SCIENTIFIC REP⚙RTS

OPEN

Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer

Martin Vallières [1], **Emily Kay-Rivest**[2], **Léo Jean Perrin**[3], **Xavier Liem**[4], **Christophe Furstoss**[5], **Hugo J. W. L. Aerts** [6], **Nader Khaouam**[5], **Phuc Felix Nguyen-Tan**[4], **Chang-Shu Wang**[3], **Khalil Sultanem**[2], **Jan Seuntjens**[1] & **Issam El Naqa** [7]

Quantitative extraction of high-dimensional mineable data from medical images is a process known as radiomics. Radiomics is foreseen as an essential prognostic tool for cancer risk assessment and the quantification of intratumoural heterogeneity. In this work, 1615 radiomic features (quantifying tumour image intensity, shape, texture) extracted from pre-treatment FDG-PET and CT images of 300 patients from four different cohorts were analyzed for the risk assessment of locoregional recurrences (LR) and distant metastases (DM) in head-and-neck cancer. Prediction models combining radiomic and clinical variables were constructed via random forests and imbalance-adjustment strategies using two of the four cohorts. Independent validation of the prediction and prognostic performance of the models was carried out on the other two cohorts (LR: AUC = 0.69 and CI = 0.67; DM: AUC = 0.86 and CI = 0.88). Furthermore, the results obtained via Kaplan-Meier analysis demonstrated the potential of radiomics for assessing the risk of specific tumour outcomes using multiple stratification groups. This could have important clinical impact, notably by allowing for a better personalization of chemo-radiation treatments for head-and-neck cancer patients from different risk groups.

Precision oncology promises to tailor the full spectrum of cancer care to an individual patient, notably in terms of personalization of cancer prevention, screening, risk stratification, therapy and response assessment. With sufficient infrastructure support and concerted efforts from the different stakeholders, it is possible to foresee that personalized therapy would become the standard of care in oncology[1]. Cancer mechanisms are increasingly elucidated as functions of different biomarkers or tumour genetic mutations, thereby changing the way we design clinical trials to achieve better cancer management efficacy in specific patient sub-populations[2]. On the other hand, rapid learning paradigms (i.e., knowledge-driven healthcare) consisting of reusing routine clinical data to develop knowledge in the form of models that can predict treatment outcomes for a larger portion of the population have also gained popularity in the oncology community[3, 4]. Although most research approaches to precision oncology are centered on genomics technologies[5, 6], it is thought that only the integration of multiple-omics, i.e., panomics data (genomics, transcriptomics, proteomics, metabolomics, etc.) could efficiently unravel biological mechanisms[7, 8].

The importance of panomics integration for cancer risk assessment emerges from the tremendous extent of heterogeneous characteristics expressed at multiple levels of tumours. Genes, proteins, cellular microenvironments, tissues and anatomical landmarks within tumours exhibit considerable spatial and temporal variations that could potentially yield valuable information about tumour aggressiveness. Tumours are generally composed

[1]Medical Physics Unit, McGill University, Cedars Cancer Centre, McGill University Health Centre – Glen Site, 1001 boulevard Décarie, Montréal, QC, H4A 3J1, Canada. [2]Radiation Oncology Division, Hôpital général juif, 3755 Côte Ste. Catherine, Montréal, QC, H3T 1E2, Canada. [3]Department of Radiation Oncology, Centre hospitalier universitaire de Sherbrooke, Hôpital Fleurimont, 3001 12e avenue N, Sherbrooke, QC, J1H 5N4, Canada. [4]Department of Radiation Oncology, Centre hospitalier de l'Université de Montréal, Hôpital Notre-Dame, 1560 Sherbrooke E, Montréal, QC, H2L 4M1, Canada. [5]Department of Radiation Oncology, Hôpital Maisonneuve-Rosemont, 5415 boulevard Assomption, Montréal, QC, H1T 2M4, Canada. [6]Departments of Radiation Oncology & Radiology, Dana-Farber Cancer Institute, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, 02215-5450, USA. [7]Department of Radiation Oncology, Physics Division, University of Michigan, 519 W. Williman St, Argus Bldg, Ann Arbor, Michigan, 48103-4943, USA. Correspondence and requests for materials should be addressed to M.V. (email: mart.vallieres@gmail.com)

of multiple clonal sub-populations of cancer cells forming complex dynamic systems that exhibit rapid evolution as a result of their interaction with their microenvironment and therapy perturbations[9]. Differing properties can be attributed to the different sub-populations in terms of growth rate, expression of biomarkers, ability to metastasize and immunological characteristics[10]. These properties could be described by differences in metabolic activity, cell proliferation, oxygenation levels, pH, blood vasculature and necrotic areas observed within the tumour. Such intratumoural differences are related to the concept of tumour heterogeneity, a characteristic that can be observed with significantly different extents even amongst tumours of the same histopathological type. Tumours exhibiting such heterogeneous characteristics are thought to be associated with high risk of resistance to treatment, progression, metastasis or recurrence[11–13].

Nowadays, medical imaging plays a central role in the investigation of intratumoural heterogeneity, as radiological images are acquired as routine practice for almost every patient with cancer. Medical images such as 2-deoxy-2-[^{18}F]fluoro-D-glucose (FDG) positron emission tomography (PET) and X-ray computed tomograph (CT) are minimally invasive and they carry an immense source of potential data for decoding tumour phenotypes[14]. The quantitative extraction of high-dimensional mineable data from all types of medical images and whose subsequent analysis aims at supporting clinical decision-making is a process coined with the term "radiomics"[15–18]. The demonstration that gene-expression signatures and clinical phenotypes could be inferred from tumour imaging features[19–21] has led to an exponential growth of this field in the past few years[22, 23]. The underlying hypothesis of radiomics is that the genomic heterogeneity of aggressive tumours could translate into heterogeneous tumour metabolism and anatomy, thereby envisioning the quantitative analysis of diagnostic medical images as an essential prognostic tool for cancer risk assessment and as an integral part of panomic tumour signature profiling.

The translation of radiomics analysis into standard cancer care to support treatment decision-making involves the development of prediction models integrating clinical information that can assess the risk of specific tumour outcomes[24] (Fig. 1). In this work, our main objective is to construct prediction models using advanced machine learning to evaluate the risk of locoregional recurrences and distant metastases prior to chemo-radiation of head-and-neck (H&N) cancers, a group of biologically similar neoplasms originating from the squamous cells that line the mucosal surfaces in the oral cavity, paranasal sinuses, pharynx or larynx. The locoregional control of H&N cancers is usually good, but this is, however, not matched by improvements in survival, as the development of distant metastases and second primary cancers are the leading causes of treatment failure and death[25, 26]. In order to improve patient survival and outcomes, the importance of identifying relevant prognostic factors that can better assess the aggressiveness of tumours at the moment of diagnosis is crucial.

We hypothesize that radiomic features are important prognostic factors for the risk assessment of specific H&N cancer outcomes[27]. The machine learning strategy employed in this work involves the extraction of 1615 different radiomic features from a total of 300 patients from four different institutions. Two cohorts are used to construct the prediction models by combining radiomics (intensity, shape, textures) and clinical attributes (patient age, H&N type, tumour stage) via random forests classifiers and imbalance-adjustments of training samples, and the remaining two cohorts are reserved to evaluate the prediction (binary assessment of outcome) and prognostic (time-to-event assessment) performance of the corresponding models (Fig. 2). Throughout this study, results obtained for locoregional recurrences and distant metastases are also compared against prediction models constructed for the general risk assessment of overall survival in H&N cancer. A comprehensive comparison of the prediction/prognostic performance of radiomics versus clinical and volumetric variables is also performed. Our results suggest that the integration of radiomic features into clinical prediction models has considerable potential for assessing the risk of specific outcomes prior to treatment of H&N cancer. Accurate stratification of locoregional recurrence and distant metastasis risks could eventually provide a rationale for adapting the radiation doses and chemotherapy regimens that the patients receive. Overall, combining quantitative imaging information with other categories of prognostic factors via advanced machine learning could have a profound impact on the characterization of tumour phenotypes and would increase the possibility of translation of outcome prediction models into the clinical environment as a means to personalize treatments.

Results

Summary of presentation of results.
To ease reading and the understanding of this study, a summary of how results are presented in the text is provided in Supplementary Fig. S1.

Association of variables with tumour outcomes.
In order to assess the value of quantitative pre-treatment imaging to predict specific tumour outcomes in H&N cancer, we performed a comprehensive univariate analysis of the association of radiomic features with locoregional recurrences ("LR" or "Locoregional"), distant metastases development ("DM" or "Distant") and overall survival ("OS" or "Survival" or "Death"). A total of 1615 radiomic features (Fig. 1b and Supplementary Methods section 2.5 for complete description) were first extracted from the gross tumour volume ($GTV_{primary} + GTV_{lymph\ nodes}$) of the FDG-PET and CT images (Fig. 1a), for all 300 patients from the four H&N cancer cohorts (Fig. 2): (I) 10 first-order statistics features (intensity); II) 5 morphological features (shape); and III) 40 texture features each extracted using 40 different combinations of parameters. We also compared these results to the predictive power of the tumour volume ("*Volume*") and of the following clinical variables: *Age*, *T-Stage*, *N-Stage*, *TNM-Stage* and human papillomavirus status (*HPV status*), where *HPV status* was available for 120 of the 300 patients (Supplementary Tables S5–S8). The association of the different variables with the different H&N cancer outcomes (binary endpoints) was then analyzed using Spearman's rank correlations (r_s) computed on all patients, and significance was assessed by applying multiple testing corrections using the Benjamini-Hochberg procedure[28] with a false discovery rate of 10%.

