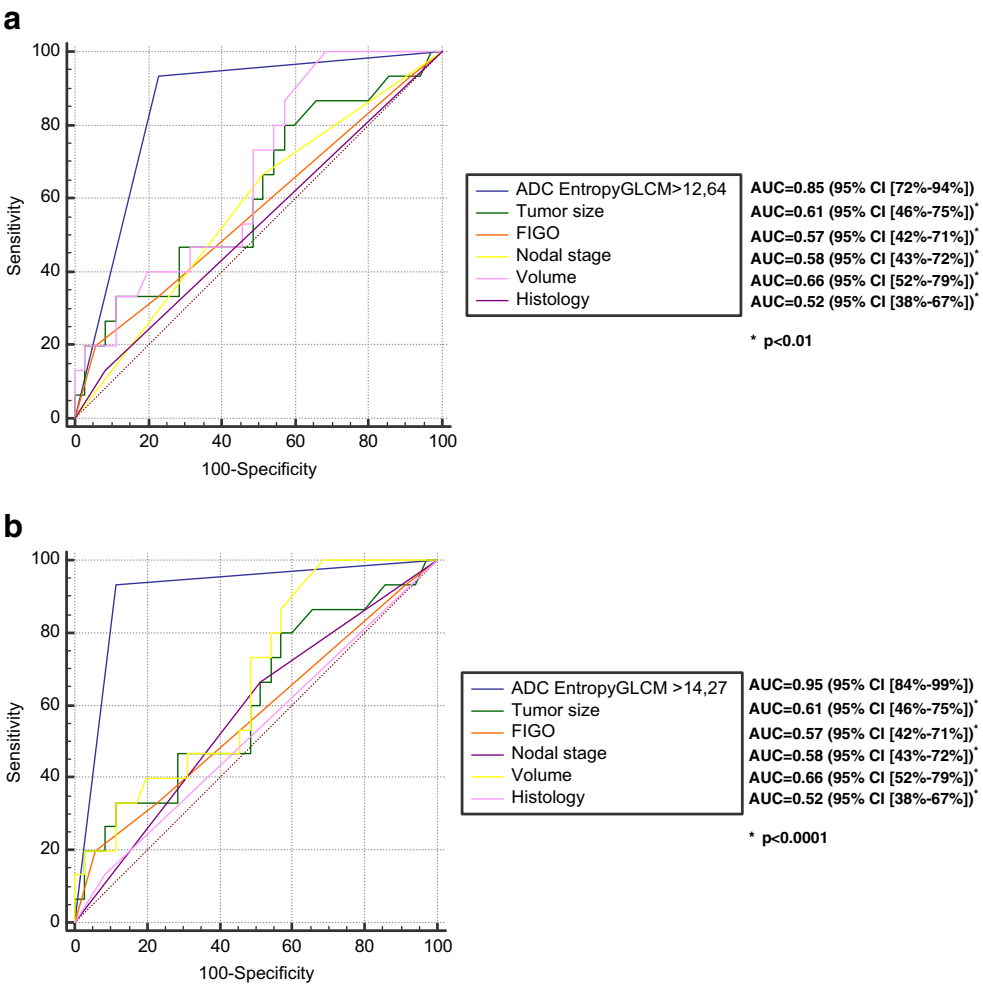

Fig. 3 DFS prediction in testing cohorts from Nantes ($N = 50$) and McGill ($N = 28$). Comparison of ROC curves for (**a**, **b**) Nantes and (**c**, **d**) McGill according to MRI ADC $Entropy_{GLCM}$ in comparison with clinico-pathological features before (**a**–**c**) and after (**b**–**d**) ComBat harmonization

