Article

Prediction of radiation-induced hypothyroidism using radiomic data analysis does not show superiority over standard normal tissue complication models

Urszula Smyczynska [1,‡], Szymon Grabia [1,‡], Zuzanna Nowicka [1], Anna Papis-Ubych [2], Robert Bibik [3], Tomasz Latusek [4], Tomasz Rutkowski [5], Jacek Fijuth [2,6], Wojciech Fendler [1,7]* and Bartłomiej Tomasik [1,7]

1 Department of Biostatistics and Translational Medicine, Medical University of Lodz, 92-215 Lodz, Poland
2 Department of Radiotherapy, N. Copernicus Memorial Regional Specialist Hospital, 93-513 Lodz, Poland
3 Department of Radiation Oncology, Oncology Center of Radom, 26-600 Radom, Poland
4 Radiotherapy Department, Maria Sklodowska-Curie National Research Institute of Oncology (MSCNRIO)—branch in Gliwice, 44-101 Gliwice
5 I Radiation and Clinical Oncology Department, Maria Sklodowska-Curie National Research Institute of Oncology (MSCNRIO)—branch in Gliwice, 44-101 Gliwice, Poland
6 Department of Radiotherapy, Chair of Oncology, Medical University of Lodz, 93-509 Lodz, Poland
7 Department of Radiation Oncology, Dana-Farber Cancer Institute, Boston, MA 02115, USA
* Correspondence: wojciech_fendler@dfci.harvard.edu
‡ These authors contributed equally to this work.

Abstract: State-of-art normal tissue complication probability (NTCP) models do not take into account more complex individual anatomical variations, which can be objectively quantitated and compared in radiomic analysis. The goal of this project was development of radiomic NTCP model for radiation-induced hypothyroidism (RIHT) using imaging biomarkers (radiomics). We gathered CT images and clinical data from 98 patients, who underwent intensity-modulated radiation therapy (IMRT) for head and neck cancers with a planned total dose of 70.0 Gy (33–35 fractions). During the 28-month (median) follow-up 27 patients (28%) developed RIHT. For each patient, we extracted 1316 radiomic features from original and transformed images using manually contoured thyroid masks. Creating models based on clinical, radiomic features or a combination thereof, we considered 3 variants of data preprocessing. Based on their performance metrics (sensitivity, specificity), we picked best models for each variant ((0.8, 0.96), (0.9, 0.93), (0.9, 0.89) variant-wise) and compared them with external NTCP models ((0.82, 0.88), (0.82, 0.88), (0.76, 0.91)). Our models reach accuracy comparable with or better than previously presented non-radiomic NTCP models. The benefit of our approach is obtaining the RIHT predictions before treatment planning to adjust IMRT plan to avoid the thyroid region in most susceptible patients.

Keywords: radiomics, radiation-induced hypothyroidism, NTCP, head and neck cancer

1. Introduction

Patients with head and neck cancer (HNC) treated with radiation therapy (RT) may experience multiple adverse normal tissue effects [1], including hypothyroidism. Radiation-induced hypothyroidism (RIHT) has been reported in 25-65% of patients and typically develops during the first 2 years from RT completion. Since the symptoms of RIHT are non-specific and insufficient levels of thyroid hormones not only negatively impact patients' quality of life [2], but also mortality [3,4], adequate identification of patients at risk of this effect is of paramount importance [5–7].

Clinical and dosimetric parameters are typically used in normal tissue complication probability (NTCP) models to predict RIHT [1]. Recently, we [7] and others [6] have applied published NTCP models in homogenous cohorts of patients with oropharyngeal cancer (OPC). We concluded that two models based on thyroid mean dose and volume, published by Rønjom et al. [8] and by Boomsma et al. [9], performed best in terms of accuracy (84 to 87%), highlighting the feasibility of dose-response models to predict RIHT and their potential utility in the clinical setting.

Radiomics is a relatively new discipline that aims at deriving biomarkers from medical images [10]. These biomarkers are features describing shape, intensity or texture of specific region(s) of interest (ROI), typical an organ or lesion. Radiomics emerged in the field of cancer studies where large number of medical images are generated even by routine diagnostic process. In this context radiomics can supplement standard radiologic assessment with localized, quantitative information on interesting structures with low cost and without any additional burden or discomfort to patients. In recent few years, radiomics researchers community took considerable efforts to standardize methodology aiming at translatability of radiomic studies that resulted in publication of the first reference manual for image biomarker studies [11].

Radiomic features were shown to reflect cancer biology in terms of histologic type [12], overall and progression-free survival [12–14], probability of recurrence [15], activity of biological pathways [12,16], human papillomavirus infection [17], probability lymph node [18] or distant metastases [19], CD8 cells infiltration [20]. Some studies extended radiomic analysis outside tumor volume, including also peritumoral regions for prediction of patients' survival [21] or adjuvant chemotherapy [22]. Similarly, lymph nodes were sometimes considered as ROI in studies that aimed at identifying lymph nodes metastases [23,24]. Some radiomic NTCP models based on computed tomography (CT) were developed, for instance for prediction of xerostomia after head-and-neck cancer radiotherapy [25,26], radiation-induced pneumonitis [27]. Alternatively, radiomic features for radiation-induced NTCP models could be calculated in 3D dose distribution that was successfully done for cervical cancer [28], prostate cancer [29] and lung cancer [30]. Given the very good performance of some NTCP models in predicting RIHT in patients with OPC, the question whether there would be an added benefit of an in depth radiomic analysis was unresolved and thus we decided to apply CT-based radiomics in this group of patients and compare predictive performance with the available models based on dosimetric and clinical features.

2. Materials and Methods

2.1. Patients and CT images

The studied cohort was a subgroup of patients included in previous study on validation of NTCP models for radiation-induced hypothyroidism by Nowicka et al. [7], recruited between 01.05.2016 and 31.12.2018 and followed up until 02.2020. In addition to already collected clinical data, for this study we retrieved CT images from radiotherapy planning with thyroid glands retrospectively contoured by two experienced radiation oncologists according to the guidelines for organs at risk (OARs), as described previously. Patients for whom the CT image was unavailable, thyroid contour was missing or thyroid region contained CT artifacts, were excluded. Finally, among 108 patients recruited in 3 centers 98 had complete data and were included in this study (38 from center A - Copernicus Regional Specialist Hospital in Lodz, Poland; 12 from center B - Radom Oncology Centre and Maria Sklodowska-Curie National Research Institute of Oncology, Radom, Poland and 48 from center C - Radiotherapy Department, Maria Sklodowska-Curie National Research Institute of Oncology — branch in Gliwice, Poland); selection of cases is summarized in Fig. 1A.

All patients underwent intensity-modulated radiation therapy (IMRT) of OPC before which their thyroid function was normal. After RT, they were monitored for a median of 28 months on average (median). During this follow-up period, 27 patients developed hypothyroidism. To avoid bias, laboratory assessment, contouring, outcome assessments, statistical and radiomic analysis were performed by independent researchers. Details of treatment protocol were described previously [7].

For all 98 patients, CT images paired with RT plans and anatomical structure contouring files were retrieved from PACS of treating centers. All images were stored in DICOM format in CT and RTSTRUCT modalities; however, some differences in acquisition protocol were identified. First, each center used a different CT scanner (center

A: Somatom Sensation Open, Siemens; center B: Optima CT580 RT, GE Healthcare; center C: SOMATOM Definition AS and Somatom Sensation Open, Siemens). In each center, slice thickness and pixel spacing were selected by radiotherapist and radiologist so that images were sufficient for treatment planning. Summary of selected settings in each center is presented in Table 1 and anonymized raw patients data are in Table S1 in supplementary material.

The study was approved by the Bioethics Committee of the Medical University of Lodz (KE/7/10, RNN/65/18).

Table 1. Patient, disease and CT images characteristics.

