Predicting tumor relapse from anatomical and ADC-derived radiomics biomarkers in brain tumor patients

L. Ackermann

Oktober 2025

Abstract

Background and Motivation: Accurate prediction of radiation therapy response remains difficult due to inter-patient variability and tumor heterogeneity. Identification of quantitative imaging biomarkers that can characterize these differences and improve outcome prediction is needed. *Materials and Methods*: This study investigated the association between radiomics features, clinical parameters and local treatment failure in patients with skull base chordoma and chondrosarcoma cancers treated with proton therapy. The final cohort consisted of 31 patients, of whom 6 experienced local treatment failure. Baseline diffusion weighted, T1-weighted and FLAIR-weighted magnetic resonance images, together with clinical parameters, were collected for all patients. A pipeline was developed for multimodal image registration, intensity normalization, and radiomics feature extraction. After removing highly correlated features, 50 radiomics features and one clinical parameter were included in a survival analysis. Feature robustness was evaluated using bootstrap resampling and a Cox proportional hazards model was fitted. Model performance was determined using the concordance index, the integrated Brier score, and Kaplan-Meier analysis. *Results*: The discriminative performance of the model plateaued after the inclusion of the five most robust features. Calibration performance was highest using the two most robust features, but remained comparably stable for five. The final Cox model incorporating five features achieved good risk stratification, as confirmed by Kaplan-Meier analysis. *Conclusion*: The presented pipeline provides a reproducible framework to extract radiomics features from multimodal MRI data and perform survival analysis in combination with clinical parameters. The features identified in this study show potential to capture valuable biological information associated with treatment outcome. Future studies with a larger dataset are needed to confirm these findings and to reliably identify imaging biomarkers to help make clinical decisions.

Contents

1 Introduction

1.1 Response prediction in radiotherapy

Radiation therapy is a cornerstone of cancer treatment, particularly for solid tumors, as it involves the precise delivery of ionising radiation to inhibit the proliferation of malignant cells. The primary objective of radiation therapy is to achieve optimal tumor control while minimizing toxicity to surrounding healthy tissues. Continuous advances in clinical practice and beam delivery technologies have improved overall treatment quality and success rate of current practices. Nevertheless, substantial inter-patient heterogeneity persists, resulting in variable therapeutic responses and, in some cases, tumor control failure. Consequently, accurate and individualized prediction of tumor response to treatment could guide clinical decision-making between alternative therapeutic strategies, facilitate risk–benefit assessment, and ultimately support the development of personalized radiation therapy aiming for improved clinical outcomes.[1].

Tumor control is typically estimated using dose-based tumor control probability (TCP) models that are specific to the treatment plan, indication, and fractionation scheme, rather than truly patient specific. In contrast, patient-level probabilities can be derived from statistical survival models that incorporate clinical, biological, or imaging covariates. The following sections provide a brief introduction to these complementary approaches.

Conventional TCP models

TCP models are based on radiobiology theory and generally use Poisson statistics to predict the probability of tumor control. They typically assume that tumor control is achieved if none of the initial n_0 clonogenic tumor cells survive after irradiation, leading to:

$$TCP = P(n = 0) = e^{-n_0 S} \tag{1}$$

where n_0 is the initial number of clonogenic tumor cells and $S = S(D)$ the surviving fraction of cancer cells following the application of $D[Gray, Gy]$ amount of radiation dose. $S(D)$ is commonly modeled by the linear-quadratic (LQ) model:

$$S(D) = e^{-\alpha D - \beta G D^2} \tag{2}$$

In the LQ model, the linear component represents damage caused by single-track events and depends on factors such as oxygenation, cell cycle phase, and linear energy transfer (LET) of the applied radiation. Its contribution is weighted by the coefficient α, which reflects the probability of lethal damage from a single radiation track. The quadratic term reflects damage resulting from the interaction of two independent radiation tracks, weighted by the coefficient β. This term characterises the sensitivity of tissue to fractionation and dose rate. The dose protraction factor G accounts for the repair of sublethal damage between dose fractions and is influenced by the intrinsic repair kinetics of the cells[2].

Conventional survival (time-to-event) analysis

An alternative approach to clinical outcome analysis focuses on tracking and predicting the time to treatment failure. This enables the generalisation of treatment response across a broad patient population and allows assessment of the clinical effectiveness of a given therapy by analysing failure events over time. Importantly, such methods can properly account for incomplete follow-up, such as patients lost to follow-up or deceased before failure, through the use of censored data, ensuring that these cases do not bias the estimations.

The most common regression model for survival analysis is the Cox Proportional Hazards (CPH) model, which estimates the relationship between covariates and the time to the event. Covariates can be categorical or continuous and describe patient specific characteristics (or clinical parameters), treatment parameters, or tumor characteristics.

The output of the CPH model is the hazard ratio (HR), which quantifies how the risk of experiencing the event of interest differs between two groups at any time point. An HR greater than 1 indicates a higher event risk, whereas an HR less than 1 indicates a lower risk relative to the reference group. This framework can be applied to analyse time-to-event endpoints such as time to local failure, disease progression, or overall survival[3].

The performance of prediction models is commonly evaluated using the concordance index (C-Index), which assesses the ability of a model to correctly rank the patients according to their predicted risks. The C-Index measures the proportion of all pairs of comparable patients in which the patient with higher predicted risk experienced the event earlier. Comparable pairs are defined as pairs in which at least one patient experiences the event[4]. A C-index of 0.5 corresponds to random prediction, while a value of 1.0 indicates perfect concordance between predicted and observed outcomes.

Emerging approaches based on imaging biomarkers

Conventional survival models are a valuable tool for clinical follow-up studies. In current practice, these studies most often are based on treatment plan dosimetry and clinical parameters (e.g. dose, age, tumor stage), but they rarely incorporate biological and functional heterogeneity information within the tumor.

Advances in quantitative imaging have enabled the extraction of radiomics features and quantitative imaging biomarkers (QIBs), providing new possibilities to characterize tumor phenotypes. Simply put, QIBs are objective, non-invasive indicators derived from medical images that describe tissue properties such as texture or function[5]. Unlike qualitative image interpretation by humans, QIBs provide standardized and reproducible measurements suitable for outcome analysis studies. There exist initiatives, such as the Image Biomarker Standardisation Initiative (IBSI), that ensure methodological consistency and comparability between studies. In particular, radiomics approaches can extract high-dimensional features from imaging modalities such as MRI and CT, or QIB maps, providing characteristic descriptors of tumor tissue not discernible to the human eye [6].

Therefore, when combined with clinical parameters, omics analyses of imaging data and QIBs can provide valuable complementary information for prognosis and prediction of treatment response in radiation therapy.

1.2 Imaging for outcome modeling

Imaging modalities

Medical imaging is essential for both the planning and delivery of radiation therapy treatments. For protons in particular, but also for photons to some extent, Treatment Planning Systems (TPS) require voxel-level electron density information to calculate the dose distribution within the patient. In common practice, the electron density is derived from the Hounsfield units (HU) of computed tomography (CT) imaging. HUs are converted to electron densities using calibration curves that establish the relationship between CT intensity values and the corresponding physical quantities of interest [7]. Despite its critical role in dose calculation, CT data provide primarily anatomical information, with inherently limited soft-tissue contrast, and are not ideally suited for tissue characterisation.