Overall, we found that 0%, 63% and 12% of the total radiomic features extracted from PET scans, and that 0%, 61% and 34% of the total radiomic features extracted from CT scans were significantly associated with LR, DM
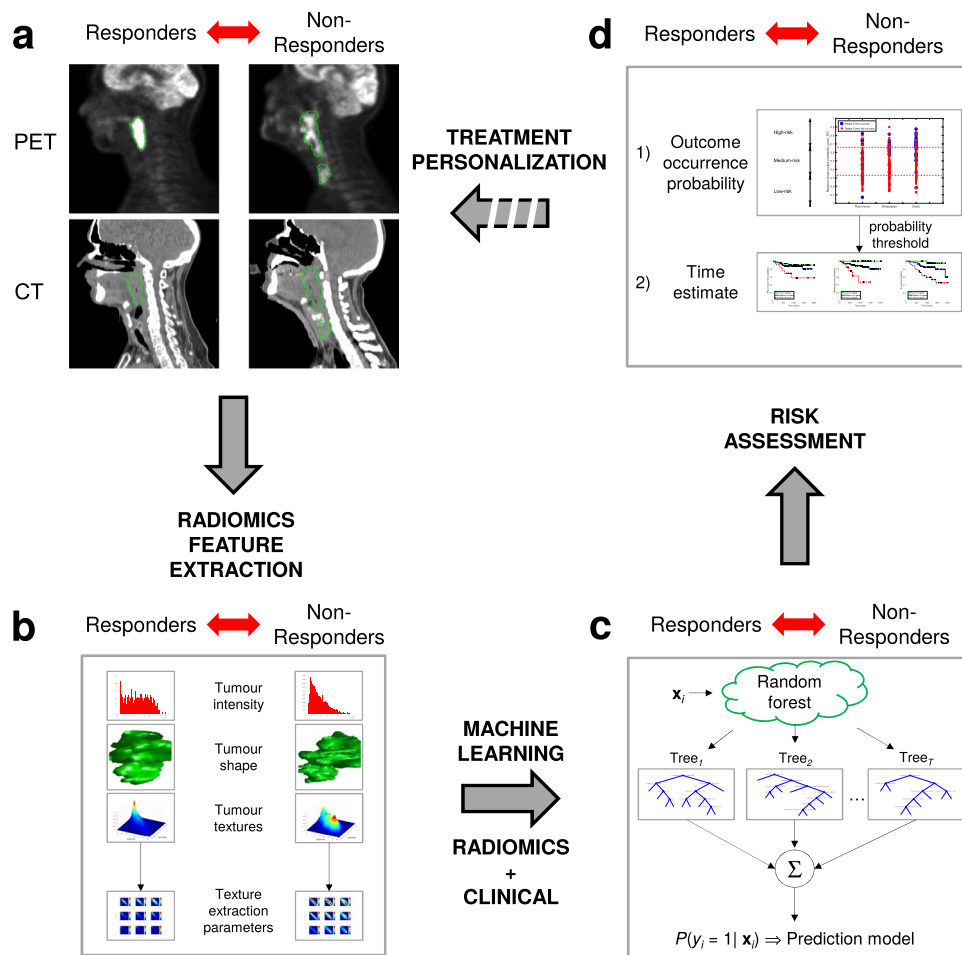
Figure 1. From radiomics analysis to treatment personalization. (**a**) Example of diagnostic FDG-PET and CT images of two head-and-neck cancer patients with tumour contours. The patient that did not respond well to treatment (right) has a more heterogeneous intratumoural intensity distribution in both FDG-PET and CT images than the patient that responded well to treatment (left). (**b**) The radiomics analysis strategy involves the extraction of features differentiating responders from non-responders to treatment. Features are extracted from the FDG-PET and CT tumour contours and quantify tumour shape, intensity, and texture. (**c**) Advanced machine learning combines radiomic features and patient clinical information via a random forest algorithm. The classifier is trained to differentiate between responders and non-responders to treatment (prediction model). (**d**) The output probability of the random forest classifier computed on new patients can be used to assess the risk of non-response to treatment via probabilities of occurrence of outcome events and time estimates. Eventually, accurate risk assessment of specific tumour outcomes via radiomics analysis could help to better personalize cancer treatments.

and OS, respectively (after multiple testing corrections). The radiomic features (PET or CT) with the highest associations with LR, DM and OS were $LZHGE_{GLSZM}$ from CT scans ($r_s = -0.15$, $p = 0.007$), ZSN_{GLSZM} from CT scans ($r_s = -0.29$, $p = 2 \times 10^{-7}$) and GLV_{GLRLM} from CT scans ($r_s = 0.24$, $p = 4 \times 10^{-5}$), respectively (Supplementary Table S1). Tumour volume was not found to be significantly associated with LR ($r_s = -0.04$, $p = 0.48$), but was significantly associated with DM ($r_s = 0.24$, $p = 3 \times 10^{-5}$) and OS ($r_s = -0.18$, $p = 2 \times 10^{-3}$). Finally, we found that {*Age*, *T-Stage*, *N-Stage*, *HPV status*}, {*N-Stage*} and {*Age*, *T-Stage*, *HPV status*} were significantly associated with LR, DM and OS, respectively. The clinical variables with the highest associations with LR, DM and OS were $HPV-$ ($r_s = 0.39$, $p = 8 \times 10^{-6}$), higher *N-Stage* ($r_s = 0.18$, $p = 1 \times 10^{-3}$) and higher *T-Stage* ($r_s = 0.21$, $p = 3 \times 10^{-4}$), respectively (Supplementary Table S2).

Construction of prediction models. The construction of prediction models for LR, DM and OS was carried out using a training set consisting of the combination of 194 patients from the H&N1 and H&N2 cohorts (Fig. 2). Three initial radiomic feature sets were considered: I) the 1615 radiomic features extracted from PET scans ("*PET*" feature set); II) the 1615 radiomic features extracted from "*CT*" scans ("*CT*" feature set); and III) a combined set containing all "*PET*" and "*CT*" radiomic features used in feature sets I and II ("*PETCT*" feature set).