		Center A (n=38)		Center B (n=12)		Center C (n=48)	
RIHT		NO	YES	NO	YES	NO	YES
Sex	Female	7	7	1	1	5	2
	Male	21	3	8	2	29	12
Age	Median	62.0	60.0	61.0	57.0	57.5	58.0
	IQR	57.0-66.2	56.8-61.8	60.0-68.0	55.5-60.5	53.0-62.0	52.2-66.5
Stage	I-II	6	1	2	0	12	1
	III-IV	22	9	7	3	22	13
Mean thyroid dose, D_{mean} (Gy)	Median	54.8	57	55.2	57.3	47.2	52.5
	IQR	51.9-56.3	52.7-59.3	53.5-56.3	56.2-58.4	43.8-49.5	49.7-56.2
Minimal thyroid dose, D_{min} (Gy)	Median	42.5	46.5	43	51.8	30.9	44.3
	IQR	29.1-46.6	43.3-47.5	41.1-48.2	50.1-54.3	24.6-39.0	32.7-46.9
Median thyroid dose, D50 (Gy)	Median	55.0	55.5	54.5	58.5	47.0	52.2
	IQR	53.7-56.3	52.9-58.7	53.9-54.9	57.2-59.0	43.8-50.4	50.5-56.4
Maximal thyroid dose, D_{max} (Gy)	Median	62.5	69.4	61.7	61.6	60.2	65.1
	IQR	57.6-70.1	63.3-71.9	60.8-62.1	59.6-61.8	52.6-68.0	53.6-72.2
Thyroid volume (ml)	Median	21.7	11.8	29	12.6	19.1	10.6
	IQR	19.0-32.9	7.7-13.9	21.7-37.4	10.6-14.0	14.6-27.6	8.3-13.3
Baseline fT4 (pg/mL)	Median	6.5	6.1	9.3	8.2	7.2	8.1
	IQR	5.3-7.4	5.1-7.6	8.0-10.1	7.7-10.7	6.3-8.4	7.9-9.9
Baseline TSH (mIU/L)	Median	0.5	1.3	0.7	0.4	0.7	1.1
	IQR	0.3-0.8	0.8-1.7	0.6-1.2	0.4-0.7	0.5-1.5	0.6-1.2
Mean pituitary dose (Gy)	Median	4.0	3.8	4.0	3.8	3.8	3.7
	IQR	3.0-4.5	3.0-5.3	3.2-4.8	3.6-3.8	3.0-4.4	3.0-4.8
Time to follow-up (months)	Median	29.5	15	22	13	38	19
	IQR	26.0-37.2	14.0-15.8	21.0-24.0	12.0-13.5	31.2-41.0	16.0-21.0
Pixel spacing (mm^2)	0.98x0.98	25	9	0	0	26	11
	1.07x1.07	1	0	0	0	3	3
	1.09x1.09	0	0	0	0	1	0
	1.11x1.11	0	0	0	0	1	0
	1.13x1.13	0	0	0	0	1	0
	1.17x1.17	0	0	0	0	1	0
	1.27x1.27	1	0	9	3	1	0
	1.56x1.56	1	1	0	0	0	0
Slice thickness (mm)	1.5	1	0	0	0	0	0
	2	1	1	0	0	2	1
	2.5	0	0	9	3	0	0
	3	24	9	0	0	26	9
	4	0	0	0	0	6	4
	5	2	0	0	0	0	0

2.2. Image preprocessing and radiomic features calculation

Due to diversity of pixel spacing and slice thicknesses, all images and thyroid masks (generated from contours using dcmrtstruct2nii library [31]) were resampled to 1x1x1mm^3 isotropic voxels. Then we used PyRadiomics v3.0 [32] to calculate radiomic features that in the applied version of the library included: 14 shape features, 18 first order statistics, 24 gray level cooccurrence matrix-based (GLCM) features, 16 gray

level run length matrix-based (GLRLM) features, 16 gray level size zone matrix-based (GLSZM) features, 5 neighborhood gray tone difference matrix-based (NGTDM) features, 14 gray level dependence matrix-based (GLDM) features (full list of features in Appendix A). Radiomic features were extracted from original images and from their filtered versions, applying all filters available in PyRadiomics feature extractor: square, square root, logarithm, gradient, exponential and 8 wavelet decomposition filters. Default values of PyRadiomics feature extractor settings were used, including filtration parameters. Resegmentation was not applied. In total, 1316 features were calculated for each image (14 shapes features extracted from image mask, 93 intensity-based features for original images and 13 filtered ones).

The processing of images and NTCP models derivation are summarized in Fig. 1B, while raw radiomic features values are reported in Table S2.

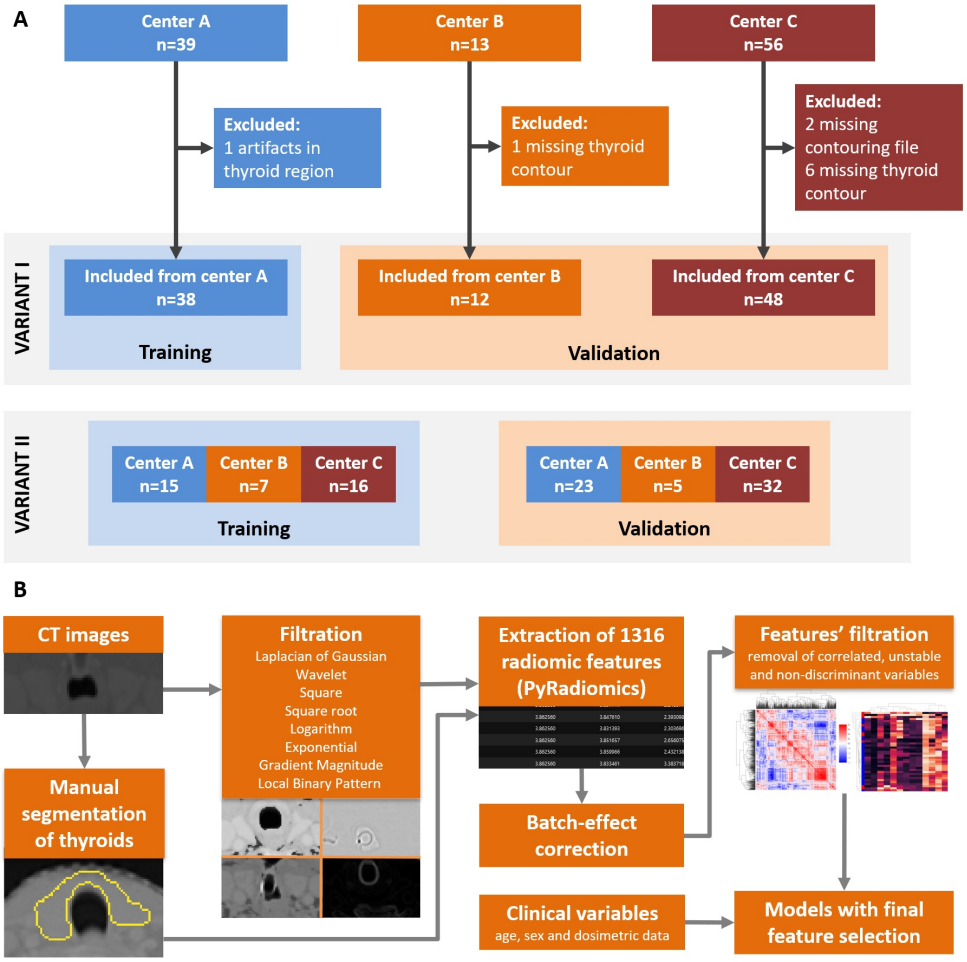

Figure 1. Summary of patients' selection and data subsets (A). Outline of the study from image segmentation to NTPC model derivation (B).

2.3. *Stability assessment*

The purpose of stability analysis was to identify features that were unaffected (or affected only slightly) by minor differenced in contouring of thyroid glands. As the segmentations by multiple radiotherapists or radiologists were not available, we decided to perform a simulation study. We modelled inaccuracies of contouring as affine deformations of thyroid mask, consisting of:

- 3 translations by up to 1 mm in either direction along each of the 3 main axes,

- 3 rotations by up to 2° in either direction around each of the 3 main axes,
- 3 zooms by up to 2% of either dimension along each of the 3 main axes.

We generated 100 such transformations by randomly choosing order and parameters of the above 9 operations. Then, we applied every transformation to all thyroid masks in our image data set and used the transformed masks to calculate new sets of radiomic features.

Inter-class correlation coefficient (ICC) served as a measure of agreement between original features (calculated using unchanged mask) and those calculated using each of 100 transformed masks. Later, 100 ICC values obtained for different transformed masks were averaged to give single measure of stability for each feature. Stability was classified as excellent when mean ICC exceeded 0.9, good (ICC between 0.75 and 0.9), moderate (ICC between 0.5 and 0.75) or poor (ICC not exceeding 0.5).

2.4. Feature preprocessing and splitting data set

Raw feature values were firstly scaled to a [-1,1] range, then subjected to Yeo–Johnson transformation [33] and scaled to [-1,1] range once again to facilitate machine learning model derivation. As average values and distributions of features values varied across the samples (Fig. S1) we normalized them dividing each value by sample mean. While performing normalization, we followed the reasoning common for other high-throughput studies (transcriptomics, proteomics) that usually only minority of features should be expected to differ between conditions, thus the distribution of all feature values should not differ significantly between samples.

We observed that even after normalization, we could still observe batch effect related to clinics that performed CT scans (Fig. S2). Thus, to account for this and check the impact of batch on NTCP model performance, we decided to implement different variants of data preprocessing and splitting:

- Variant I without batch effect correction, later referred to as Variant Ia: center A as training set, centers B and C as validation set as shown in Fig. 1A.
- Variant I with batch effect removed by ComBat [34], referred to also as Variant Ib: center A as training set, centers B and C as validation set.
- Variant II: no batch effect removal, but data from three centers were joined and subsequently divided into training and validation sets (Fig. 1A). Training set contained 38 cases to match the size of center A data set so that both variants of splitting data are comparable.