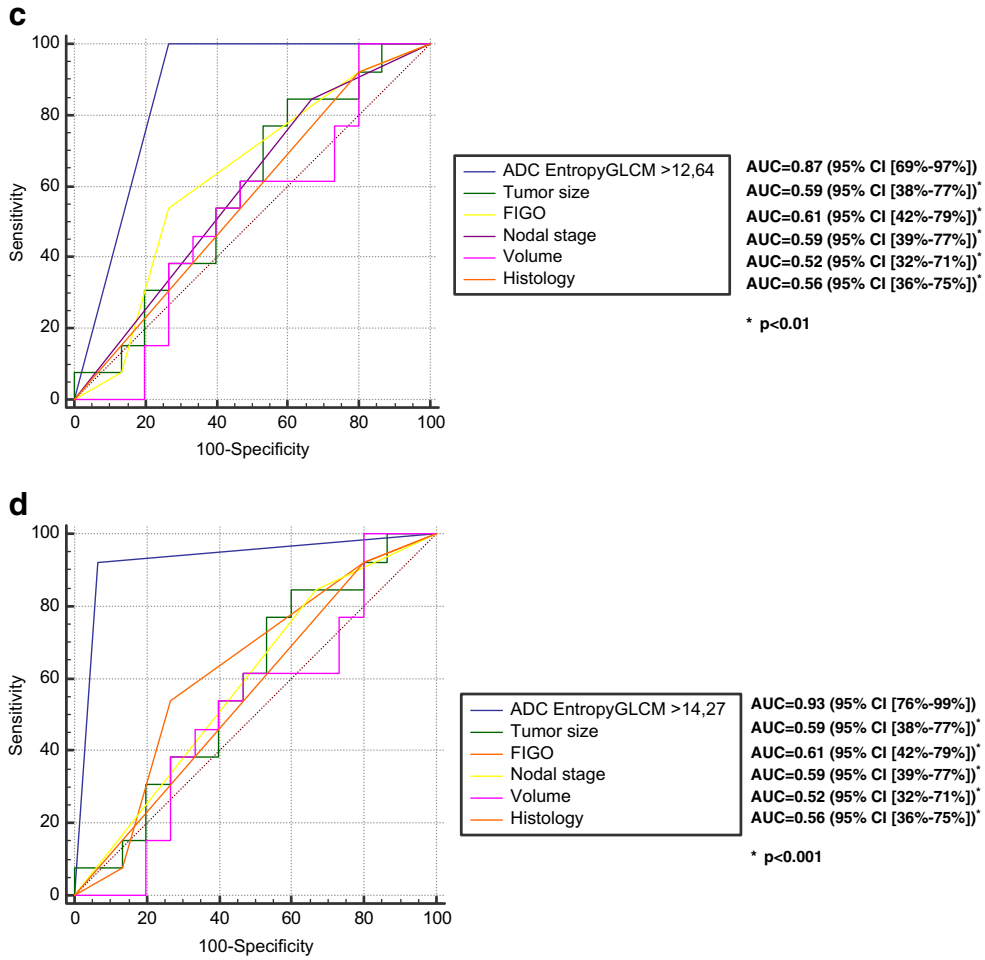

Fig. 3 (continued)

62%) to predict recurrence with an HR undefined ($p = 0.002$). After ComBat harmonization, the accuracy improved to 90% at the cost of a slightly reduced sensitivity (93%) but with improved specificity (88%). This led to an HR of 29.0 (95% CI [13.0–64.5], $p < 0.0001$). Again, by comparison, the best accuracy reached by standard clinical factors was 60% only (supplemental figure 7).

LRC Before ComBat harmonization, the model combining PET $GLNU_{GLRLM}$ and ADC $Entropy_{GLCM}$ reached an accuracy of 81% (sensitivity 85%, specificity 80%) with an HR of 19.6 (95% CI [5.3–72.7], $p = 0.0004$). After ComBat harmonization, the accuracy increased to 97% with the same sensitivity of 85% but a specificity increased to 100%. This led to a HR of 50.9 (95% CI [8.1–218.3], $p < 0.0001$). By comparison, the best accuracy obtained using standard clinical factors was only 58% (supplemental material, figure 8).

Figure 5 shows how the previously proposed flowchart for personalized treatment management of LACC patients based on these two image features [12] could be implemented.

Discussion

Our previous study suggested that two textural features, namely $GLNU_{GLRLM}$ extracted from FDG PET and $Entropy_{GLCM}$ from ADC maps calculated relying on DWI MRI, are powerful predictors of the efficacy of CRT before treatment of LACC with higher accuracy than standard post-treatment metabolic response assessment [12]. Higher values of these radiomics features were indeed associated with worse outcome, confirming that more heterogeneous tumors have a poorer prognosis. From these findings, we derived a flowchart that could be relied upon to tailor treatment (Fig. 5). Accordingly, more aggressive loco-regional treatment could be offered to patients with a high risk of an isolated loco-regional relapse, whereas for patients with high risk of distant relapse, a systemic adjuvant treatment would be more beneficial. It can be observed that some patients with high ADC $entropy_{GLCM}$ and low PET $GLNU_{GLRLM}$ would receive unnecessary treatment (Fig. 5). An alternative could be to recommend more intensive surveillance for the patients classified in this subgroup.