Prediction models consisting of radiomic information only were first constructed for each of the three H&N outcomes and the three initial radiomic feature sets. *Feature set reduction*, *feature selection*, *prediction performance*
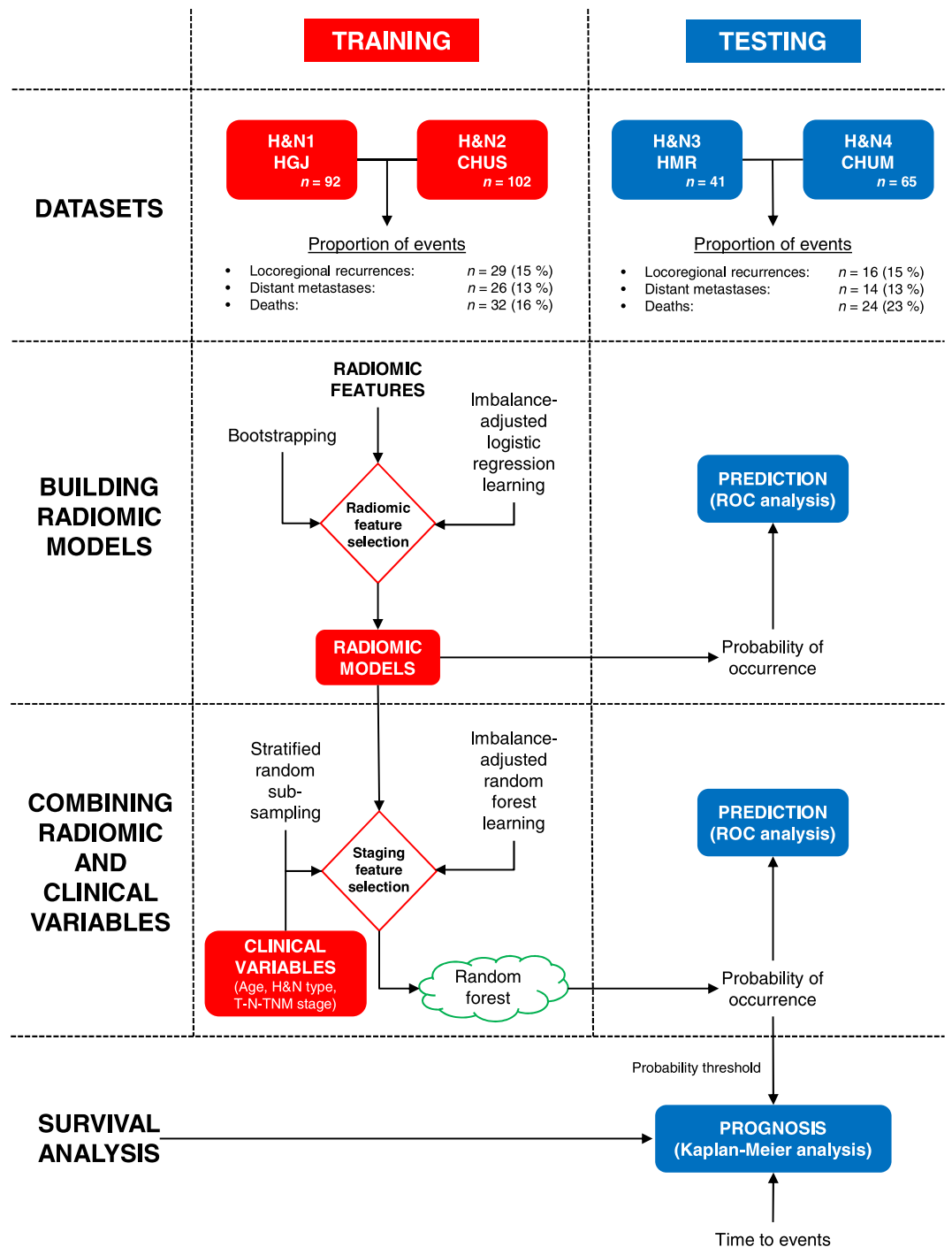
Figure 2. Models construction strategy and analysis workflow. Four different cohorts were used to demonstrate the utility of radiomics analysis for the pre-treatment assessment of the risk of locoregional recurrence and distant metastases in head-and-neck cancer. The H&N1 and H&N2 cohorts were combined and used as a single training set ($n = 194$), whereas the H&N3 and H&N4 cohorts were combined and used as a single testing set ($n = 106$). The best combinations of radiomic features were selected in the training set using imbalance-adjusted logistic regression learning and bootstrapping validations. These radiomic features were combined with selected clinical variables in the training set using imbalance-adjusted random forest learning and stratified random sub-sampling validations. Independent prediction analysis was performed in the testing set for all classifiers fully constructed in the training set. Independent prognosis analysis and Kaplan-Meier risk stratification was carried out in the testing set using the output probability of occurrence of events of random forests fully constructed in the training set.
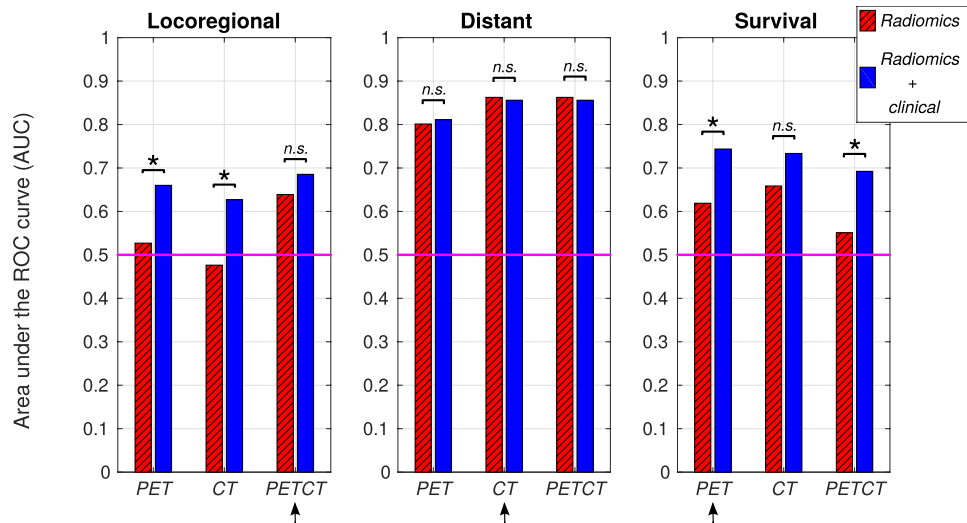
Figure 3. Prediction performance of selected models. All prediction models were selected and built using the training set (H&N1 and H&N2; $n = 194$) for three initial radiomic feature sets: I) PET radiomic features (*PET*); II) CT radiomic features (*CT*); and III) PET and CT radiomic features (*PETCT*). The prediction performance is evaluated here in terms of the area under the receiver operating characteristic curve (AUC) for patients of the testing set (H&N3 and H&N4; $n = 106$), for two types of prediction models: I) Radiomic models constructed using logistic regression (*Radiomics*); and II) Radiomic models combined with clinical variables via random forests (*Radiomics + clinical*). Significant increase in AUC from *Radiomics* to *Radiomics + clinical* models is identified with an asterisk (*), and non-significant increase is identified by *n.s.* The radiomic feature sets providing the prediction models with highest performance in this study are identified with an arrow for each outcome.

estimation, *choice of model complexity* (Supplementary Fig. S2) and *final model computation* processes were carried out using logistic regression and bootstrap resampling, similarly to the methodology developed in the study of Vallières *et al.*[29]. To account for the disproportion of occurrence of events and non-occurrence of events in the training set (15% LR, 13% DM, 16% deaths), an imbalance-adjustment strategy adapted from the study of Schiller *et al.*[30] was also applied during the training process. Overall for the *PET*, *CT* and *PETCT* feature sets, the number of variables forming the final radiomic models for each outcome were, respectively: I) 8, 3 and 3 radiomic variables for the LR outcome; II) 6, 3 and 3 radiomic variables for the DM outcome; and III) 4, 3 and 6 radiomic variables for the OS outcome.

The construction of prediction models combining radiomic and clinical variables was then carried out for the nine identified radiomic models (3 feature sets × 3 outcomes). By estimating prediction performance via stratified random sub-sampling in the training set, the following group of clinical variables were first selected for each outcome: I) {*Age*, *H&N type*, *T-Stage*, *N-Stage*} for LR prediction; II) {*Age*, *H&N type*, *N-Stage*} for DM prediction; and III) {*Age*, *H&N type*, *T-Stage*, *N-Stage*} for OS prediction. Final prediction models were ultimately constructed for each radiomic feature set and H&N outcome by combining the selected radiomic and clinical variables via random forests and imbalance adjustments.

Performance of prediction models. The performance of the radiomic prediction models constructed using logistic regression and of the prediction models constructed by combining radiomic and clinical variables via random forests was validated in a testing set consisting of the combination of 106 patients from the H&N3 and H&N4 cohorts (Fig. 2) using receiver operating characteristic (ROC) metrics (binary endpoints).

Figure 3 presents the performance results (AUC: area under the ROC curve) obtained in the testing set for the *radiomics* and *radiomics + clinical* models, where the significance of the increase in AUC when combining clinical to radiomic variables is assessed using the method of DeLong *et al.*[31]. Sensitivity, specificity and accuracy of predictions are also presented in Supplementary Fig. S3. Overall, it can be observed that there is a general increase in prediction performance for many of the different categories of models that we constructed in this work. For LR prediction, the increase in AUC is significant for prediction models from the *PET* ($p = 0.03$) and the *CT* ($p = 0.01$) radiomic feature sets. For DM prediction, none of the radiomics models show a significant AUC increase when combined with clinical variables. For OS (death) prediction, the increase in AUC is significant for prediction models from the *PET* ($p = 0.01$) and the *PETCT* ($p = 0.006$) radiomic feature sets. Furthermore, we verified that the increase in performance is not explained by the use of a more complex and potentially more predictive learning algorithm: random forests classifiers constructed with radiomic variables alone preserved the predictive properties obtained by logistic regression models constructed with the same variables, but without improving them (Supplementary Table S3). These results point to the potential of random forests in successfully combining the complementary value of different categories of prognostic factors such as radiomic and clinical variables.

In Fig. 3, the highest performance for LR prediction was obtained using the model combining the *PETCT* radiomic and clinical variables, with an AUC of 0.69. For DM prediction, the highest performance was obtained

Outcome	Variables	Prediction				Prognosis	
		AUC[a]	Sensitivity[a]	Specificity[a]	Accuracy[a]	CI[b]	p-value[c]
Locoregional	Radiomics$_{PETCT}$	0.64	0.56	0.67	0.65	0.63	0.28
	Volume	0.43	0.31	0.58	0.54	0.40	0.80
	Clinical	0.72	0.50	0.76	0.72	0.69	0.05
	Radiomics$_{PETCT}$ + Clinical	*0.69*	*0.63*	*0.68*	*0.67*	*0.67*	*0.03*
	Volume + Clinical	0.71	0.50	0.76	0.72	0.68	0.06
Distant	Radiomics$_{CT}$	0.86	0.79	0.77	0.77	0.88	0.0001
	Volume	0.80	0.86	0.65	0.68	0.83	0.10
	Clinical	0.55	0.64	0.46	0.48	0.60	0.61
	Radiomics$_{CT}$ + Clinical	*0.86*	*0.86*	*0.76*	*0.77*	*0.88*	*0.000003*
	Volume + Clinical	0.78	1	0.50	0.57	0.80	0.0004
Survival	Radiomics$_{PET}$	0.62	0.58	0.66	0.64	0.60	0.03
	Volume	0.68	0.67	0.57	0.59	0.67	0.29
	Clinical	*0.78*	*0.92*	*0.57*	*0.65*	*0.76*	*0.00003*
	Radiomics$_{PET}$ + Clinical	0.74	0.79	0.57	0.62	0.71	0.002
	Volume + Clinical	0.79	0.88	0.52	0.60	0.76	0.0006
Survival[d]	Radiomics$_{CTcompleteSign}$[e]	—	—	—	—	0.66	0.70
	Radiomics$_{CTsign}$[f]	0.68	0.71	0.50	0.55	0.66	0.05
	Radiomics$_{CTsign}$[g] + Clinical	0.80	0.96	0.38	0.51	0.75	0.001

Table 1. Comparison of prediction/prognostic performance of models constructed in this work with other combinations of variables. → Performance is shown for models constructed in the training set (H&N1 and H&N2; $n = 194$) and independently evaluated in the testing set (H&N3 and H&N4; $n = 106$). → Models involving *Radiomic* variables only or the *Volume* variable only were optimized using logistic/cox regression. All models involving *Clinical* variables were optimized using random forests. → The best predictive/ prognostic and balanced models for each outcome (final models) are identified in italic and are fully described in Supplementary Table S4. [a]Binary prediction of outcomes using logistic regression/random forest output responses. [b]Concordance-index between cox regression/random forest output responses and time to events. [c]Log-rank test from Kaplan-Meier curves with a risk stratification into two groups (thresholds: median hazard ratio for cox regression, output probability of 0.5 for random forests). [d]Radiomic signature variables as defined in the study of Aerts & Velazquez *et al.*[21]. [e]Using the original definition of the radiomic signature variables, and the original cox regression coefficients and median hazard ratio trained from the Lung1 cohort in the study of Aerts & Velazquez *et al.*[21]. [f]Using a revised version of the radiomic signature variables (Supplementary Methods section 2.6.2) and new cox/logistic regression coefficients trained using the current training set of this work. [g]Using a revised version of the radiomic signature variables (Supplementary Methods section 2.6.2) and a random forest classifer trained using the current training set of this work.

using the *CT* radiomic model, with an AUC of 0.86. These results demonstrate that different radiomic-based models could successfully be used to predict specific outcomes such as locoregional recurrences and distant metastases in H&N cancer. Finally, the highest performance for OS (death) prediction was obtained using the model combining the *PET* radiomic and clinical variables, with an AUC of 0.74. For subsequent analysis in the next section, only the prediction models (*radiomics* and *radiomics + clinical*) constructed from these radiomic features sets (*PETCT* for LR, *CT* for DM, *PET* for OS) are used. The complete description of these identified radiomic models (specific features, texture extraction parameters, logistic regression coefficients) is given in Supplementary Results section 1.4.2.