2.5. Feature filtration

Due to the large number of features, we decided to filter them before derivation of NTCP models. First, only features with excellent or good stability were considered as candidates for NTCP model predictors. We used ICC classes from center A for filtration in Variant I and, analogously ICC classes from training set for Variant II.

Then, in order to retain only features that actually differentiate patient with and without RIHT in follow-up, t-tests were performed to compare values of each stable feature between these group. Benjamini-Hochberg correction was applied to p-values to control false discovery rate (FDR) in this large set of comparisons. Next, hierarchical clustering of features was performed with 1 – r (correlation coefficient) as distance measure and average linkage to identify groups of highly correlated features. These groups of features were extracted by setting a threshold of 0.3 to distances in the dendrogram. From each such group we retained a single feature with lowest FDR-corrected p-value (later referred to as FDR) on the condition that this FDR did not exceed 0.1. Additionally, we kept all features with FDR<0.01 even if they were correlated with others. This shortened list of features was the basis for models' training that included final model-based feature selection.

2.6. Model training and evaluation

Model training was performed with the use of scikit-learn Python library [35]. Model architectures considered in our analysis included: logistic regression, multilayer perceptron (MLP), k nearest neighbors classifier, support vector classifier, decision tree, random forest, AdaBoost classifier, Gaussian process classifier, Gaussian Naive Bayes classifier, Quadratic discriminant analysis. Whenever possible (in logistic regression, support vector classifiers, decision trees and random forest) two methods of weighting cases during training were applied: 1) equal weights of all cases 2) weights inversely proportional to classes' frequencies to account for imbalanced data set with less than 1/3 of cases in RIHT group. Full list of tested model parameters is collected in Table S3.

Models were derived using three feature sets: 1) only clinical and dosimetric features (clinical model), 2) only radiomic features (radiomic model), 3) radiomic, clinical and dosimetric features (radiomic+clinical model). Final set of input features for each model was determined by forward method of Sequential Feature Selector from mlxtend library [36]. Feature selector used area under the ROC curve (AUC) as a measure of models' performance and was allowed to choose from 2 to 5 features.

All models were first derived using training data and then their performance was tested on validation data. AUC, accuracy, sensitivity, specificity and F-score were calculated for each model. The analysis was performed for Variants Ia, Ib and II of clinical and radiomic features derivation. Models with the highest F-score were selected for each variant of analysis and input feature set. Finally, ensembles of best clinical and radiomic models were considered that combined them in following ways:

- logical conjunction (AND): positive prediction only when both models predicted RIHT,
- logical disjunction (OR): positive prediction when any of the two models predicted RIHT,
- averaged probability (PROBA): probability (raw output) of two models were averaged and new decision threshold selected using ROC curve for training set.

3. Results

3.1. Feature stability analysis

Results of feature stability analysis differed slightly between variant I and II. In variant I, where all images were acquired in the same center, we identified more features as highly stable (ICC>0.9) than in variant II (721, Fig. 2A vs 621 features, Fig. 2B). In line with this observation, lowest stability class was assigned to 211 in variant II and only to 138 features in variant I (stability analysis was performed before batch correction, thus at this stage variant Ia and Ib are not distinguished). Independently of the variant, all shape features were excellently stable even though our test consisted essentially in disturbing thyroid mask shape. Excellent or good stability was observed for majority of first order, GLCM, GLRLM, and GLDM features, while for GLSZM and NGTDM the fraction of unstable features was higher. In majority, stable features overlapped between variants I and II (Fig. 2C). Stability of features depended also on the applied filter (Fig. 2D), with exponential and gradient filter ensuring on average higher stability than square, square root, logarithm and some wavelet filters.

In further analysis, we included all features with excellent and good stability: 926 in variant I and 869 in variant II. Stability classes for all features are reported in Table S4.

3.2. Feature processing and filtration

Before filtration, features were scaled, normalized (Fig. S1) and in variant Ib batch effect was corrected (Fig. 2S). First stage of filtration consisted of exclusion of features that in univariate analysis did not differentiate patients who did or did not develop RIHT (detailed results of this analysis in Table S5). It resulted in selection of 165 radiomic features in variant I and 166 in variant II, among which 153 were common to both variants (Fig. 3A). Having a limited number of patients, we decided to reduce these

A

	Stability category in variant I			
feature group	excellent	good	moderate	low
shape	14	0	0	0
first order	141	37	53	21
GLCM	195	60	60	21
GLRLM	139	23	36	26
GLSZM	94	47	53	30
GLDM	116	19	37	24
NGTDM	22	19	13	16
total	721	205	252	138

B

	Stability category in variant II			
feature group	excellent	good	moderate	low
shape	14	0	0	0
first order	121	47	46	38
GLCM	161	82	46	47
GLRLM	128	31	36	29
GLSZM	74	49	55	46
GLDM	103	26	38	29
NGTDM	20	13	15	22
total	621	248	236	211

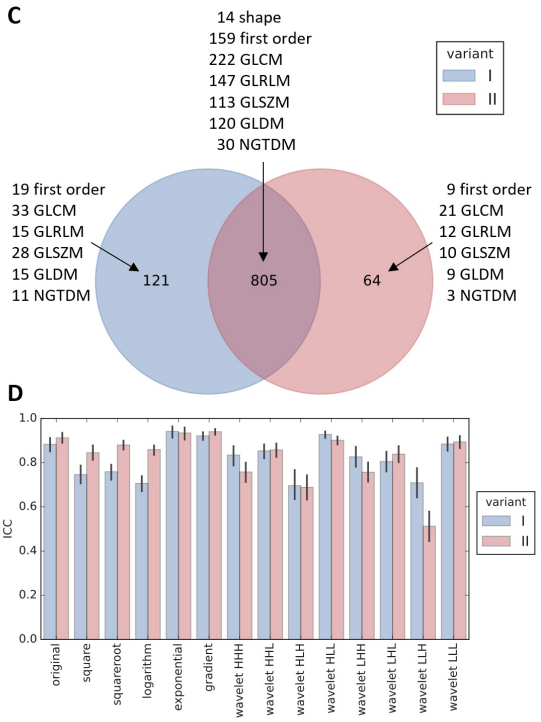

Figure 2. Results of feature stability assessment. Number of features in each group classified to different stability categories for variant I (A) and II (B). Overlap of features with at least good stability between variant I and II (C). Average stability of all radiomic features with respect to applied image filter (D, shown mean with 95% confidence interval).n (B).

numbers further before model development by elimination of highly correlated features. Selection of representatives from each group of such related variables left 67 features in variant Ia, 68 in variant Ib and 66 in variant II (lists of features in Table S6). Again, the overlap between variants, consisting of 61 radiomic features, greatly outnumbered distinct variables for each variant (Fig. 3C). At both stages of filtration, features from original image and the one with exponential filter applied were preferred over other groups (Fig. 3BD).

3.3. Models

The training of all the considered models on the reduced set of features and their validation resulted in selection of top 9 models, one for each variant and a feature set. Their names, along with the features they were built upon, are presented in Table 2, while details of all analysed models can be found in Table S7.

For all the variants, when considering models derived only with clinical features, Gaussian process classifier was chosen. Likewise, the same set of features was selected, i.e., mean of thyroid dose, median of thyroid dose and volume of the thyroid. For radiomic and clinical+radiomic feature sets, the selected models were: logistic regression without regularization, with equal class weights and excluded intercept for variant Ia and multilayer perceptron with 4 hidden neurons for variants Ib and II. For those feature sets, selected features were more diverse.