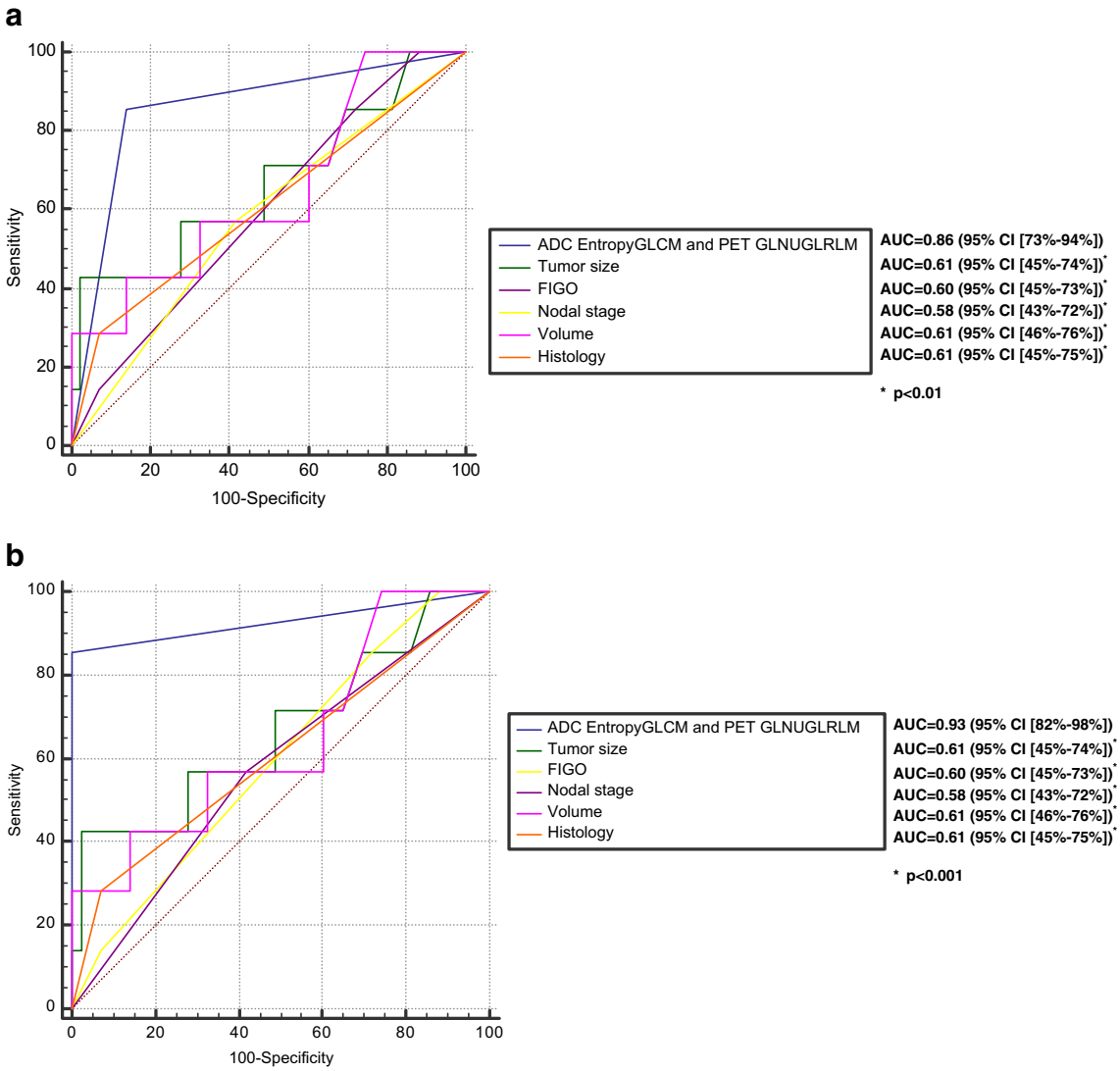

Fig. 4 LRC prediction in testing cohort from Nantes ($N = 50$) and McGill ($N = 28$). Comparison of ROC curves for (**a**, **b**) Nantes and (**c**, **d**) McGill according to the combination of MRI ADC $Entropy_{GLCM}$ and PET $GLNU_{GLRLM}$ in comparison with clinico-pathological features before (**a**–**c**) and after (**b**–**d**) ComBat Harmonization