Comparison with other prognostic factors. The performance of the best radiomic prediction models and the best prediction models combining radiomic and clinical variables identified in this study (shown with arrows in Fig. 3) were further compared against other prognostic factors: I) *Volume*; II) "clinical-only" models; III) combination of *Volume* and clinical variables; and IV) a validated radiomic signature developed for the prognosis assessment of overall survival[21, 32]. In addition to the prediction performance evaluated using ROC metrics, the prognostic performance of the models was also assessed using: I) the concordance index (CI)[33] between the output probability of occurrence of an event (LR, DM, death) of prediction models and the time-to-event, defined as the time elapsed between the date the treatment ended and the date when an event occurred or the date of last-follow-up; and II) the p-value obtained from Kaplan-Meier analysis using the log-rank test between two risk groups. The models consisting of only radiomic or the *Volume* variables were optimized using logistic or cox regression, and all models involving clinical variables were optimized using random forest classifiers, still using the defined training set of this work (H&N1 and H&N2 cohorts; $n = 194$). Fully independent results are then presented in Table 1 for models evaluated in the testing set (H&N3 and H&N4 cohorts; $n = 106$).

For locoregional recurrences, we found that the model combining the *PETCT* radiomic and clinical variables provided the best performance in terms of predictive/prognostic power and balance of classification of occurrence and non-occurrence of events, notably with an AUC of 0.69, a sensitivity of 0.63, a specificity of 0.68, an accuracy of 0.67, a CI of 0.67 and a Kaplan-Meier p-value of 0.03. Using random permutation tests, each variable

was calculated to be approximately of equal importance in the random forest model (Supplementary Table S4). Similarly to univariate analysis, *Volume* was not found to be a significant prognostic factor for LR. On the other hand, clinical variables alone had high performance with an AUC of 0.72 and a CI of 0.69, but this type of modeling did not provide sufficient balance between the prediction of occurrence and non-occurrence of events (sensitivity of 0.50, specificity of 0.76).

For distant metastases, we found that the model combining the *CT* radiomic and clinical variables provided the best overall performance, notably with an AUC of 0.86, a sensitivity of 0.86, a specificity of 0.76, an accuracy of 0.77, a CI of 0.88 and a Kaplan-Meier *p*-value of 3×10^{-6}. However, radiomic variables were found to be of much higher importance than the clinical variables in the random forest model (Supplementary Table S4). In fact, the model composed of clinical variables alone did not perform well. *Volume* was again found to be a significant prognostic factor for DM, but radiomic variables outperformed it.

For overall survival, we found that the model composed of clinical variables alone provided the best overall performance, notably with an AUC of 0.78, a sensitivity of 0.92, a specificity of 0.57, an accuracy of 0.65, a CI of 0.76 and a Kaplan-Meier *p*-value of 3×10^{-5}. Furthermore, the *H&N type* variable had the highest and *N-Stage* the lowest importance in the model (Supplementary Table S4). Another important finding was that *Volume* alone provided similar or better prognosis assessment of OS than any of the following radiomic-based models: I) the best radiomic model for OS constructed in this work; II) the original radiomic signature using the cox regression coefficients employed in the work of Aerts & Velazquez *et al.*[21]; and III) a revised version of the radiomic signature computation (Supplementary Section 2.6.2) using new sets of regression coefficients trained with the current training set of this work.

Risk assessment of tumour outcomes. The work performed in this study leads to the identification of three prediction models based on three final random forest classifiers, one for each of the three outcomes studied here (identified with italic fonts in Table 1): I) {*PET-GLN*$_{GLSZM}$, *CT-Correlation*$_{GLCM}$, *CT-LGZE*$_{GLSZM}$, *Age*, *H&N type*, *T-Stage*, *N-Stage*} for LR; II) {*CT-LRHGE*$_{GLRLM}$, *CT-ZSV*$_{GLSZM}$, *CT-ZSN*$_{GLSZM}$, *Age*, *H&N type*, *N-Stage*} for DM; and III) {*Age*, *H&N type*, *T-Stage*, *N-Stage*} for OS. A property of a random forest is that the binary prediction of each of its decision trees can be averaged to serve as an output probability of occurrence of a given event (*prob*$_{RF}$). This output probability, similarly to other machine learning algorithms, can constitute one of the tools to be used for the risk assessment of specific tumour outcomes. For example, the final random forest classifiers constructed with the training set (H&N1 and H&N2 cohorts; $n = 194$) can be used to stratify the risk of occurrence of the outcome events for each patient of the testing set (H&N3 and H&N4 cohorts; $n = 106$) into three groups (Fig. 4a): I) low-risk group $\rightarrow 0 \leq prob_{RF} < \frac{1}{3}$; II) medium-risk group $\rightarrow \frac{1}{3} \leq prob_{RF} < \frac{2}{3}$; and III) high-risk group $\rightarrow \frac{2}{3} \leq prob_{RF} < 1$. Thereafter, this stratification scheme can be used to evaluate the probability of non-occurrence of the events after a given time for the different risk groups via Kaplan-Meier analysis. Standard Kaplan-Meier analysis using two risk groups ($prob_{RF} \leq 0.5$, $prob_{RF} > 0.5$) is first shown in Fig. 4b for all patients of the testing set. These curves demonstrate the possibility of prognostic risk assessment of specific outcomes in H&N cancer such as locoregional recurrences ($p = 0.03$) and distant metastases ($p = 3 \times 10^{-6}$) using specific prediction models combining different radiomic and clinical variables, but also of the general outcome of overall survival ($p = 3 \times 10^{-5}$) using a prediction model composed of clinical variables only. More accurate prognostic risk assessment can then be further performed using Kaplan-Meier analysis with three risk groups (as defined above: low-risk, medium-risk, high-risk) as shown in Fig. 4c for all patients of the testing set. For the risk assessment of LR, the developed prediction model is, however, not powerful enough to significantly separate the patients between the high/medium ($p = 0.62$) and medium/low ($p = 0.10$) risk groups. In the case of DM, the developed prediction model allows to significantly separate the patients between the high/medium ($p = 0.05$) and medium/low ($p = 0.03$) risk groups. For OS, the developed prediction model does not significantly separate the patients between the high/medium risk groups ($p = 0.07$), but it does significantly separate the patients between the medium/low risk groups ($p = 0.02$).

Discussion

Increasing evidence suggests that the genomic heterogeneity of aggressive tumours could translate into intratumoural spatial heterogeneity exhibited at the anatomical and functional scales[19–21]. This constitutes the central idea of the emerging field of "radiomics", in which large amounts of information via advanced quantitative analysis of medical images are used as non-invasive means to characterize intratumoural heterogeneity and to reveal important prognostic information about the cancer[15–18]. Ultimately, the objective is to narrow down this extensive quantity of information into simple prediction models that can aid in the identification of specific tumour phenotypes for improved treatment management. In this study, we were able via advanced machine learning to develop two prediction models combining PET/CT radiomics and clinical information for the early assessment of the risk of locoregional recurrences and distant metastases in head-and-neck cancer.

First, we extracted a total of 1615 radiomic features from PET and CT pre-treatment images of 300 patients with head-and-neck cancer from four different cohorts. These features are composed of 10 intensity features, 5 shape features and 40 textures computed using 40 different combinations of extraction parameters (five isotropic voxel sizes, two quantization algorithms and four numbers of gray levels). In general, different texture features better represent the underlying tumour biology using different extraction parameters, and the optimal set of parameters to use is application-specific and depends on many factors such as the clinical endpoint studied and the imaging modalities employed. Texture optimization has the potential to enhance the predictive value of the extracted features as Vallières *et al.*[29] have previously shown, and we suggest to incorporate this step in the texture extraction workflow of future similar studies.