Magnetic resonance imaging (MRI) represents an important alternative, offering excellent soft tissue contrast essential for the delineation of target volumes (TV), as well as functional information, depending on the imaging protocol used. MRI provides a variety of acquisition sequences, such as T1-, T2-, and FLAIR-weighted imaging, each highlighting different tissue properties relevant for diagnostic and treatment planning tasks.

Fluid-attenuated inversion recovery (FLAIR) imaging is a T2-weighted sequence designed to suppress the signal from cerebrospinal fluid (CSF), thereby improving the detection of lesions and edema near CSF spaces. This technique is particularly effective in identifying pathologies associated with increased water content, including edema, gliosis, or tumor infiltration[8].

In addition to anatomical imaging, MRI also enables functional imaging, that visualises physiological processes within tissues. One such functional MRI technique is diffusion-weighted imaging (DWI), which can be used to generate QIB maps, such as Apparent Diffusion Coefficient (ADC) maps. ADC maps quantify the diffusion of water molecules in tissues, providing information about the cellular organization and density. ADC values are expressed in $[mm^2/s]$. A major advantage of ADC maps is that their voxel intensities correspond to physical quantities, making them, to some extent, less sensitive to scanner- and sequence-specific variations [9][5].

Radiomics features

Radiomics offers a systematic and reproducible approach to analyze imaging data. A large number of quantitative descriptors can be extracted and integrated into predictive models to expand the information base with an unprecedented level of detail derived from non-invasive anatomical and functional imaging.

Radiomic features can be broadly grouped into the following categories:

- First-order statistics, which describe intensity distributions within a region of interest (ROI) and provide measures such as intensity range, entropy, kurtosis, maximum, mean and uniformity.
- Texture features, which quantify spatial relationships and gray-level patterns between neighboring voxels. They describe how voxel intensities change in the ROI and are able to capture intra-tumoral heterogeneity not immediately visible to the human eye.

- Shape features, which describe the geometric and morphological characteristics of the tumor, including parameters such as surface area, volume, compactness, and sphericity. These features can describe irregular growth patterns and infiltration of the tumor in adjacent tissue[6].

1.3 Study Goals

The purpose of this study was to investigate the role of image-derived omics features in time to local failure in a cohort of skull-base chordoma and chondrosarcoma patients treated with proton therapy. The project covers the full analysis pipeline, from imaging data preparation to feature extraction and survival model fitting.

Retrospective clinical data were collected, including baseline T1-weighted images, diffusion-derived apparent diffusion coefficient (ADC) maps, and FLAIR images, along with treatment volume information and clinical outcomes related to the local tumor control, time to failure or censoring. Imaging data were processed to align all modalities within a consistent reference frame and normalised on a per-patient basis prior to radiomic feature extraction. Finally, the extracted features were integrated into a comprehensive survival analysis to identify predictors of local tumor control.

2 Materials and Methods

2.1 Clinical Cohort

The clinical cohort comprised 48 patients diagnosed with skull-base chordoma and chondrosarcoma. All received proton therapy at the Paul Scherrer Institute (PSI), where baseline diffusion weighted imaging (DWI), T1-weighted and FLAIR MRI, and CT scans were acquired. The gross tumor volumes (GTVs) were delineated on the planning CT and served as regions of interest for radiomic feature extraction.

Patient-specific characteristics collected included sex, date of birth, age at the start of proton therapy, the occurrence of an in-field local failure, and the corresponding time to event. The latter was defined as either the time to local failure or the time to last follow-up (i.e censoring). Of the initial 48 patients, 1 was excluded due to missing imaging data, 14 were excluded due to a partial field of view in T1-weighted MRI or DWI, and 2 due to their relatively small GTV size which, combined with registration uncertainty, would have yielded unreliable radiomics features. The final study cohort consisted of 31 patients, of whom six (19%) experienced in-field treatment failure. The patient selection process is illustrated in the PRISMA-style diagram fig. 1.

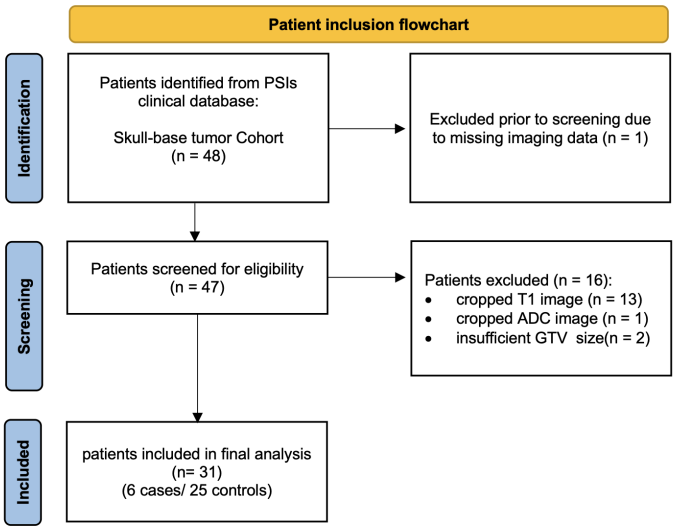

Figure 1: Prisma diagram visualizing identification and screening procedures for the patient selection of the final study cohort.

2.2 Radiomics Pipeline

To extract radiomics features from multiple imaging modalities, all images were registered to the treatment planning CT, on which the GTVs were defined. This ensured anatomical consistency of the radiomics features across modalities. Since both T1-weighted and FLAIR MRI images contain anatomical information, they could be directly registered to the CT. However, DWI images pose a greater challenge due to the functional nature of their intensity information. Therefore, a dedicated image registration pipeline was implemented to address this issue as an extension of a previously developed analysis framework[10].

Spatial alignment

Since patient positioning on the scanner table can vary considerably between imaging modalities, registration required the estimation of an appropriate initial condition. TotalSegmentator[11] was used to generate binary brain masks from the T1-weighted, FLAIR, and CT images, which were subsequently converted to signed distance fields (SDF). Rigid registration of the moving T1 and FLAIR SDFs to the fixed CT SDF was performed to obtain initial transforms. These initial transforms were then used as starting points for affine registration of the T1-weighted and FLAIR images to the CT, using normalised mutual information as the similarity metric.

For ADC maps, additional processing was required to enable registration. A lower and upper threshold of $0.3 \cdot 10^{-3} mm^2/s$ and $1.5 \cdot 10^{-3} mm^2/s$ was applied, and values were normalized to this range. By inverting the normalized intensities, the ADC maps were made to resemble T1-weighted MRI images, allowing the previously described registration procedure to be applied. Finally, the ADC-T1 and T1-CT transforms were concatenated to obtain the ADC-CT transform. This registration procedure is presented in fig. 2. All registration steps were implemented in Python using SimpleITK library[12].