Our results are in line with other studies in CC relying on PET or MRI. High pretreatment PET $GLNU_{GLRLM}$ was associated with poorer prognosis [7] and ^{18}F-FDG PET/CT features could predict local recurrence of LACC better than SUV_{max} [11]. DCE-MRI second-order textures were able to predict treatment outcome [8]. Another recent study developed an ADC MRI-based prognostic model including T and M stages [31]. However, the majority of patients were 1b and some were metastatic, which is different from our cohorts. In our study, only the radiomics features remained significant after correction for multiple testing and led to much higher predictive power than all clinical variables. Even though our study was the first to demonstrate the complementary prognostic value of radiomics features from both PET/CT and MRI images in patients with LACC, the developed models were only trained and internally validated in a single monocentric cohort. Although many features in both MRI and PET were significantly associated with both LRC and DFS, we selected the smallest subsets of features available to achieve high accuracy for each clinical endpoint. As a result, a single MRI feature (without the need for the PET feature) was sufficient to predict DFS, whereas the addition of a PET feature was necessary to reach the best accuracy to predict LRC. Our goal for the present study was thus to further validate these models in external cohorts.

One of the main challenges in validating radiomics-based models in cohorts from different centers is the need to pool imaging data generated with different imaging and reconstruction protocols, as it has been shown that radiomics features are often very sensitive to the acquisition and reconstruction parameters, amongst other factors [22, 32]. As a result of this variability of radiomics features, a radiomics-based model

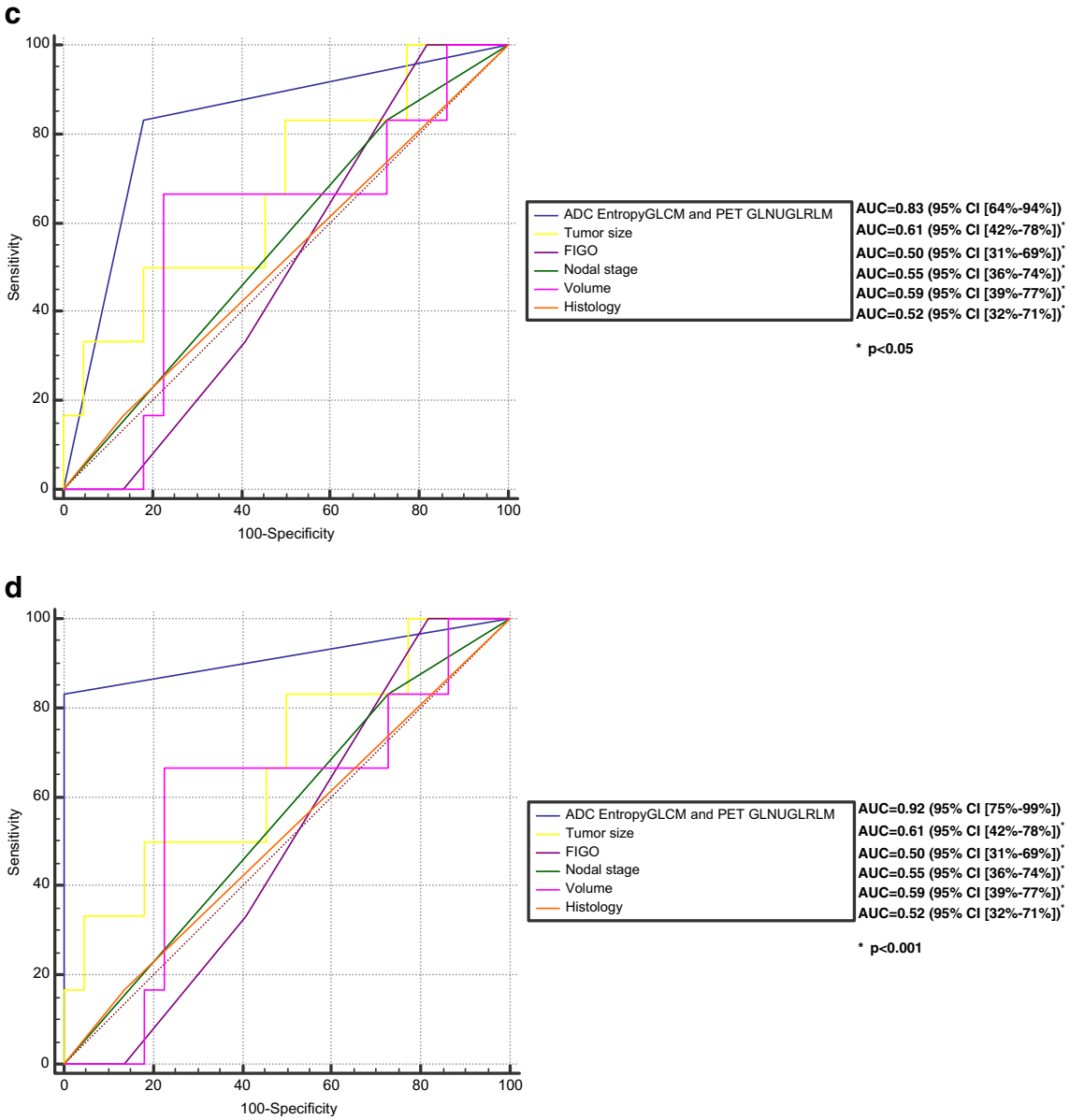

Fig. 4 (continued)