Univariate analysis showed that the majority of the features extracted from both PET and CT images are significantly associated with the development of distant metastases, suggesting that the metastatic phenotype
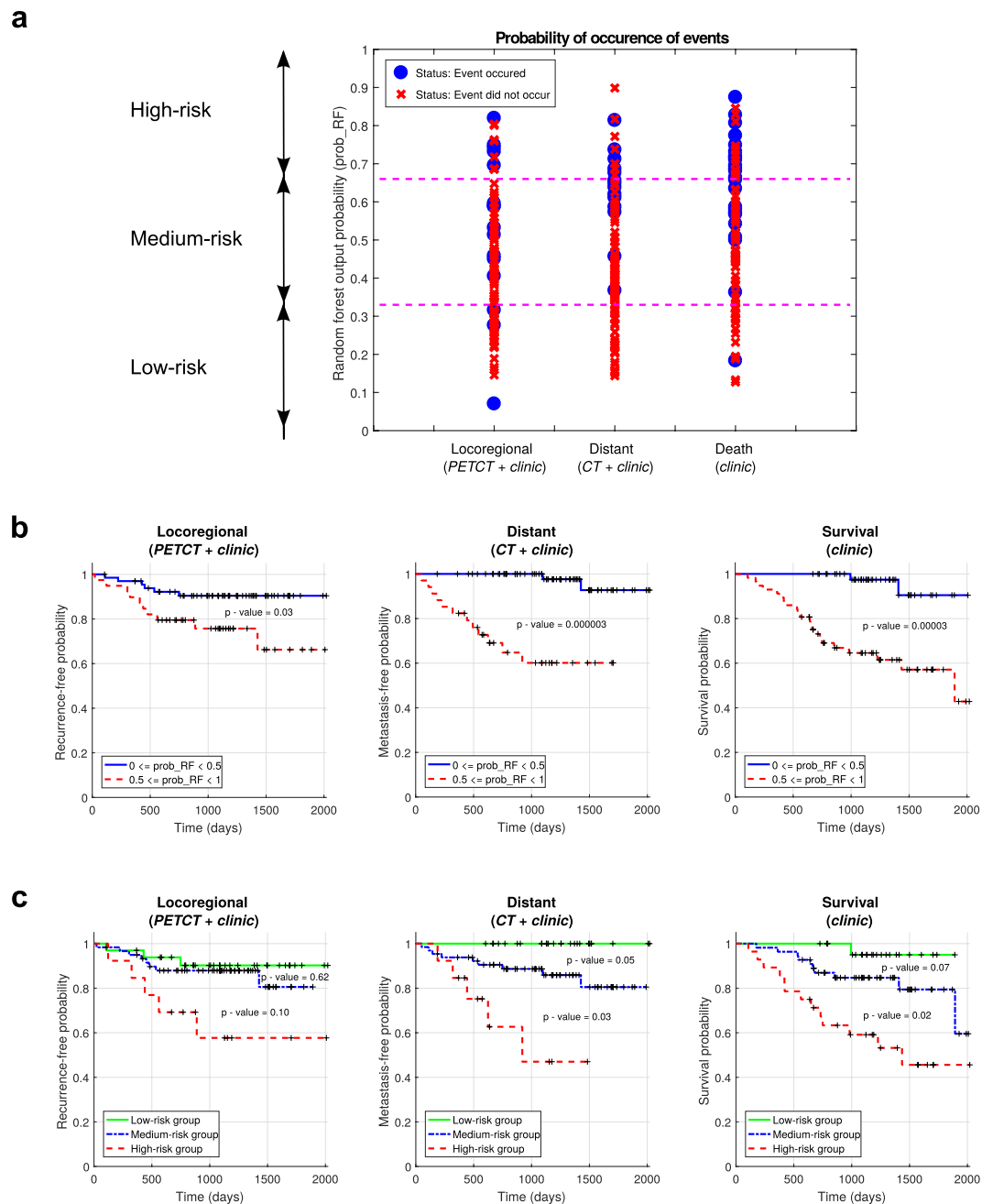
Figure 4. Risk assessment of tumour outcomes. (**a**) Probability of occurrence of events (locoregional recurrence, distant metastases, death) for each patient of the testing set (H&N3 and H&N4; $n = 106$) as determined by the random forest classifiers built using the training set (H&N1 and H&N2; $n = 194$). The output probability of occurrence of events of random forests allows for risk stratification; for example, three risk groups can be defined (low, medium, high) using probability thresholds of $\frac{1}{3}$ and $\frac{2}{3}$. (**b**) Kaplan-Meier curves of the testing set using a risk stratification into two groups as defined by a random forest output probability threshold of 0.5. All curves have significant prognostic performance, thus demonstrating the possibility of outcome-specific risk assessment in head-and-neck cancer. (**c**) Kaplan-Meier curves of the testing set using a risk stratification into three groups as defined by random forest output probability thresholds of $\frac{1}{3}$ and $\frac{2}{3}$. Some pair of curves have significant prognostic performance, thus demonstrating the possibility of risk stratification into multiple groups for treatment escalation/personalization in head-and-neck cancer.

of tumours could be captured via quantitative image analysis. On the other hand, none of the radiomic features were significantly associated with locoregional recurrences after multiple testing corrections with a FDR of 10%. Although combinations of these metrics still proved useful in subsequent multivariable analysis (Table 1: AUC of 0.64), it does reveal the need of using other types of metrics such as radiation dose characteristics to enhance the

predictive properties of the radiomic-based models constructed for locoregional recurrences. In addition to radiomic features, we also investigated the association of clinical variables with the different head-and-neck cancer outcomes studied in this work. The most significant association was found between *HPV status* and locoregional recurrences, a currently known result that agrees with other studies[34, 35]. However, this result was obtained with only 120 of the 300 patients with available *HPV status*, and this variable could not be used in the subsequent multivariable analysis. Similarly, the performance of radiation dose metrics for the prediction of locoregional recurrences could not be investigated in this study due to the non-availability of radiation plans.

Next, we constructed multivariable prediction models from radiomic information alone using the methodology developed by Vallières *et al.*[29]. All models were entirely produced from the defined training set of this work combining two head-and-neck cancer cohorts (H&N1 and H&N2; $n = 194$). The best radiomic model for locoregional recurrences was found to possess significant predictive properties (Table 1: AUC of 0.64) in the defined testing set of this work combining two other head-and-neck cancer cohorts (H&N3 and H&N4; $n = 106$), thereby suggesting that radiomics analysis of PET/CT images prior to chemo-radiation treatments could have a role to play in decoding the local control tumour phenotype in head-and-neck cancer. This model is composed of one metric extracted from PET images (GLN_{GLSZM}: *gray-level nonuniformity*$_{GLSZM}$) and two metrics extracted from CT images ($correlation_{GLCM}$ and $LGZE_{GLSZM}$: *low gray-level zone emphasis*$_{GLSZM}$). On the other hand, the best radiomic model for distant metastases was found to possess highly significant predictive properties in the testing set (Table 1: AUC of 0.86), and is composed of three metrics extracted from CT images ($LRHGE_{GLRLM}$: *long run high gray level emphasis*$_{GLRLM}$, ZSV_{GLSZM}: *zone size variance*$_{GLSZM}$ and ZSN_{GLSZM}: *zone size nonuifomity*$_{GLSZM}$). Overall, the prediction results obtained for these two models suggest that radiomics analysis can be specific enough to assess the risk of different outcomes in head-and-neck cancer. The models we developed for locoregional recurrences and distant metastases are in fact different and they may capture specific tumour phenotypes.

Of important note, all the selected radiomic features of the locoregional and distant metastases prediction models are textural. Aggressive tumours tend to show increased intratumoural heterogeneity[11–13], an effect that may be captured by the different textural metrics of the models when extracted from PET and CT images. For example, the intrinsic definitions of the GLN_{GLSZM}, $LRHGE_{GLRLM}$, ZSV_{GLSZM} and ZSN_{GLSZM} metrics imply that these texture features are related (with different extents) to the quantitative description of the size and intensity variations of the connected sub-regions of a given region of interest. These metrics are thus related to the heterogeneity in size and intensity characteristics of the different tumour sub-regions in PET and CT images, a result in agreement with a previous study describing the importance of zone-size nonuniformities for the prognostic assessment of head-and-neck tumours[36]. Furthermore, the intrinsic definitions of the $Correlation_{GLCM}$ and $LGZE_{GLSZM}$ metrics imply that these texture features are related to the extent of the spatial correlation or uniformity of gray-levels within a region of interest, and to the emphasis on zones of low gray-levels, respectively. In terms of medical imaging of cancer, these features may be capturing the presence of large necrotic regions within the core of tumours. Such types of tumours could be rapidly increasing in size and could be more at risk to metastasize[37–39].

We also attempted in this study to improve the predictive power of our prediction models by combining radiomic variables with clinical data. The first step of our method is based on a fast mining of radiomic variables using logistic regression. Then, random forests[40] are used as a means to combine radiomic (continuous inputs) and clinical information (categorical inputs) into a single classifier. It would also be feasible to only use random forests to mine the radiomic variables, but our method is advantageous in terms of computation speed. Our results showed that the combination of clinical variables with the optimal radiomic variables via random forests had a positive impact on the prediction and prognostic assessment of locoregional recurrences and distant metastases, although with minimal impact in the latter case (Fig. 3, Table 1). As seen in Supplementary Table S4, this can be explained from the fact that the identified radiomic features are the strong and dominant variables in the model for distant metastases prediction.

In this work, we also performed a comprehensive comparison of the prediction/prognostic performance of radiomics versus clinical models and volumetric variables (Table 1). Metabolic tumour volume has already been shown to be an independent predictor of outcomes in head-and-neck cancer[41], but it was also suggested by Hatt *et al.*[42] that heterogeneity quantification via texture analysis may provide valuable complementary information to the tumour volume variable for volumes above 10 cm^3. In this study, 85% of the patients had a gross tumour volume greater than 10 cm^3 and we consequently found that radiomic models performed considerably better than tumour volume alone for the prediction of locoregional recurrences and distant metastases. On the other hand, clinical variables alone did not perform well on their own for distant metastases, but they had good performance for locoregional recurrences by outperforming radiomic models, thus suggesting that our radiomic models need to be improved (e.g., by including radiation dose metrics) to better model locoregional recurrences.