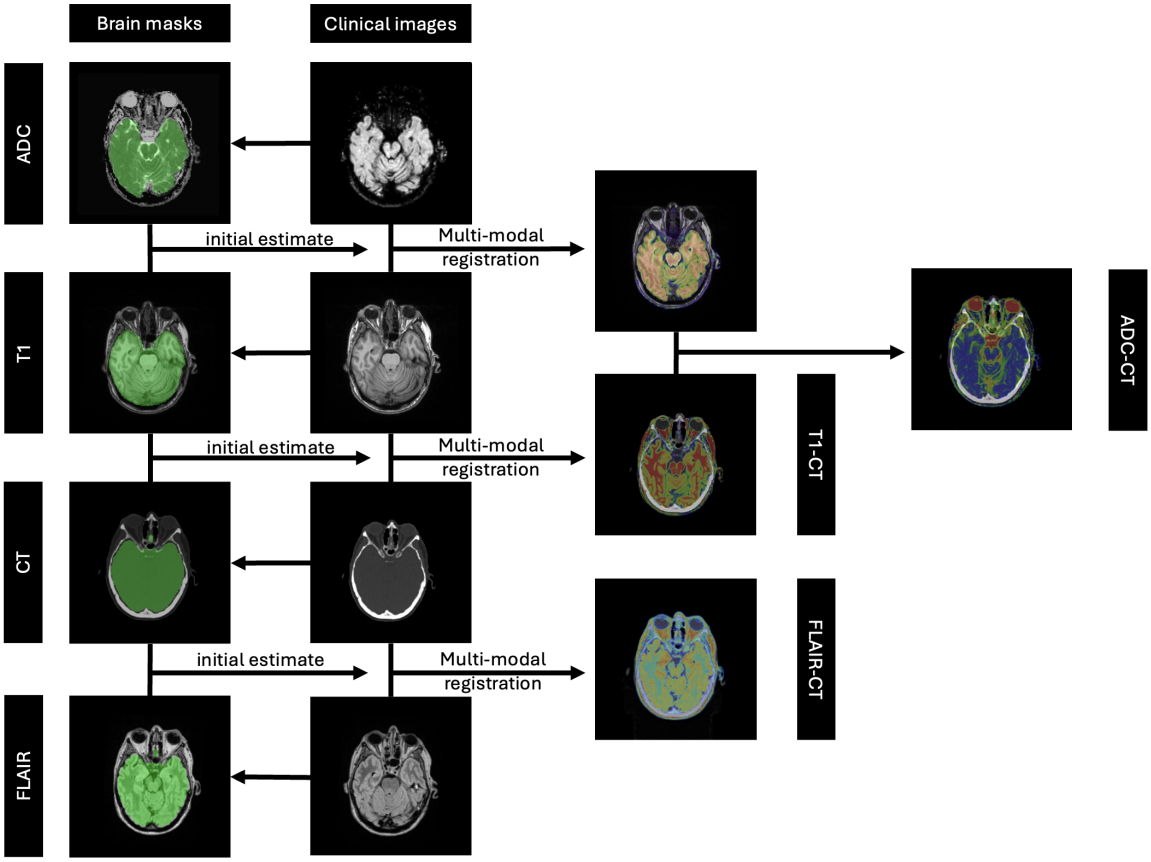

Figure 2: Image registration steps of ADC, T1-weighted, and FLAIR images to the CT reference. ADC images were first converted to anatomical-like surrogates to enable reliable registration.

To extract the image intensities from the GTV region of interest, corresponding masks were generated based on the treatment planning contours and converted into binary label masks. Since the tumors in this study were located at the skull base, the GTVs often included regions containing air pockets or bony structures, which could bias the radiomics analysis. To mitigate this, air and bone masks were generated and subsequently subtracted from the GTVs. Air masks were generated by thresholding the CT image between –1024 and –500 HUs, corresponding to voxels containing air. Bone masks were generated using TotalSegmentator. Subtracting these regions ensured that only soft-tissue voxels corresponding to the tumor were included in the radiomics analysis.

Resampling and Intensity Harmonization

A major challenge when extracting quantitative imaging biomarkers (QIB) and radiomics features from MRI images is their standardization across datasets. MRI signal intensities are not expressed in physical units and strongly depend on imaging parameters[13]. To address this issue, all images were normalized prior to analysis. To account for different voxel spacings, images and corresponding masks were resampled to a common voxel spacing

of (1mm,1mm,2mm) prior to normalization. For ADC maps derived from the DWI, robust z-scoring was performed using the median and median absolute deviation (MAD):

$$\mathrm{ADC}_{\mathrm{normalized}} = 0.6745 \cdot \frac{\mathrm{ADC} - \mathrm{ADC}_{\mathrm{median}}}{\mathrm{MAD}} \tag{3}$$

$\mathrm{ADC}_{\mathrm{median}}$ corresponds to the median ADC within the brain (excluding GTV) and MAD = Median($|\mathrm{ADC} - \mathrm{ADC}_{\mathrm{median}}|$) is the median absolute deviation[10].

T1-weighted and FLAIR MRI images were instead normalized to normal appearing white matter (NAWM). This region provides a homogeneous reference tissue, reducing the influence of intensity variations caused by tumors or edema [13]. NAWM masks were obtained using a component of the FMRIB (also known as FSL) toolkit, namely the Automated Segmentation Tool(FAST)[14]. The resulting masks were then used to calculate the median and MAD to normalize T1-weighted and FLAIR MRI images according to eq. (3) above.

2.3 Feature extraction

Image intensities from the GTV were used to extract radiomics features and to analyze their distribution across the cohort. In addition, GTV-derived metrics were investigated in the context of survival analysis for local treatment failure. Feature extraction was performed using PyRadiomics[15]. For all features, PyRadiomics ensures compliance with the Image Biomarker Standardisation Initiative (IBSI) guidelines or uses equivalently validated definitions[16].

To improve stability of radiomics features across the dataset, a fixed number of 128 bins was used for intensity discretization instead of a fixed bin size approach[17]. For each of the three imaging modalities, 53 radiomics features were extracted, restricting the analysis to *first-order Features*, *Gray Level Size Zone Matrix(GLSZM)*, *Gray Level Dependence Matrix(GLDM)*, and *Neighboring Gray Tone Difference Matrix (NGTDM)* features. These features capture image intensity distributions and texture patterns. Shape features were excluded, as the removal of bone and air from the GTV masks rendered geometric measurements unreliable.

2.4 Survival Analysis

The extracted and harmonized radiomics features from baseline imaging, together with patient demographic data, and clinical progression from follow-up consultations, were consolidated into a final dataset of covariates for survival analysis. The Cox proportional hazards model was used to model the relationship between the covariates and the outcome, appropriately accounting for the different observed events. All survival analyses were performed using scikit-survival package [18].