trained on data from a given clinical centre using a specific scanner and associated acquisition/reconstruction protocol, might not be directly applicable to data from another scanner, as recently demonstrated for FDG PET in CC [11]. This is a severe limitation for a broader transfer of radiomics to clinical practice. The genomics field faced a similar problem called "batch effect", where batch refers to the settings used to acquire the data, and hence is identical to the imaging protocol effect in radiomics. The ComBat harmonization method is now widely used in genomics, and has the advantage over other methods to provide satisfactory results even for small datasets with a limited number of features [33]. The ComBat method was also recently used to harmonize features from histopathological images for cancer diagnosis [34] or the cortical thickness measurements from MRI [26]. Regarding radiomics, a recent study showed that the ComBat method could successfully harmonize radiomics features extracted from PET images with different acquisition and reconstruction properties [27]. The ComBat method is fast and also easy to use. It only requires features extracted from patient data acquired in different departments, without requiring any phantom experiment, which makes it suitable for the analysis of retrospective data. Alternative methods have been considered previously to deal with the variability of radiomics features according to different reconstruction and/or acquisition settings, such as interpolating images to harmonize voxel sizes [35], or restrict the statistical analysis to the most robust subset of features [36]. Harmonizing acquisition and reconstruction

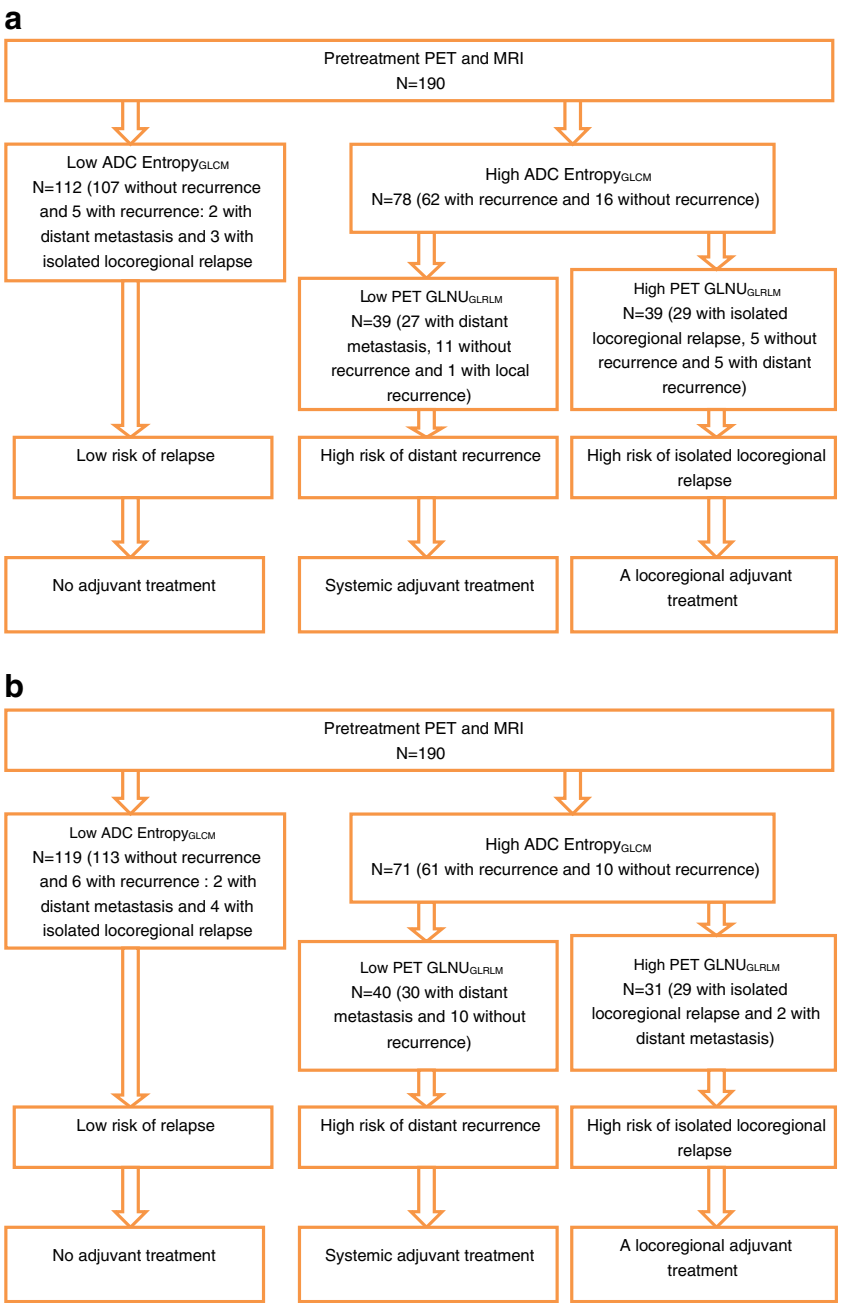

Fig. 5 Flow diagram of risk-stratification strategy based on pretreatment FDG PET and DWI MRI illustrated in the pooled set before (**a**) and after (**b**) Combat harmonization. The first step separates patients into two groups: low (first group) and high risk of relapse thanks to ADC Entropy$_{GLCM}$ from DWI MRI. The second step further discriminates within the high-risk group between metastatic (second group) or pelvic (third group) relapse. The first group would not require additional treatment. The second group could benefit from a complementary systemic treatment or intensified surveillance, and the third group could be treated with an additional locoregional treatment (like surgery or additional boost in brachytherapy)

protocols is only possible for prospective data collection, and is also limited since different centers can use different devices and may be reluctant to change their workflows. The use of the ComBat method is the most efficient way of addressing this issue with retrospectively collected images. Harmonizing voxel sizes by interpolating images before radiomics computation or restricting the analysis to the most robust features are only incomplete solutions.