In terms of overall survival assessment, our results indicate that the tumour volume variable matched or outperformed all radiomic models thus far we developed or tested in this work, including a previously validated radiomic signature[21, 32]. For one, it is unsurprising that the original radiomic signature[21] did not perform better than tumour volume, as it can be verified that all its feature components are very strongly correlated with tumour volume: the Pearson linear coefficients between tumour volume and the four features of the signature[21] were calculated to be 0.62 (*Energy*), 0.80 (*Compactness*), 0.99 (GLN_{GLRLM}) and 0.94 (GLN_{GLRLM} *HLH*) using the whole set of 300 patients of this study, all with $p \ll 0.001$. On the other hand, all the features forming the other radiomic models developed in this work showed potential complementarity value to tumour volume: all the features of the revised version of the radiomic signature (Supplementary Section 2.6.2) had a Pearson linear coefficient lower than 0.5 except one (*Energy*), and all the variables forming the final radiomic models constructed in this work (italic fonts in Table 1, including those for locoregional recurrences and distant metastases) had linear coefficients lower than 0.40. This suggests that overall survival may be harder to model than specific tumour outcomes due to a larger number of confounding factors being involved, and it may thus be more prone to overfitting. As a

consequence, tumour volume may currently be a more robust and reproducible metric than imaging features for modeling this general outcome. In the end, the best global performance for overall survival was however obtained with clinical variables alone. This would emphasize that clinical data remains the important source of information to consider for the evaluation of overall survival, and that more work is required to understand how to adequately model this outcome using radiomic features.

Overall, our results demonstrate the possibility of using radiomics for decoding specific tumour phenotypes for the risk assessment of specific outcomes in head-and-neck cancer. As seen in Fig. 4, the output probability of occurrence of events of our prediction models allow to significantly separate (testing) patients into two locoregional recurrence risk groups and into three distant metastases risk groups. The clinical impact of these results could be substantial, as it could allow for a better personalization of head-and-neck cancer treatments. For example, higher radiation doses could be considered for patients at higher risks of locoregional recurrences. For distant metastases, the chemotherapy regimens could be strengthened for patients in the high risk group to reduce potential metastatic invasion, and lessened for patients in the low risk group to improve quality of life. These are hypothetical scenarios that, at the moment, are not ready to be implemented in the clinical environment, as our models first need to be constructed and validated on larger patient cohorts, and robust clinical trials are required to validate their benefits on patient survival.

In conclusion, we showed in this study that radiomics provide important prognostic information for the risk assessment of locoregional recurrences and distant metastases in head-and-neck cancer. The combination of radiomics data into clinically-integrated prediction models should allow to more comprehensively assess cancer risks and could improve how we adapt treatments for each patient. As the standardization efforts of radiomics analysis continue to rapidly progress[43, 44], we can envision the clinical implementation of radiomic-based decision-support systems. Full transparency on data and methods is the key for the progression of the field, and our research efforts need to include large-scale collaborations and reproducibility practices to accelerate the translation of radiomics into the clinical environment[45].

Methods

Data sets availability.   Our analysis was conducted on imaging and clinical data of a total of 300 H&N cancer patients from four different institutions who received radiation alone ($n = 48$, 16%) or chemo-radiation ($n = 252$, 84%) with curative intent as part of treatment management. The median follow-up period of all patients was 43 months (range: 6–112). The Institutional Review Boards of all participating institutions approved the study. Retrospective analyses were performed in accordance with the relevant guidelines and regulations as approved by the Research Ethics Committee of McGill University Health Center (Protocol Number: MM-JGH-CR15-50).

- The H&N1 data set consists of 92 head-and-neck squamous cell carcinoma (HNSCC) patients treated at Hôpital général juif (HGJ) de Montréal, QC, Canada. During the follow-up period, 12 patients developed a locoregional recurrence (13%), 16 patients developed distant metastases (17%) and 14 patients died (15%). This data set was used as part of the *training set* of this work.
- The H&N2 data set consists of 102 head-and-neck squamous cell carcinoma (HNSCC) patients treated at Centre hospitalier universitaire de Sherbrooke (CHUS), QC, Canada. During the follow-up period, 17 patients developed a locoregional recurrence (17%), 10 patients developed distant metastases (10%) and 18 patients died (18%). This data set was used as part of the *training set* of this work.
- The H&N3 data set consists of 41 head-and-neck squamous cell carcinoma (HNSCC) patients treated at Hôpital Maisonneuve-Rosemont (HMR) de Montréal, QC, Canada. During the follow-up period, 9 patients developed a locoregional recurrence (22%), 11 patients developed distant metastases (27%) and 19 patients died (46%). This data set was used as part of the *testing set* of this work.
- The H&N4 data set consists of 65 head-and-neck squamous cell carcinoma (HNSCC) patients treated at Centre hospitalier de l'Université de Montréal (CHUM), QC, Canada. During the follow-up period, 7 patients developed a locoregional recurrence (11%), 3 patients developed distant metastases (5%) and 5 patients died (8%). This data set was used as part of the *testing set* of this work.

All patients underwent FDG-PET/CT imaging scans within a median of 18 days (range: 6–66) before treatment. For 93 of the 300 patients (31%), the radiotherapy contours were directly drawn on the CT of the PET/CT scan by expert radiation oncologists and thereafter used for treatment planning. For 207 of the 300 patients (69%), the radiotherapy contours were drawn on a different CT scan dedicated to treatment planning and were propagated/resampled to the FDG-PET/CT scan reference frame using intensity-based free-form deformable registration with the software MIM (MIM software Inc., Cleveland, OH).

Further information specific to each patient cohort (e.g., treatment details) is presented in Supplementary Methods section 2.4 and Supplementary Tables S5–S8. Pre-treatment FDG-PET/CT imaging data, clinical data, radiotherapy contours (*RTstruct*) and MATLAB routines allowing to read imaging data and their associated region-of-interest (ROI) are made available for all patients on The Cancer Imaging Archive (TCIA)[46] at: http://www.cancerimagingarchive.net. The Research Ethics Committee of McGill University Health Center approved online publishing of clinical and imaging data following patient anonymisation.

Sample size and division of cohorts.   Patients with recurrent H&N cancer or with metastases at presentation, and patients receiving palliative treatment were excluded from the study. Patients that did not develop a locoregional recurrence or distant metastases during the follow-up period and that had a follow-up time smaller than 24 months were also excluded from the study. The four patient cohorts were then divided into two groups to create one combined training set (H&N1 and H&N2; $n = 194$) and one combined testing set (H&N3 and H&N4; $n = 106$). Bootstrap resampling and stratified random sub-sampling were always performed with patients from

the training set to estimate the relevant performance metrics of interest and to construct the final prediction models, and fully independent validation results were computed with patients from the testing set. This precise type of division of patient cohorts allowed to: I) Train on a combined set of different cohorts to allow the models to take into account some institutional variability; II) Reduce the number of testing results reported; III) Create a training set size to testing set size ratio of approximately 2:1; and IV) Conduct partition sampling such that the proportions of occurrence of events (locoregional recurrences, distant metastases) are approximately the same in the training and testing sets.

Extraction of radiomic features. Starting from the original FDG-PET/CT imaging data and associated radiotherapy contours in DICOM format, the complete set of data was read and transferred into MATLAB (MathWorks, Natick, MA) format using in-house routines. PET images were converted to standard uptake value (SUV) maps and CT images were kept in raw Hounsfield Unit (HU) format. In this work, we then extracted a total of 1615 radiomic features for both the PET and CT images from the tumour region defined by the "$GTV_{primary} + GTV_{lymph\ nodes}$" contours as delineated by the radiation oncologists of each institution. These features can be divided into three different groups: I) 10 first-order statistics features (intensity); II) 5 morphological features (shape); and III) 40 texture features each computed using 40 different combinations of extraction parameters.

Intensity features are computed from histograms ($n_{bins} = 100$) of the intensity distribution of the ROI. The features extracted in this work were the *variance*, the *skewness*, the *kurtosis*, *SUVmax*, *SUVpeak*, *SUVmean*, the *area under the curve of the cumulative SUV-volume histogram*[47], the *total lesion glycolysis*, the *percentage of inactive volume* and the *generalized effective total uptake*[48]. Shape features describe geometrical aspects of the ROI. The features extracted in this work were the *volume*, the *size* (maximum tumour diameter), the *solidity*, the *eccentricity* and the *compactness*.

Texture features measure intratumoural heterogeneity by quantitatively describing the spatial distributions of the different intensities within the ROI. In this work, 9 features from the Gray-Level Co-occurrence Matrix (GLCM)[49], 13 features from the Gray-Level Run-Length Matrix (GLRLM)[50–52], 13 features from the Gray-Level Size Zone Matrix (GLSZM)[50–53] and 5 features from the Neighbourhood Gray-Tone Difference Matrix (NGTDM)[54] were computed. All texture matrices were constructed using 3D analysis/26-voxel connectivity of the tumour region resampled to a defined isotropic voxel size. For each of the four texture types, only one matrix was computed per scan by simultaneously taking into account the neighbouring properties of voxels in the 13 directions of 3D space. However, the 6 voxels at a distance of 1 voxel, the 12 voxels at a distance of $\sqrt{2}$ voxels, and the 8 voxels at a distance of $\sqrt{3}$ voxels around center voxels were treated differently in the calculation of the matrices to take into account discretization length differences.