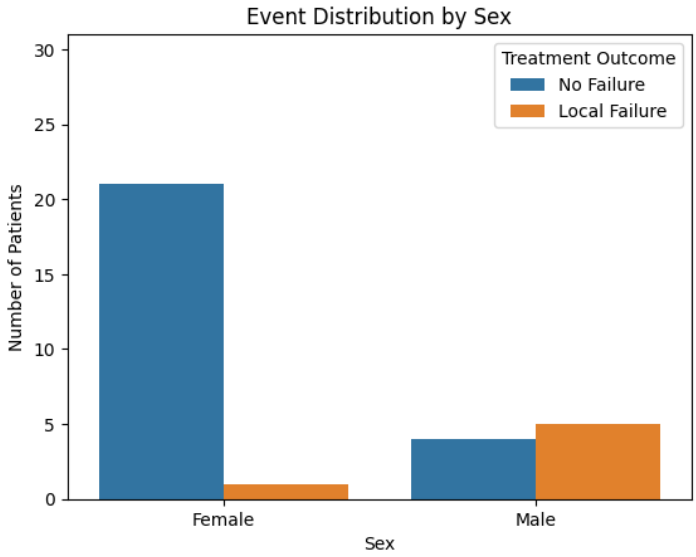

Figure 3: Distribution of local treatment failure events stratified by sex.

Data Preparation

Given the large number of radiomics features extracted (n = 159), a meaningful subset was first selected. Constant features that provided no information for the analysis were excluded. Subsequently, following the approach described by Schroeder et al. [19], highly correlated features were then grouped together using Spearman's rank correlation coefficient ($r > 0.8$). Within each group, the feature showing the strongest correlation with local treatment failure was retained for further processing, while the others were removed. This collinearity reduction step, commonly applied to minimise bias from highly correlated variables in survival analysis, reduced the feature pool to 50 relevant ones.

Among clinical variables, sex was excluded despite its apparent prognostic value, as shown in Figure 3, which was attributed to small-sample effects rather than true statistical significance. Consequently, the final dataset used for the analysis included 50 radiomics features and the age at the start of proton therapy as the only clinical variable. These remaining features were used as parameters in a Cox proportional hazards model.

Cox Proportional Hazards Model

To evaluate the robustness of each parameter, 2000 bootstrap samples of the data set were generated following an internal validation approach. For each sample, a penalized Cox regression model was fitted using the least absolute shrinkage and selection operator(LASSO) penalty. The regularization parameter was tuned so that each model used only five of the 51 parameters at a time. Subsequently, the selection frequency of each parameter, i.e. the proportion of bootstrap models in which a given parameter was selected, was calculated.

To determine the optimal number of parameters for the final model, Cox regression models were fitted with an iteratively increasing number of covariates. Parameters were forward added, in descending order of selection frequency, until a total of 15 features were included.

Model performance was measured using leave-one-out cross-validation (LOOCV) and the concordance index (C-Index).

The optimal number of parameters included in the final model was determined by assessing the variation in the C-index as a function of the number of included parameters and visually identifying the point at which the C-Index no longer improved significantly. The performance of the model was further evaluated using LOOCV and the Integrated Brier Score (IBS), which evaluates the overall accuracy and calibration of survival predictions. The maximal follow-up time used to compute the IBS was set to the longest observed event time to avoid extrapolation beyond the observed data.

To investigate the clinical relevance of the predicted risk, a final model was fitted to the complete dataset. Patients were grouped into tertiles based on predicted risk, and Kaplan-Meier survival curves were generated to visualize differences between the groups.

3 Results

3.1 Imaging Intensity Analysis

For each imaging modality, the median, skewness, and kurtosis of the normalized voxel intensity distributions were calculated across the cohort. Subsequently, grouped by treatment outcome, the minimum, maximum, median, and interquartile range (IQR) values were computed across each group. Results are presented in tables 1, 2 and 3.

Notable differences were observed in the normalized ADC distributions between patients with local failure (LF) and those in the control group (no LF). The IQR of the median ADC is considerably larger in the control group (5.39) than in the LF group(0.517), indicating greater inter-patient variability in the control group. Similarly, the median ADC is higher in the control group (5.16) compared to the LF group (2.89). The LF group (0.685) also showed a right-skewed distribution, suggesting a lot of low ADC values and a few high-value outliers. The median kurtosis, which describes how peaked a distribution is, is also higher in the LF group, supporting the observation of clustering of values around the mean.

Table 1: Summary statistics of normalized ADC distribution in the GTV, grouped by treatment outcome.

	Median ADC		Skewness ADC		Kurtosis ADC	
	no LF	LF	no LF	LF	no LF	LF
Min	0.663	1.732	-1.107	-0.507	-1.348	-0.612
Max	11.104	16.040	0.708	1.377	2.168	1.864
Median	5.164	2.889	0.102	0.685	-0.113	0.448
IQR	5.394	0.517	0.568	1.176	0.982	1.443

The median of the median FLAIR values is higher for the LF group (0.54) compared to the control group(-0.12). A smaller IQR in the LF group (0.80 vs. 1.77) suggests reduced inter-patient variability. However, the magnitude of these differences is small, and the distributions of skewness and kurtosis are largely comparable between the groups.

Table 2: Summary statistics of normalized FLAIR distribution in the GTV, grouped by treatment outcome.

	Median FLAIR		Skewness FLAIR		Kurtosis FLAIR	
	no LF	LF	no LF	LF	no LF	LF
Min	-3.726	-2.360	-1.277	-1.235	-1.120	-1.325
Max	2.265	1.984	1.030	0.477	2.457	0.988
Median	-0.124	0.542	-0.165	-0.267	-0.527	-0.551
IQR	1.769	0.803	0.762	0.521	0.484	0.596

Comparison of distributions of T1 values of control and LF groups shows lower median T1 values (resp. -2.60 and -1.29) and larger IQR (resp. 1.70 and 0.69). Kurtosis is higher in the LF group (1.56 vs. 0.94), suggesting a more peaked intensity distribution and tighter clustering of voxel intensities around the mean.

Table 3: Summary statistics of normalized T1 distribution in the GTV, grouped by treatment outcome..

	Median T1		Skewness T1		Kurtosis T1	
	no LF	LF	no LF	LF	no LF	LF
Min	-5.987	-4.785	-0.554	-1.378	-1.350	-0.637
Max	3.965	0.450	2.671	1.163	12.954	4.124
Median	-2.600	-1.290	0.655	0.434	0.935	1.555
IQR	1.703	0.687	1.731	0.339	5.465	2.032

3.2 Radiomics Feature Analysis

Figures 4 to 6 present the Spearman's rank correlation coefficients between the extracted radiomics features and the treatment outcome for each imaging modality. To improve readability, only features with a p-value below 0.2 are shown, together with their 95% confidence intervals, as including all features would make the figures overly crowded.