In all variants of analysis, radiomic models performed similarly to clinical models (Fig. 4A-C); comparisons of ROC AUC by DeLong test never showed superiority of models using radiomic features vs clinical/dosimetric model (Table 3. Slight improvement of radiomic models is observed after batch effect correction (variant Ib). The statistical measures (sensitivity, specificity, accuracy and f-score) for each of the selected models and the ensembles of clinical and radiomic models were calculated (Fig. 4D-I). Radiomic+clinical model was not included in ensembles, because it contained a very similar feature set

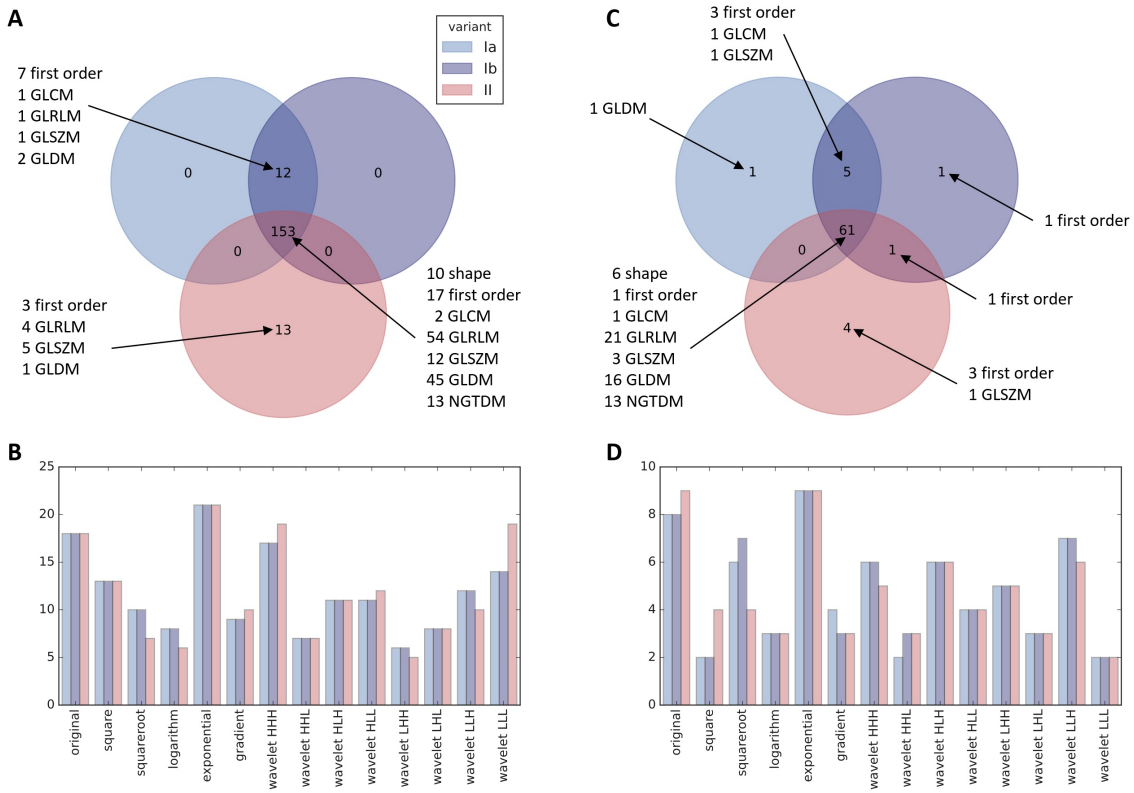

Figure 3. Features with FDR<0.1 in univariate analysis grouped by radiomics category (A) and by image filter (B). Features selected for model grouped by radiomics category (C) and by image filter (D).

to radiomic and clinical models taken together. Based on those measures, we picked clinical, radiomic and OR models for the comparison with the external NTCP models for radiation-induced hypothyroidism [8,9,37–39].

Table 2. Model architectures and features selected for each of the variants and feature sets.

	VARIANT Ia		**VARIANT Ib**		**VARIANT II**	
	model	features	model	features	model	features
clinical (same for Ia and Ib)	GPC	D_{mean} D50 $V_{thyroid}$	GPC	D_{mean} D50 $V_{thyroid}$	GPC	D_{mean} D50 $V_{thyroid}$
radiomic	LR_E	wavelet HHH GLSZM zone percentage logarithm NGTDM coarseness	MLP_4	original NGTDM coarseness wavelet LLL NGTDM coarseness exponential GLDM small dependence low gray level emphasis logarithm NGTDM coarseness	MLP_4	exponential GLDM small dependence low gray level emphasis logarithm NGTDM coarseness
clinical+radiomic	LR_E	sex original shape least axis length exponential GLRLM run percentage exponential GLDM small dependence low gray level emphasis logarithm NGTDM coarseness	MLP_4	original NGTDM coarseness wavelet LLL NGTDM coarseness exponential GLDM small dependence low gray level emphasis logarithm NGTDM coarseness	MLP_4	sex original shape least axis length exponential GLRLM run percentage exponential GLDM small dependence low gray level emphasis logarithm NGTDM coarseness

GPC – Gaussian process classifier, LR_E – logistic regression with equal cases weights, MLP_4 – MLP network with 4 neurons in single hidden layer

Table 3. Comparison between best models with and without radiomic features.

	VARIANT Ia		VARIANT Ib		VARIANT II	
Model	AUC ± SE	p	AUC ± SE	p	AUC ± SE	p
clinical	0.90±0.07	-	0.90±0.07	-	0.95±0.05	-
radiomic	0.89±0.07	0.9196	0.94±0.05	0.6471	0.91±0.07	0.6263
radiomic+clinical	0.85±0.08	0.6386	0.94±0.05	0.6471	0.55±0.12	0.0008
PROBA	0.90±0.07	1.0000	0.95±0.05	0.5549	0.93±0.06	0.7940

p-values for comparison of each model with radiomic features vs clinical model

Compared with previous examples (Fig. 5), our models tend to be slightly less sensitive, but more specific and accurate. Models by Cella et al. and Vogelious et al. significantly overestimate risk of RIHT, declaring all or almost all patients as having high risk of this complication. The model by Ronjom et al. seemed comparable to our models.

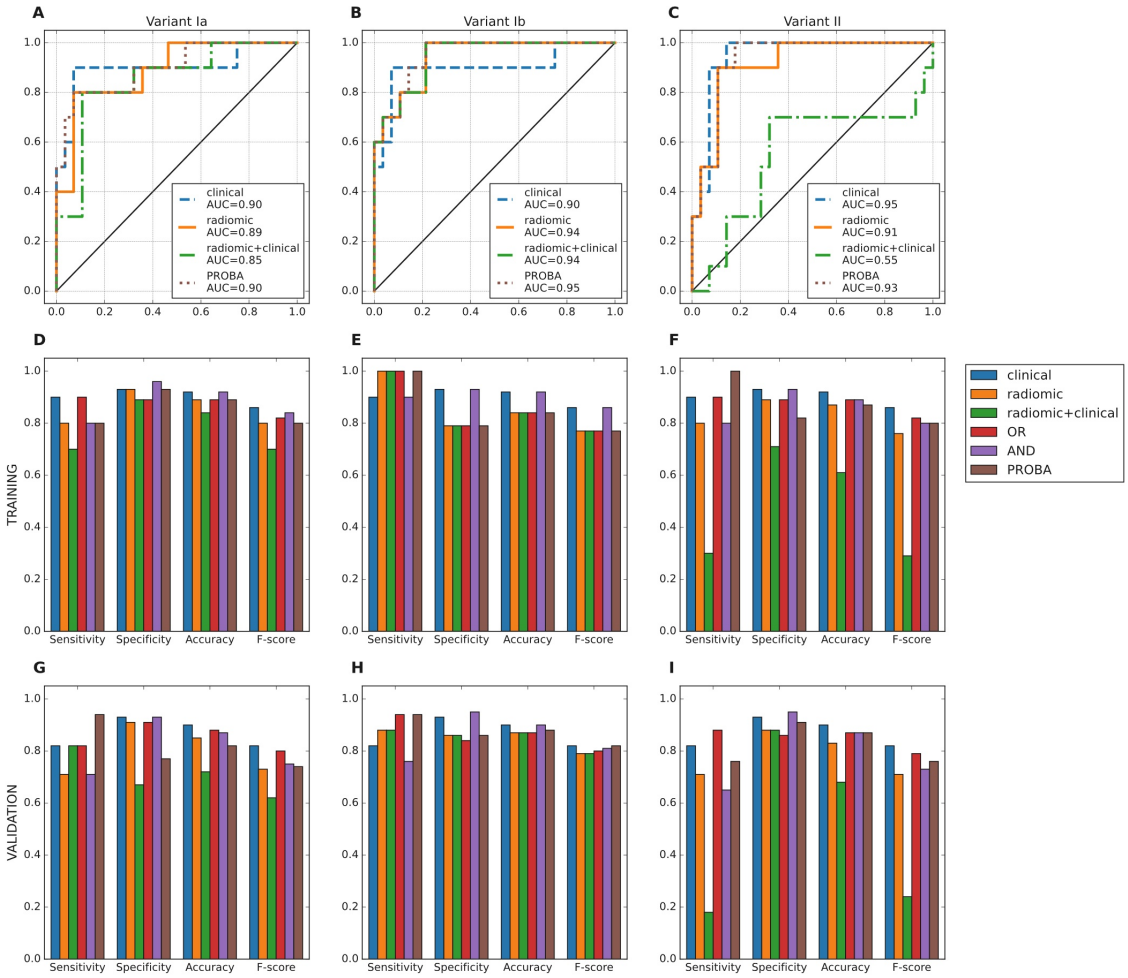

Figure 4. AUC ROC of the selected models (clinical, radiomic, radiomic+clinical) and their ensembles for the training set (A, B, C). Comparison of statistical measures for the selected model architectures of the considered feature sets and ensembles of those models: training (D, E, F) and validation (G, H, I).

Radiomic features included in the final models came from original, logarithm, exponential and wavelet (LLL, HHH) images. Majority of these features measured nonuniformity of the thyroid region (e.g. coarseness, zone percentage) and were higher

in patients who developed RIHT (Fig. 6). The only feature lower in these patients was least axis length of thyroid (Fig. S3).