All 40 texture features from the ROI of both PET and CT volumes were extracted using all possible combinations (40) of the following parameters:

- Isotropic voxel size (5): Voxel sizes of 1 mm, 2 mm, 3 mm, 4 mm and 5 mm.
- Quantization algorithm (2): *Equal-probability* (equalization of intensity histogram) and *Uniform* (uniform division of intensity range) quantization algorithms with fixed numbers of gray levels.
- Number of gray levels (4): Fixed numbers of gray levels of 8, 16, 32 and 64 in the quantized ROI.

Detailed description with supplementary references and methodology used to extract all radiomic features is further provided in Supplementary Methods section 2.5.

Construction of radiomic models. The construction of prediction models from the total set of radiomic features for each of the three initial feature sets (I: PET features; II: CT features; and III: PET and CT features) and three H&N cancer outcomes was performed from the defined training set of this work (H&N1 and H&N2 cohorts; $n = 194$) using the methodology developed in the work of Vallières *et al.*[29]. The process of combining radiomic features into a multivariable model was achieved using the logistic regression utilities of the software DREES[55]. The general workflow is presented in Supplementary Fig. S5.

First, feature set reduction was performed for each of the initial feature sets via a stepwise forward feature selection scheme in order to create reduced feature sets containing 25 different features balanced between predictive power (Spearman's rank correlation) and non-redundancy (maximal information coefficient[56]). This procedure was carried out using the *Gain* equation[29], which is detailed in Supplementary Methods section 2.2.2.

From the reduced feature sets, stepwise forward feature selection was then performed by maximizing the 0.632+ bootstrap AUC[57, 58] (100 samples). For each of the reduced feature sets, combinations of features were selected for model orders (i.e., number of combined variables) of 1 to 10. The whole feature selection process is detailed in Supplementary Methods section 2.2.3 and in Supplementary Fig. S6.

Once optimal combinations of features were identified for model orders of 1 to 10 for all feature sets, prediction performance was estimated using the 0.632+ bootstrap AUC (100 samples). By inspecting the prediction estimates shown in Supplementary Fig. S2, a single combination of features (i.e., model order) potentially possessing the best parsimonious properties was then chosen for each feature set and each outcome (identified as circles in Supplementary Fig. S2). The final logistic regression coefficients of these selected radiomic prediction models (3 feature sets × 3 outcomes) were then found by averaging all coefficients computed from another set of 100 bootstrap samples. These prediction models in their final form were thereafter directly tested in the defined testing set of this work (H&N3 and H&N4 cohorts; $n = 106$).

Combination of radiomic and clinical variables. The construction of prediction models combining radiomic and clinical variables was also carried out using the training set consisting of the combination of 194

patients from the H&N1 and H&N2 cohorts (Fig. 2). First, random forest classifiers[40] containing only the following clinical variables were constructed for the LR, DM and OS outcomes: I) *Age*; II) *H&N type* (oropharynx, hypopharynx, nasopharynx or larynx); and III) Tumour stage. The selection of the following best groups of tumour stage variables to be incorporated into the clinical-only random forest classifiers was performed: I) *T-Stage*; II) *N-Stage*; III) *T-Stage* and *N-Stage*; and IV) *TNM-Stage*. Estimation of prediction performance for feature selection and subsequent random forest training was performed in the training set using stratified random sub-sampling and imbalance adjustments to account for the disproportion between the occurrence and non-occurrence of events. Overall, the following staging variables were estimated to possess the highest prediction performance in the training set when combined into random forest classifiers with *Age* and *H&N type*, and were thereafter used for the rest of the work accordingly for each outcome: (I) *T-Stage* and *N-Stage* for LR prediction; (II) *N-Stage* for DM prediction; and (III) *T-Stage* and *N-Stage* for OS prediction. Finally, the variables of the previously identified radiomic prediction models (3 feature sets × 3 outcomes) were incorporated with the corresponding clinical variables identified for each outcome via the separate construction of final random forest classifiers.

Imbalance-adjustment strategy. To obtain models with predictive power equally balanced between the prediction of occurrence and non-occurrence of events, an imbalance adjustment strategy adapted from the work of Schiller *et al.*[30] was used in this work (Supplementary Fig. S4). Imbalance adjustments become an essential part of the training process when the proportion of instances (e.g., patients) of a given class (e.g., occurrence of an event) is much lower than the proportion of instances of the other class (e.g., non-occurrence of an event). This is the case in this work for the proportion of locoregional recurrences, distant metastases and death events in the training and testing sets (Fig. 2).

In this work, every time a different bootstrap sample was drawn from the training set to construct a logistic regression or a random forest classifier, a different ensemble of multiple balanced classifiers was used in the training process instead of using only one unbalanced classifier. The ensemble classifier is composed of a number of $P = \lceil N^-/N^+ \rceil$ partitions, where N^- is the number of instances from the majority class and N^+ the number of instances from the minority class in a particular bootstrap sample. The N^+ instances are copied and used in every partition, and the N^- instances are randomly sampled without replacement in the P partitions such that the number of instances of the majority class is either $\lfloor N^-/P \rfloor$ or $\lceil N^-/P \rceil$ in each partition. For example, for $N^- = 168$ and $N^+ = 32$, five partitions would be created: two would contain 33 instances from the majority class, three would contain 34 instances from the majority class, and all would contain the 32 instances from the minority class.

For the logistic regression training process, a different classifier (i.e., different coefficients) is then trained for each of the created partitions, and the final ensemble classifier consists in the average of the corresponding coefficients from each partition. For random forest training, each partition is used to create a decision tree to be appended to a final forest instead of creating only one tree per bootstrap sample.

Random forest training. The process of random forest training inherently uses bootstrapping in order to train the multiple decision trees of the forest. Conventionally, one different decision tree is trained for each bootstrap sample. In this work, we used 100 bootstrap samples to train each random forest constructed from the training set (H&N1 and H&N2 cohorts; $n = 194$). For each bootstrap sample, the imbalance-adjustment strategy detailed above was used such that each bootstrap sample produced multiple decision trees (one per partition) to be appended to a random forest. Therefore, the final number of decision trees per random forest was dependent on the actual proportion of events in each bootstrap sample for each outcome studied. The three final random forest models developed in this work (italic fonts in Table 1, Supplementary Table S4) were constructed using 582, 661 and 518 decision trees for LR, DM and OS, respectively.

In addition to the imbalance-adjustment strategy adopted in this work, under/oversampling of the instances in each partition of an ensemble was used to further correct for data imbalance in the random forest training process. Under/oversampling weights of the minority class of 0.5 to 2 with increments of 0.1 were tested in this work. Stratified random sub-sampling was used to estimate the optimal weight for a given training process (and also to estimate the optimal clinical staging variables to be used) in terms of the maximal average AUC, a process randomly separating the training set of this work into multiple sub-training and sub-testing sets ($n = 10$) with 2:1 size ratio and equal proportion of events. The final random forest models developed in this work (italic fonts in Table 1, Supplementary Table S4) used oversampling weights of 1.4, 1.6 and 1.7 (in conjunction with the previously described imbalance-adjustment strategy) to train the decision trees of the forests for LR, DM and OS, respectively. The overall random forest training process is pictured in Supplementary Fig. S7.

Calculation of prediction/prognostic performance metrics. In this work, all prediction models were fully trained in the defined training set of this work (H&N1 and H&N2 cohorts; $n = 194$). Models were then independently tested in the defined testing set of this work (H&N3 and H&N4 cohorts; $n = 106$). Prediction performance was assessed using ROC metrics in terms of the AUC, sensitivity, specificity and accuracy of classification of binary clinical endpoints (locoregional recurrences, distant metastases, deaths). Prognostic performance in terms of time estimates of clinical endpoints was assessed using the concordance-index (CI)[33] and the *p*-value obtained from Kaplan-Meier analysis using the log-rank test between two risk groups.

For prediction performance, the output of the linear combination of features of logistic regression models was directly used to calculate the AUC with binary outcome data. The multivariable response was then transformed into the posterior probability of occurrence of an event using a logit transform to calculate the sensitivity, specificity and accuracy of prediction using a probability threshold of 0.5. Similarly, the output probability of occurrence

of an event of random forest models was directly used to calculate the AUC with binary outcome data, and an output probability threshold of 0.5 was also used to calculate the remaining metrics.

For prognostic performance, the output of the linear combination of features of cox proportional hazard regression models was directly used to calculate the CI with time-to-event data (time elapsed between the date the treatment ended and the date when an event occurred or the date of last-follow-up). The median of the output of the cox regression models found in the training set was used to separate the patients of the testing set into two risk groups for Kaplan-Meier analysis. For random forests, the output probability of occurrence of an event was directly used to calculate the CI with time-to-event data, and a probability threshold of 0.5 was used to separate the patients of the testing set into two risk groups (or $\frac{1}{3}$ and $\frac{2}{3}$ for three risk groups) for Kaplan-Meier analysis.

Code availability. All software code including a single organized script allowing to run the experiments used to produce all the results presented in this work is freely shared under the GNU General Public License on the GitHub website at: https://github.com/mvallieres/radiomics.