Our results, obtained in two different external cohorts (one French, one Canadian), seem to strongly support the validity of the previously developed models. Indeed, even without the use of the harmonization method, the predictive performance was good in both training cohorts, with accuracy of 76–81%, suggesting the selected radiomics features are relatively robust to differences in acquisition and reconstruction settings. Using the ComBat method to harmonize radiomics features allowed the accuracies to improve to 81–97%. Even though the differences in PET and MRI protocols were not necessarily very important (for instance, the PET/CT scanner model and the imaging/reconstruction protocols were the same in Brest and Nantes), we demonstrated an important improvement of the models performance after ComBat harmonization. This simple additional statistical step allowed for instance specificity of the models in the testing cohorts to increase from 62 to 88%

for DFS prediction and from 80 to 100% for LRC prediction (when pooling the two testing cohorts). These results further confirm the potential value of the ComBat method to pool cohorts from different centers in radiomics studies, thus facilitating the training and validation of radiomics-based models in a larger multicentric context, whether the data are collected retrospectively or prospectively.

Our study has limitations. It was retrospective, which is however the case of most radiomics studies up to date [6], including those for CC [7–11]. Large confidence intervals were observed for most HR reported, which is explained by the relatively small size of the cohorts and the limited number of events. However, we would like to emphasize that even the lowest values of these interval remain high and clinically relevant. In the future, we intend to further validate our findings in a larger multicentric context including additional cohorts for a much larger number of patients. All images were analyzed and processed by a single expert using a single radiomics pipeline for segmentation and features calculation. The use of semi-automatic segmentation tools should reduce the user-dependency. Another analysis setting where the data would be processed in each center using different pipelines could lead to more variability and less robust models. The fact that radiomics features were calculated according to IBSI guidelines and validated with respect to the consensus obtained in that international standardization initiative should facilitate the reproducibility of our results by others if they adhere to the same guidelines.

Conclusions

Previously developed radiomics-based models relying on $GLNU_{GLRLM}$ from ^{18}F-FDG PET and $Entropy_{GLCM}$ from ADC maps derived from DW-MRI were validated in two independent external cohorts of patients. Prediction of recurrence and local regional control in LACC patients undergoing CRT could be achieved with almost perfect accuracy especially after statistical harmonization through the ComBat method, and with definitely higher predictive power than usual clinical factors such as the FIGO stage, or even the post-treatment metabolic response evaluation. The identification of high-risk patients at diagnosis can therefore allow for tailored treatment strategies involving higher doses of radiation boost, consolidation chemotherapy, and/or adjuvant hysterectomy, when indicated. This newly proposed management of patients based on these validated radiomics models should now be confirmed in future prospective studies, which, as was demonstrated in the present work, could be multicentric in nature thanks to the ability to perform a posteriori harmonization of radiomics features.