4. Discussion

Here, using data from 98 patients with OPC treated with definite RT, we identified radiomic features predicting the occurrence of RIHT in 2-year follow-up and contrasted our radiomic-based model with published NTCP models based on clinical and dosimetric parameters. Our model performed comparably to models published Rønjom et al.[8] and by Boomsma et al. [9] and better than three other external models [37–39]. Since predictions based on CT-derived radiomic features are agnostic to dose distribution, they reflect the baseline susceptibility of individual thyroid glands to damage inflicted by radiation doses typical for head and neck cancer RT. They may be therefore leveraged to identify patients in whom close attention should be paid to minimize radiation to the thyroid and the risk of RIHT; alternatively, dosimetric/clinical and radiomic models can be combined to improve prediction accuracy.

Most features in our radiomic model were calculated in filtered CT images and described non-uniformity or coarseness of thyroid region, sometimes emphasizing low grey levels. Invariably higher values of these features, indicating greater non-homogeneity, especially of darker (lower Hounsfield units) regions of thyroid, were characteristic for patients who later developed RIHT.

Differences in performance between our models and external ones as well as those between different external models likely stem from differences in patient cohort characteristics and chosen treatment protocols, which has already been discussed in details in previous study by Nowicka et al. [7] that used data from the same cohort of patients.

An important contribution of our study is analytical pipeline that adds to the recommendations from first reference manual for image biomarker studies [11] and Radiomic Quality Score (RQS) [40]. Addressing criteria from RQS, we reported imaging protocols from participating institutions and verified the stability of features by segmentation perturbation, reducing the number of features and applying multiple comparisons correction. Our simulation study showed that some features are unstable with respect to inaccuracies in thyroid contouring; however, we identified a subset of relatively stable features, corresponding with the results of study on robustness of radiomic features by Zwanenburg et al. [41]. Then, we scaled and normalized the features and performed batch effect correction in one variant of the analysis. This pipeline may be reused and extended for similar projects requiring combining data from many institutions or imaging machines.

We included non-radiomic patient characteristics in the study, reported models' quality statistics and validated models. Patients were recruited prospectively in 3 clinics for the study by Nowicka et al. with the same endpoint of RIHT [7], however radiomic analysis was included in the protocol at the later stage. In the place of validation against "golden standard", that is not established for prediction of RIHT, we compared our models with those by other authors. As reported by reviews of radiomic studies [42,43], full adherence to the guidelines is rarely achieved, however even partial compliance improves the quality of research." A TRIPOD guidelines checklist that describes the critical aspects of our work has been provided in the supplementary materials (Table S8).

Our study has several limitations inherent to radiomic studies. The observed batch effects of radiomic features related to oncologic centers (possibly due to use of different CT machines or their settings) may be, for research purposes, solved by batch effect removal tools developed for other high-throughput studies [34,44]. However, such procedure complicates the translation of results and has not been extensively validates in terms of preserving data integrity. Furthermore, our data set has a relatively small sample size and the developed models require further validation before they can be applied in the clinical setting. Although the validation groups were sufficiently numbered to detect statistically significant differences exceeding 13% of accuracy it would mandate a larger

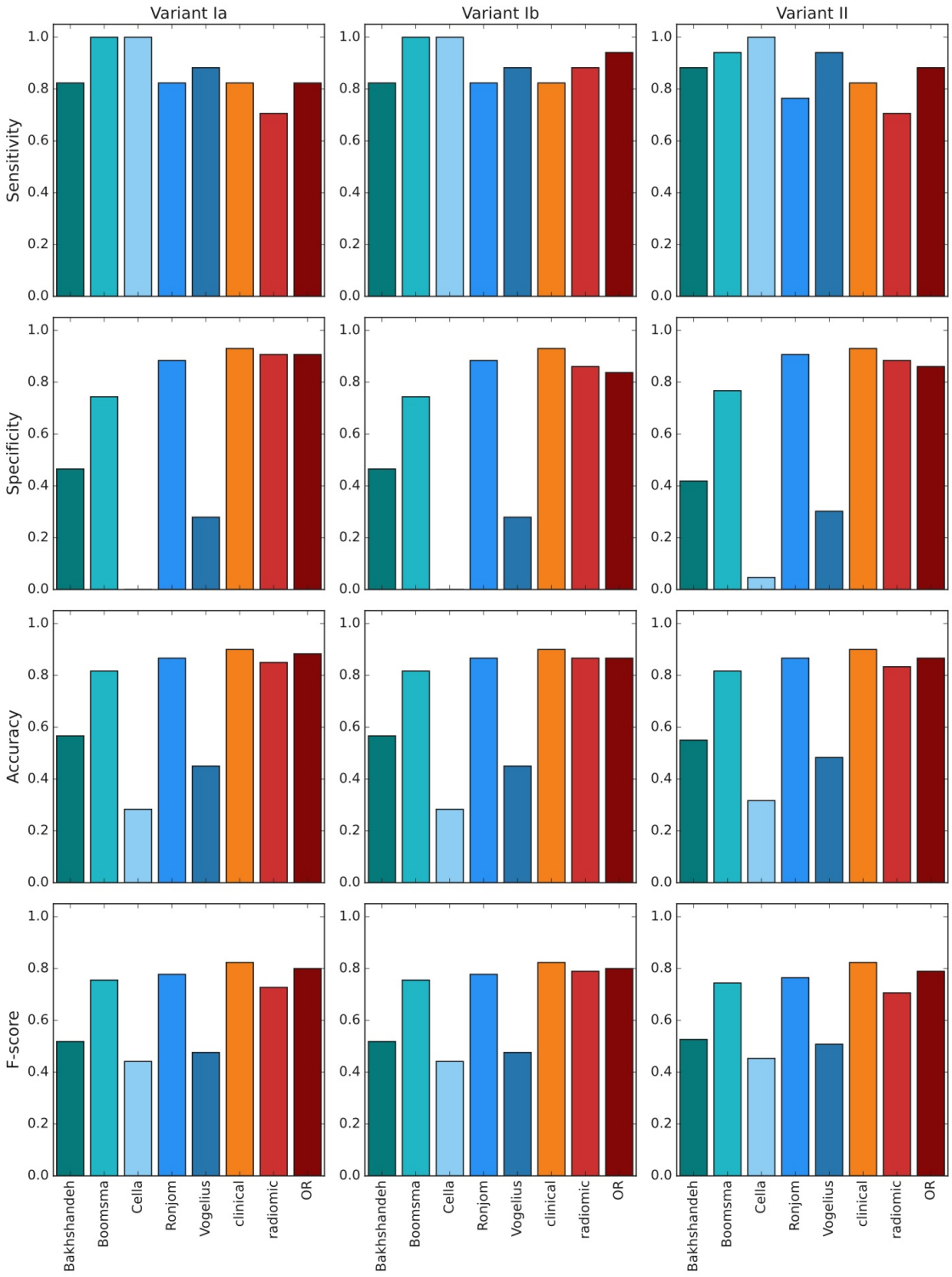

Figure 5. Comparison of quality measures between external NTCP models (cool colors, 5 bars from the left) and our top model selections (warm colors, 3 bars from the right), validation data.

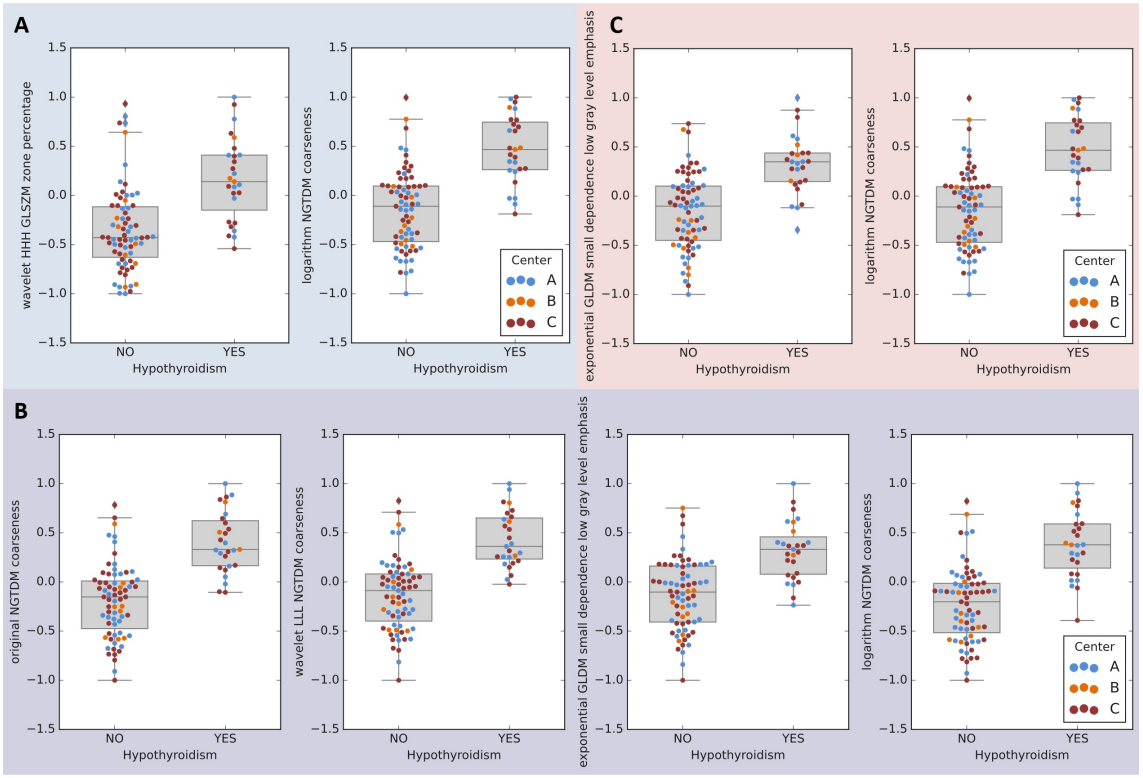

Figure 6. Transformed (normalized and scaled) values of features retained in radiomic models. A: variant Ia, B: variant: Ib, C: variant II.