References

1. Meric-Bernstam, F., Farhangfar, C., Mendelsohn, J. & Mills, G. B. Building a personalized medicine infrastructure at a major cancer center. *J. Clin. Oncol.* **31**, 1849–1857 (2013).
2. Renfro, L. A., An, M.-W. & Mandrekar, S. J. Precision oncology: a new era of cancer clinical trials. *Cancer Lett.* **387**, 121–126 (2016).
3. Lambin, P. *et al.* Rapid Learning health care in oncology – an approach towards decision support systems enabling customised radiotherapy. Ra-diother. *Oncol.* **109**, 159–164 (2013).
4. Shrager, J. & Tenenbaum, J. M. Rapid learning for precision oncology. *Nat. Rev. Clin. Oncol.* **11**, 109–118 (2014).
5. Weitzel, J. N., Blazer, K. R., MacDonald, D. J., Culver, J. O. & Offit, K. Genetics, genomics, and cancer risk assessment. *CA Cancer J. Clin.* **61**, 327–359 (2011).
6. Garraway, L. A., Verweij, J. & Ballman, K. V. Precision oncology: an overview. *J. Clin. Oncol.* **31**, 1803–1805 (2013).
7. El Naqa, I. Biomedical informatics and panomics for evidence-based radi-ation therapy. *WIREs Data Mining Knowl. Discov.* **4**, 327–340 (2014).
8. Ebrahim, A. *et al.* Multi-omic data integration enables discovery of hidden biological regularities. *Nat. Commun.* **7**, 13091 (2016).
9. Fisher, R., Pusztai, L. & Swanton, C. Cancer heterogeneity: implications for targeted therapeutics. *Br. J. Cancer* **108**, 479–485 (2013).
10. Heppner, G. H. & Miller, B. E. Tumor heterogeneity: biological impli-cations and therapeutic consequences. *Cancer Metastasis Rev.* **2**, 5–23 (1983).
11. Fidler, I. J. Critical factors in the biology of human cancer metastasis: twenty-eighth G.H.A. Clowes memorial award lecture. *Cancer Res.* **50**, 6130–6138 (1990).
12. Yokota, J. Tumor progression and metastasis. *Carcinogenesis* **21**, 497–503 (2000).
13. Campbell, P. J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, 1109–1113 (2010).
14. Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: images are more than pictures, they are data. *Radiology* **278**, 563–577 (2016).
15. El Naqa, I. *et al.* Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. *Pattern Recognit.* **42**, 1162–1171 (2009).
16. Gillies, R. J., Anderson, A. R., Gatenby, R. A. & Morse, D. L. The biology underlying molecular imaging in oncology: from genome to anatome and back again. *Clin. Radiol.* **65**, 517–521 (2010).
17. Lambin, P. *et al.* Radiomics: extracting more information from medical im-ages using advanced feature analysis. *Eur. J. Cancer* **48**, 441–446 (2012).
18. Kumar, V. *et al.* Radiomics: the process and the challenges. *Magn. Reson. Imaging* **30**, 1234–1248 (2012).
19. Segal, E. *et al.* Decoding global gene expression programs in liver cancer by noninvasive imaging. *Nat. Biotechnol.* **25**, 675–680 (2007).
20. Diehn, M. *et al.* Identification of noninvasive imaging surrogates for brain tumor gene-expression modules. *Proc. Natl. Acad. Sci. USA* **105**, 5213–5218 (2008).
21. Aerts, H. J. W. L. *et al.* Decoding tumour phenotype by noninvasive imag-ing using a quantitative radiomics approach. *Nat. Commun.* **5**, 4006 (2014).
22. Hatt, M. *et al.* Characterization of PET/CT images using texture analysis: the past, the present any future? *Eur. J. Nucl. Med. Mol. Imaging* 1–15 (2016).
23. Yip, S. S. F. & Aerts, H. J. W. L. Applications and limitations of radiomics. *Phys. Med. Biol.* **61**, R150–R166 (2016).
24. Lambin, P. *et al.* Predicting outcomes in radiation oncology–multifactorial decision support systems. *Nat. Rev. Clin. Oncol.* **10**, 27–40 (2013).
25. Ferlito, A., Shaha, A. R., Silver, C. E., Rinaldo, A. & Mondin, V. Incidence and sites of distant metastases from head and neck cancer. *ORL* **63**, 202–207 (2001).
26. Baxi, S. S. *et al.* Causes of death in long-term survivors of head and neck cancer. *Cancer* **120**, 1507–1513 (2014).
27. Wong, A. J., Kanwar, A., Mohamed, A. S. & Fuller, C. D. Radiomics in head and neck cancer: from exploration to application. *Transl. Cancer Res.* **5**, 371–382 (2016).
28. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a prac-tical and powerful approach to multiple testing. J. R. Statist. *Soc. B* **57**, 289–300 (1995).
29. Vallières, M., Freeman, C. R., Skamene, S. R. & El Naqa, I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Phys. Med. Biol.* **60**, 5471–5496 (2015).
30. Schiller, T. W., Chen, Y., El Naqa, I. & Deasy, J. O. Modeling radiation-induced lung injury risk with an ensemble of support vector machines. *Neurocomputing* **73**, 1861–1867 (2010).
31. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).
32. Leijenaar, R. T. H. *et al.* External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma. *Acta Oncol.* **54**, 1423–1429 (2015).
33. Harrell, F. E. J., Lee, K. L. & Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15**, 361–387 (1996).
34. Fakhry, C. *et al.* Improved survival of patients with human papillomavirus-positive head and neck squamous cell carcinoma in a prospective clinical trial. *J. Natl. Cancer Inst.* **100**, 261–269 (2008).
35. Ang, K. K. *et al.* Human papillomavirus and survival of patients with oropharyngeal cancer. *N. Engl. J. Med.* **363**, 24–35 (2010).
36. Cheng, N.-M. *et al.* Zone-size nonuniformity of 18F-FDG PET regional textural features predicts survival in patients with oropharyngeal cancer. *Eur. J. Nucl. Med. Mol. Imaging.* **42**, 419–428 (2014).
37. Vakkila, J. & Lotze, M. T. Inflammation and necrosis promote tumour growth. *Nat. Rev. Immunol.* **4**, 641–648 (2004).
38. Proskuryakov, S. Y. & Gabai, V. L. Mechanisms of tumor cell necrosis. *Curr. Pharm. Des.* **16**, 56–68 (2010).

39. Ahn, S.-H. *et al*. Necrotic cells influence migration and invasion of glioblas-toma via NF-κB/AP-1-mediated IL-8 regulation. *Sci. Rep.* **6**, 24552 (2016).
40. Breiman, L. Random forests. *Machine Learning* **45**, 5–32 (2001).
41. Tang, C. *et al*. Validation that metabolic tumor volume predicts outcome in head-and-neck cancer. *Int. J. Radiat. Oncol. Biol. Phys.* **83**, 1514–1520 (2012).
42. Hatt, M. *et al*. 18F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and func-tional tumor volume in a multi-cancer site patient cohort. *J. Nucl. Med.* **56**, 38–44 (2015).
43. Nyflot, M. J. *et al*. Quantitative radiomics: impact of stochastic effects on textural feature analysis implies the need for standards. *J. Med. Imaging* **2**, 041002 (2015).
44. Zhao, B. *et al*. Reproducibility of radiomics for deciphering tumor pheno-type with imaging. *Sci. Rep.* **6**, 23428 (2016).
45. Ioannidis, J. P. A. How to make more published research true. *PLoS Med.* **11**, e1001747 (2014).
46. Clark, K. *et al*. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* **26**, 1045–1057 (2013).
47. Van Velden, F. H. P. *et al*. Evaluation of a cumulative SUV-volume his-togram method for parameterizing heterogeneous intratumoural FDG up-take in non-small cell lung cancer PET studies. *Eur. J. Nucl. Med. Mol. Imaging* **38**, 1636–1647 (2011).
48. Rahmim, A. *et al*. A novel metric for quantification of homogeneous and heterogeneous tumors in PET for enhanced clinical outcome prediction. *Phys. Med. Biol.* **61**, 227 (2016).
49. Haralick, R. M., Shanmugam, K. & Dinstein, I. Textural features for im-age classification. *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-3**, 610–621 (1973).
50. Galloway, M. M. Texture analysis using gray level run lengths. *Computer Graphics and Image Processing* **4**, 172–179 (1975).
51. Chu, A., Sehgal, C. M. & Greenleaf, J. F. Use of gray value distribution of run lengths for texture analysis. *Pattern Recognition Letters* **11**, 415–419 (1990).
52. Dasarathy, B. V. & Holder, E. B. Image characterizations based on joint gray level–run length distributions. *Pattern Recognition Letters* **12**, 497–502 (1991).
53. Thibault, G. *et al. Texture indexes and gray level size zone matrix: appli-cation to cell nuclei classification in Proceedings of the Pattern Recognition and Information Processing 2009*. International Conference on Pattern Recognition and Information Processing (PRIP '09), 140–145 (Minsk, Belarus, 2009).
54. Amadasun, M. & King, R. Textural features corresponding to textural properties. *IEEE Transactions on Systems, Man, and Cybernetics* **19**, 1264–1274 (1989).
55. El Naqa, I. *et al*. Dose response explorer: an integrated open-source tool for exploring and modelling radiotherapy dose-volume outcome relationships. *Phys. Med. Biol.* **51**, 5719–5735 (2006).
56. Reshef, D. N. *et al*. Detecting novel associations in large data sets. *Science* **334**, 1518–1524 (2011).
57. Efron, B. & Tibshirani, R. Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association* **92**, 548–560 (1997).
58. Sahiner, B., Chan, H.-P. & Hadjiiski, L. Classifier performance prediction for computer-aided diagnosis using a limited dataset. *Med. Phys.* **35**, 1559–1570 (2008).

Acknowledgements

Author Contributions

M.V., J.S and I.E.N. conceived the project, analyzed the data and wrote the paper. E.K.-R., L.J.P., X.L., C.F., N.K., F.N., C.-S.W. and K.S. collected and curated the data. H.J.W.L.A. provided expert knowledge and helped in the computation of the radiomic signature. All authors edited the manuscript.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-10371-5

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.