Compliance with ethical standards

Conflict of interest Authors François Lucia, Dimitris Visvikis, Martin Vallières, Marie-Charlotte Desseroit, Omar Miranda, Philippe Robin, Pietro Andrea Bonaffini, Joanne Alfieri, Ingrid Masson, Augustin Mervoyer, Caroline Reinhold, Olivier Pradier, Mathieu Hatt, Ulrike Schick declare that they have no conflict of interest.

No financial support was received for this work.

There are no potential conflicts of interest to disclose.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all individual participants included in the study.

Statement of translational relevance describing how the results might be applied to the future practice of cancer medicine Our findings have a direct impact on patient management in clinical practice. A flowchart demonstrates how to exploit the two radiomics features necessary to guide and personalize treatment: the textural feature ($Entropy_{GLCM}$) extracted from ADC maps derived from DWI-MRI acquisitions can identify patients with low risk of recurrence, for which it could be advised to avoid adjuvant treatment. Among the patients with a higher risk of recurrence, the second textural feature ($GLNU_{GLRLM}$) extracted from the FDG PET can differentiate between patients with a risk of distant recurrence, for which a systemic adjuvant treatment or more intensive surveillance could be recommended, and those with locoregional relapse, for which a locoregional adjuvant treatment might be more beneficial. Both MRI and PET images are routinely acquired for LACC patients, and the radiomics features have standardized definition with the IBSI guidelines, therefore anyone could easily evaluate our models in their data.

References

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer. 2015;136(5):E359–86.
2. Moore KN, Java JJ, Slaughter KN, Rose PG, Lanciano R, DiSilvestro PA, et al. Is age a prognostic biomarker for survival among women with locally advanced cervical cancer treated with chemoradiation? An NRG oncology/gynecologic oncology group ancillary data analysis. Gynecol Oncol. 2016;143(2):294–301.
3. Herrera FG, Prior JO. The role of PET/CT in cervical cancer. Front Oncol. 2013;3:34.
4. Choi J, Kim HJ, Jeong YH, Lee JH, Cho A, Yun M, et al. The role of (18) F-FDG PET/CT in assessing therapy response in cervix cancer after concurrent chemoradiation therapy. Nucl Med Mol Imaging. 2014;48(2):130–6.
5. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012;48(4):441–6.
6. Yip SS, Aerts HJ. Applications and limitations of radiomics. Phys Med Biol. 2016;61(13):R150–66.
7. Ho KC, Fang YH, Chung HW, Yen TC, Ho TY, Chou HH, et al. A preliminary investigation into textural features of intratumoral metabolic heterogeneity in (18)F-FDG PET for overall survival