study to confirm true non-interiority or equivalence of the radiomic and NTCP models, but owing to the presented results planning such an endeavor is possible.

5. Conclusions

Radiomic models reliant on CT scans showed a similar or better predictive potential for RIHT in OPC patients that the best currently used clinical and NTCP models with the additional benefit of being independent of treatment planning and readily deployable across different imaging data.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/1010000/s1, Figure S1: Noramalization and scaling o radiomic features, Figure S2: Batch effect correction, Figure S3: Distribution of radiomic features included in models, Table S1: Raw patients' data, Table S2: Raw values of radiomic features, Table S3: Parameters of analysed models' architectures, Table S4: Stability of radiomic features, Table S5: Univariate analysis and filtration of radiomic features, Table S6: List of radiomic features selected after univariate analysis, Table S7: Quality metrics for all analysed models, Table S8: Tripod checklist for our study.

Author Contributions: Conceptualization, Zuzanna Nowicka, Wojciech Fendler and Bartlomiej Tomasik; Data curation, Urszula Smyczynska, Szymon Grabia, Zuzanna Nowicka, Anna Papis-Ubych, Robert Bibik, Tomasz Latusek, Tomasz Rutkowski, Jacek Fijuth and Bartlomiej Tomasik; Formal analysis, Urszula Smyczynska, Szymon Grabia and Wojciech Fendler; Funding acquisition, Zuzanna Nowicka, Wojciech Fendler and Bartlomiej Tomasik; Investigation, Urszula Smyczynska, Szymon Grabia, Zuzanna Nowicka, Anna Papis-Ubych, Robert Bibik, Tomasz Latusek, Tomasz Rutkowski, Jacek Fijuth and Bartlomiej Tomasik; Methodology, Urszula Smyczynska, Szymon Grabia, Zuzanna Nowicka and Bartlomiej Tomasik; Project administration, Wojciech Fendler and Bartlomiej Tomasik; Resources, Anna Papis-Ubych, Robert Bibik, Tomasz Latusek, Tomasz Rutkowski, Jacek Fijuth and Bartlomiej Tomasik; Software, Urszula Smyczynska and Szymon Grabia; Supervision, Wojciech Fendler; Validation, Urszula Smyczynska, Szymon Grabia and Zuzanna Nowicka; Visualization, Urszula Smyczynska and Szymon Grabia; Writing – original

draft, Urszula Smyczynska, Szymon Grabia, Zuzanna Nowicka, Wojciech Fendler and Bartlomiej Tomasik; Writing – review & editing, Urszula Smyczynska, Szymon Grabia, Zuzanna Nowicka, Anna Papis-Ubych, Robert Bibik, Tomasz Latusek, Tomasz Rutkowski, Jacek Fijuth, Wojciech Fendler and Bartlomiej Tomasik.

Funding: This research was funded by the FIRST TEAM project financed from the Smart Growth Operational Program and coordinated by the Foundation for Polish Science, by the PRELUDIUM project financed by the National Science Center (NCN) (2016/21/N/NZ5/01938) granted to BT, by the Medical University of Lodz grant number 564/1-000-00/564-20-025 granted to ZN.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Bioethics Committee of the Medical University of Lodz (KE/7/10, RNN/65/18).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Raw, anonymized clinical data and values of calculated radiomic features were included in supplementary material.

Acknowledgments: Bartłomiej Tomasik gratefully acknowledges the financial support provided by the Foundation for Polish Science and by the Polish National Agency for Academic Exchange (the Walczak Programme).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AUC	area under curve
CT	computed tomography
FDR	alse discovery rate
GLCM	gray level cooccurrence matrix
GLDM	gray level dependence matrix
GLRLM	gray level run length matrix
GLSZM	gray level size zone matrix
GPC	Gaussian process classifier
HNG	head and neck cancer
ICC	inter-class correlation coefficient
IMRT	intensity-modulated radiation therapy
MLP	multilayer perceptron
NGTDM	neighborhood gray tone difference matrix
NTCP	normal tissue complication probability
OAR	organ at risk
OPC	oropharyngeal cancer
PACS	picture archiving and communication system
ROI	region of interest
RT	radiation therapy
RIHT	radiation-induced hypothyroidism

Appendix A. List of calculated radiomic features

Shape features

- elongation
- flatness
- least axis length
- major axis length
- maximum 2D diameter column
- maximum 2D diameter row
- maximum 2D diameter slice
- maximum 3D diameter

- mesh volume
- minor axis length
- sphericity
- surface area
- surface volume ratio
- voxel volume

First order features

- 10. percentile
- 90. percentile
- energy
- entropy
- interquartile range
- kurtosis
- maximum
- mean absolute deviation
- mean
- median
- minimum
- range
- robust mean absolute deviation
- root mean squared
- skewness
- total energy
- uniformity
- variance

Gray level cooccurrence matrix-based (GLCM) features

- autocorrelation
- cluster prominence
- cluster shade
- cluster tendency
- contrast
- correlation
- difference average
- difference entropy
- difference variance
- inverse difference (ID), homogeneity 1
- inverse difference moment (IDM), homogeneity 2
- inverse difference moment normalized (IDMN)
- inverse difference normalized (IDN)
- informational measure of correlation 1 (IMC1)
- informational measure of correlation 2 (IMC2)
- inverse variance
- joint average
- joint energy
- joint entropy
- maximal correlation coefficient (MCC)
- maximum probability
- sum average
- sum entropy
- sum squares

Gray level dependence matrix-based (GLDM) features

- dependence entropy
- dependence non uniformity

- dependence non uniformity normalized
- dependence variance
- gray level non uniformity
- gray level variance
- high gray level emphasis
- large dependence emphasis
- large dependence high gray level emphasis
- large dependence low gray level emphasis
- low gray level emphasis
- small dependence emphasis
- small dependence high gray level emphasis
- small dependence low gray level emphasis

Gray level run length matrix-based (GLRLM) features

- gray level non uniformity
- gray level non uniformity normalized
- gray level variance
- high gray level run emphasis
- long run emphasis
- long run high gray level emphasis
- long run low gray level emphasis
- low gray level run emphasis
- run entropy
- run length non uniformity
- run length non uniformity normalized
- run percentage
- run variance
- short run emphasis
- short run high gray level emphasis
- short run low gray level emphasis

Gray level size zone matrix-based (GLSZM) features

- gray level non uniformity
- gray level non uniformity normalized
- gray level variance
- high gray level zone emphasis
- large area emphasis
- large area high gray level emphasis
- large area low gray level emphasis
- low gray level zone emphasis
- size zone non uniformity
- size zone non uniformity normalized
- small area emphasis
- small area high gray level emphasis
- small area low gray level emphasis
- zone entropy
- zone percentage
- zone variance