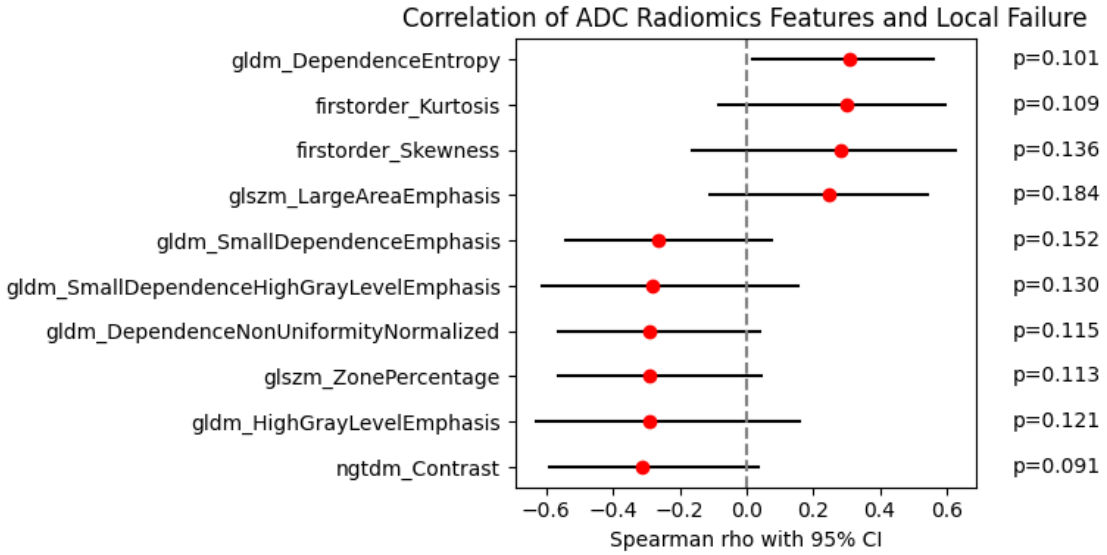

Figure 4: Spearman's rank correlations between ADC radiomics features and local treatment failure, including 95% confidence intervals.

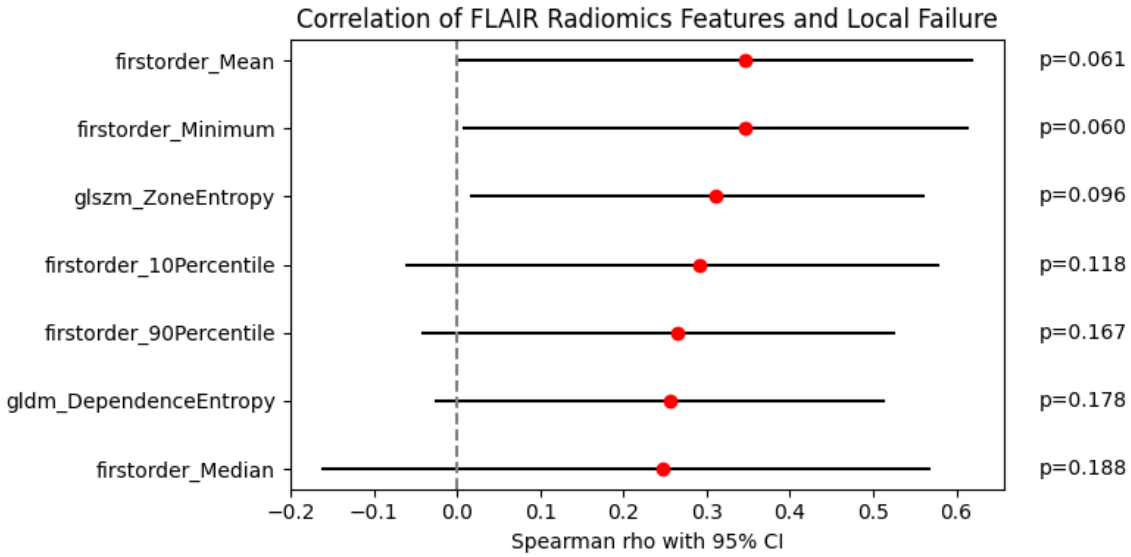

Figure 5: Spearman's rank correlations between FLAIR radiomics features and local treatment failure, including 95% confidence intervals.

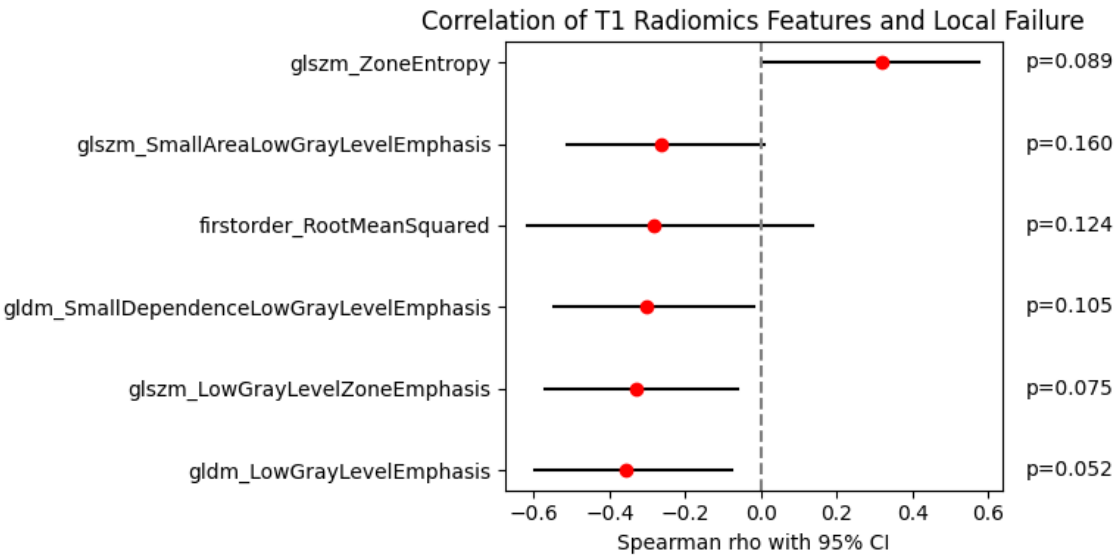

Figure 6: Spearman's rank correlations between T1 radiomics features and local treatment failure, including 95% confidence intervals.

3.3 Survival Modeling Performance

Figure 7 illustrates the clustering of radiomics features based on Spearman's rank correlation coefficient. Features with correlation coefficients > 0.8 were grouped and visualized with the same color, revealing clusters of highly correlated features. This clustering was used in the feature filtering step, where only one representative feature from each cluster was retained for survival analysis.