prognosis in patients with bulky cervical cancer treated with definitive concurrent chemoradiotherapy. Am J Nucl Med Mol Imaging. 2016;6(3):166–75.
8. Torheim T, Groendahl AR, Andersen EK, Lyng H, Malinen E, Kvaal K, et al. Cluster analysis of dynamic contrast enhanced MRI reveals tumor subregions related to locoregional relapse for cervical cancer patients. Acta Oncol. 2016;55(11):1294–8.
9. Chung HH, Kang SY, Ha S, Kim JW, Park NH, Song YS, et al. Prognostic value of preoperative intratumoral FDG uptake heterogeneity in early stage uterine cervical cancer. J Gynecol Oncol. 2016;27(2):e15.
10. Guan Y, Li W, Jiang Z, Chen Y, Liu S, He J, et al. Whole-lesion apparent diffusion coefficient-based entropy-related parameters for characterizing cervical cancers: initial findings. Acad Radiol. 2016;23(12):1559–67.
11. Reuze S, Orlhac F, Chargari C, Nioche C, Limkin E, Riet F, et al. Prediction of cervical cancer recurrence using textural features extracted from 18F-FDG PET images acquired with different scanners. Oncotarget. 2017;8(26):43169–79.
12. Lucia F, Visvikis D, Desseroit MC, Miranda O, Malhaire JP, Robin P, et al. Prediction of outcome using pretreatment (18)F-FDG PET/CT and MRI radiomics in locally advanced cervical cancer treated with chemoradiotherapy. Eur J Nucl Med Mol Imaging. 2018;45(5):768–86.
13. Zwanenburg A, Lock S. Why validation of prognostic models matters? Radiother Oncol J Eur Soc Ther Radiol Oncol. 2018;127(3):370–3.
14. Vallieres M, Zwanenburg A, Badic B, Cheze Le Rest C, Visvikis D, Hatt M. Responsible radiomics research for faster clinical translation. J Nucl Med. 2018;59(2):189–93.
15. Lim K, Small W Jr, Portelance L, Creutzberg C, Jurgenliemk-Schulz IM, Mundt A, et al. Consensus guidelines for delineation of clinical target volume for intensity-modulated pelvic radiotherapy for the definitive treatment of cervix cancer. Int J Radiat Oncol Biol Phys. 2011;79(2):348–55.
16. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors. J Nucl Med. 2009;50(Suppl 1):122S–50S.
17. Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative - feature definitions. 2017.
18. Desseroit MC, et al. Comparison of three quantization methods for the calculation of textural features in PET/CT images: impact on prognostic models in non-small cell lung cancer. IEEE Nucl Sci Symp Med Imaging Conf 2016. 2016.
19. Chalkidou A, O'Doherty MJ, Marsden PK. False discovery rates in PET and CT studies with texture features: a systematic review. PLoS One. 2015;10(5):e0124165.
20. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Eur J Clin Investig. 2015;45(2):204–14.
21. Yang F, Dogan N, Stoyanova R, Ford JC. Evaluation of radiomic texture feature error due to MRI acquisition and reconstruction: a simulation study utilizing ground truth. Phys Med Int J Devoted Appl Phys Med Biol Off J Ital Assoc Biomed Phys. 2018;50:26–36.
22. Yan J, Chu-Shern JL, Loi HY, Khor LK, Sinha AK, Quek ST, et al. Impact of image reconstruction settings on texture features in 18F-FDG PET. J Nucl Med. 2015;56(11):1667–73.
23. Hatt M, Cheze le Rest C, Turzo A, Roux C, Visvikis D. A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET. IEEE Trans Med Imaging. 2009;28(6):881–93.
24. Hatt M, Cheze le Rest C, Descourt P, Dekker A, De Ruysscher D, Oellers M, et al. Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications. Int J Radiat Oncol Biol Phys. 2010;77(1):301–8.
25. Velazquez ER, Parmar C, Jermoumi M, Mak RH, van Baardwijk A, Fennessy FM, et al. Volumetric CT-based segmentation of NSCLC using 3D-slicer. Sci Rep. 2013;3:3529.
26. Fortin JP, Cullen N, Sheline YI, Taylor WD, Aselcioglu I, Cook PA, et al. Harmonization of cortical thickness measurements across scanners and sites. Neuroimage. 2018;167:104–20.
27. Orlhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L, Soussan M, Frouin F, Frouin V, Buvat I. A post-reconstruction harmonization method for multicenter radiomic studies in PET. J Nucl Med. 2018.
28. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118–27.
29. Fortin JP, Parker D, Tunc B, Watanabe T, Elliott MA, Ruparel K, et al. Harmonization of multi-site diffusion tensor imaging data. NeuroImage. 2017;161:149–70.
30. Naik A, Gurjar OP, Gupta KL, Singh K, Nag P, Bhandari V. Comparison of dosimetric parameters and acute toxicity of intensity-modulated and three-dimensional radiotherapy in patients with cervix carcinoma: a randomized prospective study. Cancer Radiother J de la Societe Francaise de Radiother Oncol. 2016;20(5):370–6.
31. Lin G, Yang LY, Lin YC, Huang YT, Liu FY, Wang CC, Lu HY, Chiang HJ, Chen YR, Wu RC, et al. Prognostic model based on magnetic resonance imaging, whole-tumour apparent diffusion coefficient values and HPV genotyping for stage IB-IV cervical cancer patients following chemoradiotherapy. Eur Radiol. 2018.
32. Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. Acta Oncol. 2010;49(7):1012–6.
33. Goh WWB, Wang W, Wong L. Why batch effects matter in omics data, and how to avoid them. Trends Biotechnol. 2017;35(6):498–507.
34. Kothari S, Phan JH, Stokes TH, Osunkoya AO, Young AN, Wang MD. Removing batch effects from histopathological images for enhanced cancer diagnosis. IEEE J Biomed Health Inf. 2014;18(3):765–72.
35. Mackin D, Fave X, Zhang L, Yang J, Jones AK, Ng CS, et al. Harmonizing the pixel size in retrospective computed tomography radiomics studies. PLoS One. 2017;12(9):e0178524.
36. Wu J, Aguilera T, Shultz D, Gudur M, Rubin DL, Loo BW Jr, et al. Early-stage non-small cell lung cancer: quantitative imaging characteristics of (18)F fluorodeoxyglucose PET/CT allow prediction of distant metastasis. Radiology. 2016;281(1):270–8.