Neighborhood gray tone difference matrix-based (NGTDM) features

- busyness
- coarseness
- complexity
- contrast
- strength

References

1. Brodin, N.P.; Kabarriti, R.; Garg, M.K.; Guha, C.; Tome, W.A. Systematic Review of Normal Tissue Complication Models Relevant to Standard Fractionation Radiation Therapy of the Head and Neck Region Published After the QUANTEC Reports. *International Journal of Radiation Oncology Biology Physics* **2018**, *100*, 391–407. doi:10.1016/j.ijrobp.2017.09.041.
2. Vigário, P.; Teixeira, P.; Reuters, V.; Almeida, C.; Maia, M.; Silva, M.; Vaisman, M. Perceived health status of women with overt and subclinical hypothyroidism. *Medical principles and practice* **2009**, *18*, 317–322. doi:10.1159/000215731.
3. Thvilum, M.; Brandt, F.; Almind, D.; Christensen, K.; Hegedüs, L.; Brix, T.H. Excess mortality in patients diagnosed with hypothyroidism: a nationwide cohort study of singletons and twins. *The Journal of clinical endocrinology and metabolism* **2013**, *98*, 1069–1075. doi:10.1210/JC.2012-3375.
4. Hassan, A.; Altamirano-Ufion, A.; Zulfiqar, B.; Boddu, P. Sub-Clinical Hypothyroidism and Its Association With Increased Cardiovascular Mortality: Call for Action. *Cardiology Research* **2017**, *8*, 31–35. doi:10.14740/CR524W.
5. Rønjom, M.F.; Brink, C.; Bentzen, S.M.; Hegedüs, L.; Overgaard, J.; Johansen, J. Hypothyroidism after primary radiotherapy for head and neck squamous cell carcinoma: normal tissue complication probability modeling with latent time correction. *Hypothyroidism after primary radiotherapy for head and neck squamous cell carcinoma: normal tissue complication probability modeling with latent time correction* **2013**, *109*, 317–322. doi:10.1016/J.RADONC.2013.06.029.
6. Kamal, M.; Peeler, C.R.; Yepes, P.; Mohamed, A.S.; Blanchard, P.; Frank, S.; Chen, L.; Jethanandani, A.; Kuruvilla, R.; Greiner, B.; Harp, J.; Granberry, R.; Mehta, V.; Rock, C.; Hutcheson, K.; Cardenas, C.; Gunn, G.B.; Fuller, C.; Mirkovic, D. Radiation-Induced Hypothyroidism After Radical Intensity Modulated Radiation Therapy for Oropharyngeal Carcinoma. *Advances in Radiation Oncology* **2020**, *5*, 111–119. doi:10.1016/j.adro.2019.08.006.
7. Nowicka, Z.; Tomasik, B.; Papis-Ubych, A.; Bibik, R.; Graczyk, Ł.; Latusek, T.; Rutkowski, T.; Wyka, K.; Fijuth, J.; Schoenfeld, J.D.; Chałubińska-Fendler, J.; Fendler, W. Radiation-induced hypothyroidism in patients with oropharyngeal cancer treated with imrt: Independent and external validation of five normal tissue complication probability models. *Cancers* **2020**, *12*, 1–14. doi:10.3390/cancers12092716.
8. Rønjom, M.F.; Brink, C.; Bentzen, S.M.; Hegedüs, L.; Overgaard, J.; Petersen, J.B.; Primdahl, H.; Johansen, J. External validation of a normal tissue complication probability model for radiation-induced hypothyroidism in an independent cohort. *Acta Oncologica* **2015**, *54*, 1301–1309. doi:10.3109/0284186X.2015.1064160.
9. Boomsma, M.J.; Bijl, H.P.; Christianen, M.E.; Beetz, I.; Chouvalova, O.; Steenbakkers, R.J.; Van Der Laan, B.F.; Wolffenbuttel, B.H.; Oosting, S.F.; Schilstra, C.; Langendijk, J.A. A prospective cohort study on radiation-induced hypothyroidism: Development of an NTCP model. *International Journal of Radiation Oncology Biology Physics* **2012**, *84*, e351–e356. doi:10.1016/j.ijrobp.2012.05.020.
10. Lambin, P.; Rios-Velazquez, E.; Leijenaar, R.; Carvalho, S.; Van Stiphout, R.G.; Granton, P.; Zegers, C.M.; Gillies, R.; Boellard, R.; Dekker, A.; Aerts, H.J. Radiomics: Extracting more information from medical images using advanced feature analysis. *European Journal of Cancer* **2012**, *48*, 441–446. doi:10.1016/j.ejca.2011.11.036.
11. Zwanenburg, A.; Vallières, M.; Abdalah, M.A.; Aerts, H.J.W.L.; Andrearczyk, V.; Apte, A.; Ashrafinia, S.; Bakas, S.; Beukinga, R.J.; Boellaard, R.; Bogowicz, M.; Boldrini, L.; Buvat, I.; Cook, G.J.R.; Davatzikos, C.; Depeursinge, A.; Desseroit, M.C.; Dinapoli, N.; Dinh, C.V.; Echegaray, S.; El Naqa, I.; Fedorov, A.Y.; Gatta, R.; Gillies, R.J.; Goh, V.; Götz, M.; Guckenberger, M.; Ha, S.M.; Hatt, M.; Isensee, F.; Lambin, P.; Leger, S.; Leijenaar, R.T.; Lenkowicz, J.; Lippert, F.; Losnegård, A.; Maier-Hein, K.H.; Morin, O.; Müller, H.; Napel, S.; Nioche, C.; Orlhac, F.; Pati, S.; Pfaehler, E.A.; Rahmim, A.; Rao, A.U.; Scherer, J.; Siddique, M.M.; Sijtsema, N.M.; Socarras Fernandez, J.; Spezi, E.; Steenbakkers, R.J.; Tanadini-Lang, S.; Thorwarth, D.; Troost, E.G.; Upadhaya, T.; Valentini, V.; van Dijk, L.V.; van Griethuysen, J.; van Velden, F.H.; Whybra, P.; Richter, C.; Löck, S. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology* **2020**, p. 191145. doi:10.1148/radiol.2020191145.
12. Aerts, H.J.; Velazquez, E.R.; Leijenaar, R.T.; Parmar, C.; Grossmann, P.; Carvalho, S.; Bussink, J.; Monshouwer, R.; Haibe-Kains, B.; Rietveld, D.; Hoebers, F.; Rietbergen, M.M.; Leemans, C.R.; Dekker, A.; Quackenbush, J.; Gillies, R.J.; Lambin, P. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications* **2014**, *5*, 4006. doi:10.1038/ncomms5006.
13. Leijenaar, R.T.; Carvalho, S.; Hoebers, F.J.; Aerts, H.J.; Van Elmpt, W.J.; Huang, S.H.; Chan, B.; Waldron, J.N.; Osullivan, B.; Lambin, P. External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma. *Acta Oncologica* **2015**, *54*, 1423–1429. doi:10.3109/0284186X.2015.1061214.
14. Nardone, V.; Tini, P.; Pastina, P.; Botta, C.; Reginelli, A.; Carbone, S.F.; Giannicola, R.; Calabrese, G.; Tebala, C.; Guida, C.; Giudice, A.; Barbieri, V.; Tassone, P.; Tagliaferri, P.; Cappabianca, S.; Capasso, R.; Luce, A.; Caraglia, M.; Mazzei, M.A.; Pirtoli, L.; Correale, P. Radiomics predicts survival of patients with advanced non-small cell lung cancer undergoing PD-1 blockade using Nivolumab. *Oncology Letters* **2020**, *19*, 1559–1566. doi:10.3892/ol.2019.11220.
15. Vallières, M.; Kay-Rivest, E.; Perrin, L.J.; Liem, X.; Furstoss, C.; Aerts, H.J.; Khaouam, N.; Nguyen-Tan, P.F.; Wang, C.S.; Sultanem, K.; Seuntjens, J.; El Naqa, I. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Scientific Reports* **2017**, *7*, 1–14, [1703.08516]. doi:10.1038/s41598-017-10371-5.
16. Grossmann, P.; Stringfield, O.; El-Hachem, N.; Bui, M.M.; Rios Velazquez, E.; Parmar, C.; Leijenaar, R.T.; Haibe-Kains, B.; Lambin, P.; Gillies, R.J.; Aerts, H.J. Defining the biological basis of radiomic phenotypes in lung cancer. *eLife* **2017**, *6*. doi:10.7554/eLife.23421.