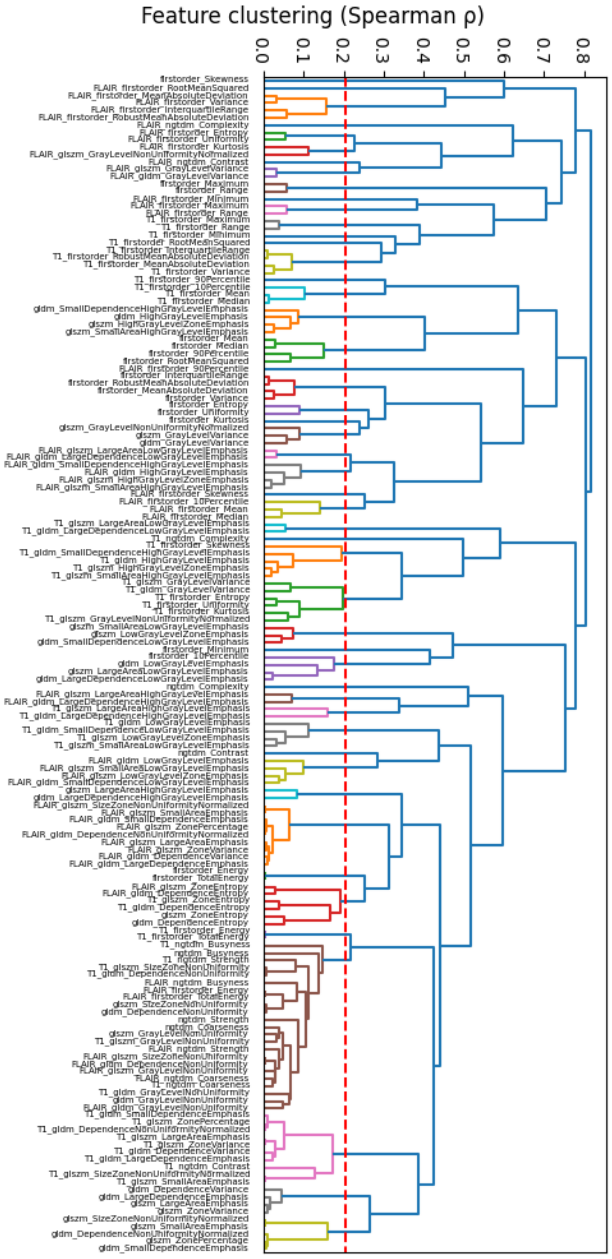

Figure 7: Clustering of radiomics features based on Spearman's rank correlation coefficient. Features with a correlation coefficient > 0.8 are visualized with the same color.

Table 4 displays the performance of the Cox regression model evaluated using LOOCV. It lists the radiomics features and clinical parameters that were iteratively added from the dataset obtained after filtering out collinear features. For each iteration, the selection frequency, the C-Index, and the integrated Brier score (IBS) are reported. The C-index increases substantially with the inclusion of up to five features, after which the model performance begins to plateau. The IBS is the lowest when using two parameters, then fluctuates and after the inclusion of 6 parameters increases slightly.

Table 4: Model performance (C-Index) and calibration quality (IBS) as features were iteratively added in descending order of selection frequency.

#	Feature Added	Selection Frequency	C-Index	IBS
1	firstorder_Skewness	0.569	0.696	0.078
2	FLAIR_firstorder_90Percentile	0.524	0.810	0.077
3	FLAIR_firstorder_Minimum	0.330	0.842	0.089
4	firstorder_Kurtosis	0.313	0.842	0.100
5	FLAIR_gldm_DependenceEntropy	0.225	0.911	0.089
6	T1_glszm_LowGrayLevelZoneEmphasis	0.208	0.892	0.084
7	glszm_LargeAreaEmphasis	0.181	0.892	0.109
8	gldm_DependenceNonUniformityNormalized	0.176	0.924	0.109
9	Age_at_PT_years	0.145	0.918	0.117

The final Cox proportional hazards model included the five most robust features identified through the bootstrap selection process as listed in table 4. The model estimated the hazard function $(t|X)$ as

$$h(t|X) = h_0(t) \cdot \exp\left(\sum_{i=1}^{5} \beta_i X_i\right) \tag{4}$$

where $h_0(t)$ denotes the baseline hazard, and X_i represent the feature values. The corresponding estimated regression coefficients (β) and hazard ratios $(HR = e^{\beta})$ are summarized in table 5.

Table 5: Estimated coefficients (β) and hazard ratios (HR) of the final CPH model.

Feature	Coefficient (β)	Hazard Ratio (HR $= e^{\beta}$)
ADC_firstorder_Skewness	0.82	2.27
FLAIR_firstorder_90Percentile	0.54	1.71
FLAIR_firstorder_Minimum	0.37	1.45
ADC_firstorder_Kurtosis	0.21	1.23
FLAIR_gldm_DependenceEntropy	0.63	1.88

describe!!!

regression model for survival analysis is the Cox Proportional Hazards

The output of the CPH model is the hazard ratio (HR), which quantifies how the risk of experiencing the event of interest differs between two groups at any time point.

The Kaplan-Meier curves in fig. 8 illustrate local failure-free survival of patients split into three risk groups based on tertiles of their predicted risk scores from the CPH model.

Within the first year after the start of proton therapy, one local failure was observed in the high-risk group, while all other patients remained local failure free. At two years, the estimated LF-free survival probability decreased to approximately 90% in the high-risk group, while remaining at 100% for the intermediate and low-risk group. After five years, the probability of LF-free survival further decreased to around 40% in the high-risk group, while it remained between 90% and 100% for the other groups.

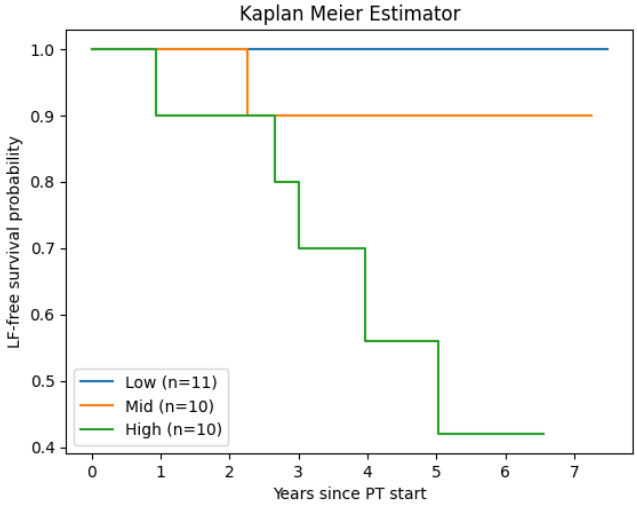

Figure 8: Kaplan-Meier curves illustrating the local failure-free survival probabilities for patients stratified into three risk groups based on predicted risk.

4 Discussion

4.1 Interpretation of the Results

Imaging biomarker analysis

In this study, patients who experienced local treatment failure (LF) exhibited a lower normalized baseline ADC within the GTV compared to those from the progression-free survival group. The research by Chen et al. demonstrated a strong inverse correlation between ADC and cellularity in the brain[20]. This indicates a more restricted diffusion of water molecules and higher cellularity in LF patients. Increased cellularity in turn, has been associated with higher-grade tumor and more aggressive biological behavior, which is consistent with the observation of lower ADC values in the LF patient group[21].

The higher kurtosis observed in patients with LF (0.45 vs. -0.11) suggests a more peaked and homogeneous distribution of normalized ADC values compared with the control group. This pattern may reflect a more uniformly dense tissue structure, consistent with restricted diffusion and increased cellularity in patients who eventually experienced relapse. A comparable increase was also observed in T1-weighted images of LF patients, further supporting the interpretation of a more uniform and densely structured tumor mass.

Similar relationships have been reported between elevated ADC kurtosis and malignancy in other types of tumors. For example, in breast lesions, malignant tumors exhibited significantly higher ADC kurtosis compared to benign lesions[22].

Median FLAIR values were higher in the LF group (-0.12 vs. 0.542). Bette et al. found that an increase in FLAIR signal intensity within the resection cavity in patients with brain metastases was a specific indicator of tumor progression and recurrence[23]. Although observed in a different cancer type, the elevation in FLAIR intensities in patients with LF in this study may also reflect tumor progression or recurrence pattern.

For all distributions the IQR of the median ADC values, which represents inter-patient variability, is notably smaller in the LF group, most notably in the median ADC values (0.52 vs. 5.39). However, this should be interpreted with caution and is likely due to the unbalanced dataset, which includes only 6 LF cases.

Radiomics features association with local failure

Among the extracted radiomics features, the most frequently selected descriptors of local failure were extracted from ADC and FLAIR images, less so for common T1 anatomical imaging.

The positive correlation of the ADC first-order kurtosis and skewness features with LF is consistent with previously observed differences in ADC distributions. Similarly, positive correlations between FLAIR first-order 90th percentile and FLAIR first-order minimum features suggest an overall elevation of FLAIR signal intensities in patients with local failure. This observation aligns with the increased FLAIR median intensities reported in Table 2.

In addition to first-order features, texture-based radiomics features from FLAIR images also showed associations with local failure. Of interest are Gray Level Dependence Matrix (GLDM) features, describing how much neighboring voxels depend on each other's intensity. The GLDM dependence entropy feature then quantifies the complexity and randomness of these intensities. Their high values indicate greater local heterogeneity in the FLAIR intensity texture. A positive correlation between this feature and local failure suggests that baseline FLAIR images of LF patients show more heterogeneous intensity patterns.

This observation was further supported by the positive correlation of Gray Level Size Zone Matrix (GLSZM) zone entropy, which measures the randomness of homogeneous zones. Higher zone entropy values indicate a wide range of small and large homogeneous regions with varying intensities, supporting increased complexity in baseline FLAIR images in the LF group.

Despite the large number of covariates (n = 51) included in the robustness analysis, the selection frequency was relatively high for the top ranking parameters, particularly for the first two (0.569 and 0.524), indicating consistent emerging pattern across bootstrapped dataset.

Model Performance

The discriminative performance of the model, evaluated using the C-Index, increased substantially as features were iteratively added. However, when more than five features were included, the C-Index reached a plateau. The integrated Brier score(IBS), which reflects the calibration of the model and decreases with better performance, was lowest when using two

parameters. As three more additional parameters were included, the IBS slightly increased, but remained nearly as low with five parameters as with two. Overall, these findings indicate that a model incorporating the five parameters with the highest selection frequency may be a resonable tradeoff between overfitting and discriminative performance or calibrationn.

The Kaplan-Meier curves illustrate a clear separation between the three risk groups derived from the CPH model. Patients in the high-risk group show markedly lower LF-free survival probabilities, with several events already occurring in the first three years after proton therapy.

The good separation between the survival curves shows that the five selected parameters capture relevant tumor characteristics associated with local treatment failure and supports their potential predictive power.

4.2 Limitations

Several limitations should be considered when interpreting the results of this study. First, the small cohort size (n = 31) limits the statistical power of the analysis and increases the risk of overfitting the predictive models. In addition, the study lacked a baseline Cox model that included only clinical variables, which would have allowed a direct comparison of the added predictive value of the radiomics features. Furthermore, no dosimetric information was available, preventing an assessment of the dose-response relationships.

The spatial registration pipeline was extended from previous work to include refined data masking and normalization procedures. It showed good visual alignment, but minor mistakes in registrations between modalities, especially between DWI and CT, could still introduce small misalignments in the boundary regions of tissues. Furthermore, air/bone removal steps may have altered tumor shape, leading to the exclusion of shape-based radiomic characteristics from the analysis.

The feature selection and modeling steps were also affected by the limited dataset size. The bootstrap based robustness analysis provided a measure of feature stability, but the small sample size and large number of features (51 covariates) presented a risk of model instability. Although cross-validation was used, external validation on an independent test dataset would be required to confirm generalizability and calibration.

Finally, the biological interpretation of the intensity distributions and radiomics features should be interpreted with caution. While associations between ADC, FLAIR, and the radiomics texture features with local failure are biologically plausible, the limited cohort size and the sometimes contradictory literature prevent definitive conclusions about tissue characteristics.

4.3 Future Work

Validation in larger and more balanced datasets will be essential to confirm these findings and improve the generalizability of the model. Future research should also aim to explain the biological mechanisms underlying the radiomics features associated with local failure. An option would be the inclusion of histological markers, such as cellularity, for biomarker validation. Since the risk estimates from the CPH model are relative to a study cohort, subsequent studies should define absolute radiomics feature thresholds that allow risk classification.

5 Conclusion

In this study, a survival analysis was performed on patients with skull base chordoma and chondrosarcoma treated with proton therapy. The analysis included diffusion-weighted derived biomarkers together with conventional anatomical T1-weighted, and FLAIR images, and a few available clinical parameters. The proposed pipeline enabled multimodal image registration, radiomics feature extraction and survival analysis. It addressed challenges such as differing information content of functional and anatomical imaging modalities, image intensity normalization, and high-dimensional feature data.

The best-fitting Cox proportional hazards model, which included five parameters, achieved good risk stratification, as confirmed by discrimination (C-Index) and calibration metrics (Brier score) as well as by Kaplan-Meier analysis. These findings suggest that radiomics features from quantitative and anatomical magnetic resonance images capture biologically relevant information associated with treatment outcomes.

Future studies with a larger and more balanced dataset are needed to confirm these findings and to reliably identify imaging biomarkers to help make clinical decisions.

References

[1] Dipesh Niraula et al. "Current status and future developments in predicting outcomes in radiation oncology". In: *The British Journal of Radiology* 95.1139 (Nov. 2022), p. 20220239. ISSN: 0007-1285. DOI: 10.1259/bjr.20220239. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9793488/ (visited on 27/10/2025).

[2] Kaat Spoormans et al. "A Review on Tumor Control Probability (TCP) and Preclinical Dosimetry in Targeted Radionuclide Therapy (TRT)". In: *Pharmaceutics* 14.10 (Sept. 2022), p. 2007. ISSN: 1999-4923. DOI: 10.3390/pharmaceutics14102007. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9608466/ (visited on 25/10/2025).

[3] Georgios Beis, Aggelos Iliopoulos and Ioannis Papasotiriou. "An Overview of Introductory and Advanced Survival Analysis Methods in Clinical Applications: Where Have we Come so far?" en. In: *Anticancer Research* 44.2 (Feb. 2024), pp. 471–487. ISSN: 0250-7005, 1791-7530. DOI: 10.21873/anticanres.16835. URL: https://ar.iiarjournals.org/content/44/2/471 (visited on 27/10/2025).