17. Leijenaar, R.T.; Bogowicz, M.; Jochems, A.; Hoebers, F.J.; Wesseling, F.W.; Huang, S.H.; Chan, B.; Waldron, J.N.; O'Sullivan, B.; Rietveld, D.; Leemans, C.R.; Brakenhoff, R.H.; Riesterer, O.; Tanadini-Lang, S.; Guckenberger, M.; Ikenberg, K.; Lambin, P. Development and validation of a radiomic signature to predict HPV (p16) status from standard CT imaging: a multicenter study. *The British Journal of Radiology* **2018**, *91*, 2017049811075.
18. Cong, M.; Feng, H.; Ren, J.L.; Xu, Q.; Cong, L.; Hou, Z.; yuan Wang, Y.; Shi, G. Development of a predictive radiomics model for lymph node metastases in pre-surgical CT-based stage IA non-small cell lung cancer. *Lung Cancer* **2020**, *139*, 73–79. doi:10.1016/j.lungcan.2019.11.003.
19. Kwan, J.Y.Y.; Su, J.; Huang, S.H.; Ghoraie, L.S.; Xu, W.; Chan, B.; Yip, K.W.; Giuliani, M.; Bayley, A.; Kim, J.; Hope, A.J.; Ringash, J.; Cho, J.; McNiven, A.; Hansen, A.; Goldstein, D.; de Almeida, J.R.; Aerts, H.J.; Waldron, J.N.; Haibe-Kains, B.; O'Sullivan, B.; Bratman, S.V.; Liu, F.F. Radiomic Biomarkers to Refine Risk Models for Distant Metastasis in HPV-related Oropharyngeal Carcinoma. *International Journal of Radiation Oncology Biology Physics* **2018**, *102*, 1107–1116. doi:10.1016/j.ijrobp.2018.01.057.
20. Sun, R.; Limkin, E.J.; Vakalopoulou, M.; Dercle, L.; Champiat, S.; Han, S.R.; Verlingue, L.; Brandao, D.; Lancia, A.; Ammari, S.; Hollebecque, A.; Scoazec, J.Y.; Marabelle, A.; Massard, C.; Soria, J.C.; Robert, C.; Paragios, N.; Deutsch, E.; Ferté, C. A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study. *The Lancet Oncology* **2018**, *19*, 1180–1191. doi:10.1016/S1470-2045(18)30413-3.
21. Pérez-Morales, J.; Tunali, I.; Stringfield, O.; Eschrich, S.A.; Balagurunathan, Y.; Gillies, R.J.; Schabath, M.B. Peritumoral and intratumoral radiomic features predict survival outcomes among patients diagnosed in lung cancer screening. *Scientific Reports* **2020**, *10*, 1–15. doi:10.1038/s41598-020-67378-8.
22. Vaidya, P.; Bera, K.; Gupta, A.; Wang, X.; Corredor, G.; Fu, P.; Beig, N.; Prasanna, P.; Patil, P.D.; Velu, P.D.; Rajiah, P.; Gilkeson, R.; Feldman, M.D.; Choi, H.; Velcheti, V.; Madabhushi, A. CT derived radiomic score for predicting the added benefit of adjuvant chemotherapy following surgery in stage I, II resectable non-small cell lung cancer: a retrospective multicohort study for outcome prediction. *The Lancet Digital Health* **2020**, *2*, e116–e128. doi:10.1016/S2589-7500(20)30002-9.
23. Coroller, T.P.; Agrawal, V.; Huynh, E.; Narayan, V.; Lee, S.W.; Mak, R.H.; Aerts, H.J. Radiomic-Based Pathological Response Prediction from Primary Tumors and Lymph Nodes in NSCLC. *Journal of Thoracic Oncology* **2017**, *12*, 467–476. doi:10.1016/j.jtho.2016.11.2226.
24. Sha, X.; Gong, G.; Qiu, Q.; Duan, J.; Li, D.; Yin, Y. Discrimination of mediastinal metastatic lymph nodes in NSCLC based on radiomic features in different phases of CT imaging. *BMC Medical Imaging* **2020**, *20*, 1–8. doi:10.1186/s12880-020-0416-3.
25. van Dijk, L.V.; Brouwer, C.L.; van der Schaaf, A.; Burgerhof, J.G.; Beukinga, R.J.; Langendijk, J.A.; Sijtsema, N.M.; Steenbakkers, R.J. CT image biomarkers to improve patient-specific prediction of radiation-induced xerostomia and sticky saliva. *Radiotherapy and Oncology* **2017**, *122*, 185–191. doi:10.1016/j.radonc.2016.07.007.
26. Gabryś, H.S.; Buettner, F.; Sterzing, F.; Hauswald, H.; Bangert, M. Design and Selection of Machine Learning Methods Using Radiomics and Dosiomics for Normal Tissue Complication Probability Modeling of Xerostomia. *Frontiers in Oncology* **2018**, *8*, 35. doi:10.3389/FONC.2018.00035.
27. Krafft, S.P.; Rao, A.; Stingo, F.; Briere, T.M.; Court, L.E.; Liao, Z.; Martel, M.K. The utility of quantitative CT radiomics features for improved prediction of radiation pneumonitis. *Medical Physics* **2018**, *45*, 5317–5324. doi:10.1002/MP.13150.
28. Lucia, F.; Bourbonne, V.; Visvikis, D.; Miranda, O.; Gujral, D.M.; Gouders, D.; Dissaux, G.; Pradier, O.; Tixier, F.; Jaouen, V.; Bert, J.; Hatt, M.; Schick, U. Radiomics Analysis of 3D Dose Distributions to Predict Toxicity of Radiotherapy for Cervical Cancer. *Journal of Personalized Medicine* **2021**, *11*, 398. doi:10.3390/JPM11050398.
29. Rossi, L.; Bijman, R.; Schillemans, W.; Aluwini, S.; Cavedon, C.; Witte, M.; Incrocci, L.; Heijmen, B. Texture analysis of 3D dose distributions for predictive modelling of toxicity rates in radiotherapy. *Radiotherapy and Oncology* **2018**, *129*, 548–553. doi:10.1016/J.RADONC.2018.07.027.
30. Bourbonne, V.; Da-ano, R.; Jaouen, V.; Lucia, F.; Dissaux, G.; Bert, J.; Pradier, O.; Visvikis, D.; Hatt, M.; Schick, U. Radiomics analysis of 3D dose distributions to predict toxicity of radiotherapy for lung cancer. *Radiotherapy and Oncology* **2021**, *155*, 144–150. doi:10.1016/J.RADONC.2020.10.040.
31. Phil, T. Sikerdebaard/dcmrtstruct2nii: v1.0.19, 2020. doi:10.5281/ZENODO.4037865.
32. Van Griethuysen, J.J.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.; Fillion-Robin, J.C.; Pieper, S.; Aerts, H.J. Computational radiomics system to decode the radiographic phenotype. *Cancer Research* **2017**, *77*, e104–e107. doi:10.1158/0008-5472.CAN-17-0339.
33. Yeo, I.K.; Johnson, R. A new family of power transformations to improve normality or symmetry. *Biometrika* **2000**, *87*, 954–959. doi:10.1093/biomet/87.4.954.
34. Johnson, W.E.; Li, C.; Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **2007**, *8*, 118–127. doi:10.1093/biostatistics/kxj037.
35. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.
36. Raschka, S. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *The Journal of Open Source Software* **2018**, *3*. doi:10.21105/joss.00638.

37. Bakhshandeh, M.; Hashemi, B.; Mahdavi, S.R.M.; Nikoofar, A.; Vasheghani, M.; Kazemnejad, A. Normal tissue complication probability modeling of radiation-induced hypothyroidism after head-and-neck radiation therapy. *International Journal of Radiation Oncology Biology Physics* **2013**, *85*, 514–521. doi:10.1016/j.ijrobp.2012.03.034.
38. Cella, L.; Liuzzi, R.; Conson, M.; D'Avino, V.; Salvatore, M.; Pacelli, R. Development of multivariate NTCP models for radiation-induced hypothyroidism: a comparative analysis. *Radiation Oncology* **2012**, *7*, 224. doi:10.1186/1748-717X-7-224.
39. Vogelius, I.R.; Bentzen, S.M.; Maraldo, M.V.; Petersen, P.M.; Specht, L. Risk factors for radiation-induced hypothyroidism: A literature-based meta-analysis. *Cancer* **2011**, *117*, 5250–5260. doi:10.1002/cncr.26186.
40. Lambin, P.; Leijenaar, R.T.H.; Deist, T.M.; Peerlings, J. Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology* **2017**, *14*, 749–762. doi:10.1038/nrclinonc.2017.141.
41. Zwanenburg, A.; Leger, S.; Agolli, L.; Pilz, K.; Troost, E.G.; Richter, C.; Löck, S. Assessing robustness of radiomic features by image perturbation. *Scientific Reports* **2019**, *9*, 1–31, [1806.06719]. doi:10.1038/s41598-018-36938-4.
42. Sanduleanu, S.; Woodruff, H.C.; de Jong, E.E.; van Timmeren, J.E.; Jochems, A.; Dubois, L.; Lambin, P. Tracking tumor biology with radiomics: A systematic review utilizing a radiomics quality score. *Radiotherapy and Oncology* **2018**, *127*, 349–360. doi:10.1016/j.radonc.2018.03.033.
43. Park, J.E.; Kim, D.; Kim, H.S.; Park, S.Y.; Kim, J.Y.; Cho, S.J.; Shin, J.H.; Kim, J.H. Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. *European Radiology* **2020**, *30*, 523–536. doi:10.1007/s00330-019-06360-z.
44. Vuong, D.; Bogowicz, M.; Denzler, S.; Oliveira, C.; Foerster, R.; Amstutz, F.; Gabryś, H.S.; Unkelbach, J.; Hillinger, S.; Thierstein, S.; Xyrafas, A.; Peters, S.; Pless, M.; Guckenberger, M.; Tanadini-Lang, S. Comparison of robust to standardized CT radiomics models to predict overall survival for non-small cell lung cancer patients. *Medical Physics* **2020**, p. mp.14224. doi:10.1002/mp.14224.