[4] *sksurv.metrics.concordance_index_censored — scikit-survival 0.25.0*. URL: https://scikit-survival.readthedocs.io/en/stable/api/generated/sksurv.metrics.concordance_index_censored.html (visited on 24/10/2025).

[5] Abeer M. Aldawsari et al. "The role and potential of using quantitative MRI biomarkers for imaging guidance in brain cancer radiotherapy treatment planning: A systematic review". In: *Physics and Imaging in Radiation Oncology* 27 (July 2023), p. 100476. ISSN: 2405-6316. DOI: 10.1016/j.phro.2023.100476. URL: https://www.sciencedirect.com/science/article/pii/S2405631623000672 (visited on 28/10/2025).

[6] Ji Eun Park and Ho Sung Kim. "Radiomics as a Quantitative Imaging Biomarker: Practical Considerations and the Current Standpoint in Neuro-oncologic Studies". In: *Nuclear Medicine and Molecular Imaging* 52.2 (Apr. 2018), pp. 99–108. ISSN: 1869-3474. DOI: 10.1007/s13139-017-0512-7. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5897262/ (visited on 27/10/2025).

[7] Camille Decoene and Frederik Crop. "Using density computed tomography images for photon dose calculations in radiation oncology: A patient study". In: *Physics and Imaging in Radiation Oncology* 27 (July 2023), p. 100463. ISSN: 2405-6316. DOI: 10.1016/j.phro.2023.100463. URL: https://www.sciencedirect.com/science/article/pii/S2405631623000544 (visited on 29/10/2025).

[8] Mai-Lan Ho, Rafael Rojas and Ronald L. Eisenberg. "Cerebral Edema". en. In: *American Journal of Roentgenology* 199.3 (Sept. 2012), W258–W273. ISSN: 0361-803X, 1546-3141. DOI: 10.2214/AJR.11.8081. URL: https://www.ajronline.org/doi/10.2214/AJR.11.8081 (visited on 01/11/2025).

[9] Marco Bologna et al. "Relevance of apparent diffusion coefficient features for a radiomics-based prediction of response to induction chemotherapy in sinonasal cancer". en. In: *NMR in Biomedicine* 35.4 (Apr. 2022), e4265. ISSN: 0952-3480, 1099-1492. DOI: 10.1002/nbm.4265. URL: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/nbm.4265 (visited on 28/10/2025).

[10] Chandra Tresch. "Development and exploratory testing of an analysis pipeline for ADC-derived radiomics biomarkers in brain tumor patients". In: (July 2025).

[11] Jakob Wasserthal et al. *TotalSegmentator: robust segmentation of 104 anatomical structures in CT images*. arXiv:2208.05868. June 2023. DOI: 10.48550/arXiv.2208.05868. URL: http://arxiv.org/abs/2208.05868 (visited on 27/10/2025).

[12] Bradley Christopher Lowekamp et al. "The Design of SimpleITK". English. In: *Frontiers in Neuroinformatics* 7 (Dec. 2013). ISSN: 1662-5196. DOI: 10.3389/fninf.2013.00045. URL: https://www.frontiersin.org/journals/neuroinformatics/articles/10.3389/fninf.2013.00045/full (visited on 23/10/2025).

[13] Alexandre Carré et al. "Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics". en. In: *Scientific Reports* 10.1 (July 2020), p. 12340. ISSN: 2045-2322. DOI: 10.1038/s41598-020-69298-z. URL: https://www.nature.com/articles/s41598-020-69298-z (visited on 22/10/2025).

[14] Y. Zhang, M. Brady and S. Smith. "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm". eng. In: *IEEE transactions on medical imaging* 20.1 (Jan. 2001), pp. 45–57. ISSN: 0278-0062. DOI: 10.1109/42.906424.

[15] Joost J.M. Van Griethuysen et al. "Computational Radiomics System to Decode the Radiographic Phenotype". en. In: *Cancer Research* 77.21 (Nov. 2017), e104–e107. ISSN: 0008-5472, 1538-7445. DOI: 10.1158/0008-5472.CAN-17-0339. URL: https://aacrjournals.org/cancerres/article/77/21/e104/662617/Computational-Radiomics-System-to-Decode-the (visited on 23/10/2025).

[16] *Frequently Asked Questions — pyradiomics v3.1.0rc2.post5+g6a761c4 documentation*. URL: https://pyradiomics.readthedocs.io/en/latest/faq.html?highlight=ibsi (visited on 25/10/2025).

[17] Zahra Khodabakhshi et al. "Magnetic resonance imaging radiomic features stability in brain metastases: Impact of image preprocessing, image-, and feature-level harmonization". eng. In: *Physics and Imaging in Radiation Oncology* 30 (Apr. 2024), p. 100585. ISSN: 2405-6316. DOI: 10.1016/j.phro.2024.100585. URL: https://www.zora.uzh.ch/handle/20.500.14742/220690 (visited on 25/10/2025).

[18] *scikit-survival — scikit-survival 0.25.0*. URL: https://scikit-survival.readthedocs.io/en/stable/ (visited on 25/10/2025).

[19] Christina Schröder et al. "NTCP Modeling for High-Grade Temporal Radionecroses in a Large Cohort of Patients Receiving Pencil Beam Scanning Proton Therapy for Skull Base and Head and Neck Tumors". eng. In: *International Journal of Radiation Oncology, Biology, Physics* 113.2 (June 2022), pp. 448–455. ISSN: 1879-355X. DOI: 10.1016/j.ijrobp.2022.01.047.

[20] Lihua Chen et al. "The Correlation between Apparent Diffusion Coefficient and Tumor Cellularity in Patients: A Meta-Analysis". In: *PLoS ONE* 8.11 (Nov. 2013), e79008. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0079008. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3823989/ (visited on 30/10/2025).

[21] Ehsan Hassannejad et al. "Correlation of ADC values of adult brain tumors with the diagnosis and pathological grade: A cross-sectional multicenter study". en. In: *Health Science Reports* 7.6 (June 2024), e2110. ISSN: 2398-8835, 2398-8835. DOI: 10.1002/hsr2.2110. URL: https://onlinelibrary.wiley.com/doi/10.1002/hsr2.2110 (visited on 30/10/2025).

[22] Dina EL-Metwally, Amina Ahmed Ahmed Sultan and Shohanda Abdelmonem Mahmoud Eltelbany. "Histogram analysis in diffusion-weighted imaging in differentiating breast masses". In: *Egyptian Journal of Radiology and Nuclear Medicine* 55.1 (Dec. 2024), p. 240. ISSN: 2090-4762. DOI: 10.1186/s43055-024-01415-8. URL: https://doi.org/10.1186/s43055-024-01415-8 (visited on 30/10/2025).

[23] Stefanie Bette et al. “Increase in FLAIR Signal of the Fluid Within the Resection Cavity as Early Recurrence Marker: Also Valid for Brain Metastases?” eng. In: *RoFo: Fortschritte Auf Dem Gebiete Der Rontgenstrahlen Und Der Nuklearmedizin* 189.1 (Jan. 2017), pp. 63–70. ISSN: 1438-9010. DOI: 10.1055/s-0042-119686.