Appendix A Supplementary Material

A.1 Dataset

The AURORA trial was conducted with the *Radiosurgery and Stereotactic Radiotherapy Working Group of the German Society for Radiation Oncology*(DEGRO). In total, seven hospitals contributed to the data collection: TUM: Klinikum rechts der Isar of the Technical University of Munich, USZ: University Hospital of Zurich, FD: General Hospital Fulda, FFM: Saphir Radiochirurgie/University Hospital Frankfurt, FR: University Hospital Freiburg, HD: Heidelberg University Hospital, KSA: Kantonsspital Aarau. The eligibility criteria to include patients in the study are shown in Figure A1.

Ethical approval was obtained from each participating institution, with the primary license granted at the Technical University of Munich under reference 119/19 S-SR. No informed consent for the research study was necessary as the retrospective analysis of patient records and data is generally allowed following Article 27 of the Bavarian Hospital Act (*Bayerisches Krankenhausgesetz*) from the law *Landeskrankenhausgesetz des Freistaates Bayern*. Informed consent for treatment was obtained from every patient.

Based on prior findings showing an area under the curve (AUC) of 0.79 for LF prediction in a single-center study with a 15% event rate [8], the minimum sample size for the test set was determined as 55 patients. To increase heterogeneity, we augmented the test set with data from multiple smaller centers. This resulted in a training cohort of 253 individuals from two centers (TUM, USZ) and an external, multi-center international test group of 99 patients from five centers (FD, FFM, FR, HD, and KSA). The training cohort was divided into 201 training and 52 validation cases for each five-fold cross-validation.

In the training and test cohorts, five and 29 patients received stereotactic radiosurgery (median dose 20 Gy and 16 Gy, respectively), while 248 and 70 patients underwent fractionated SRT (median seven fractions of 5 Gy in the training cohort and six fractions of 5 Gy in the test cohort). Table A2 summarizes all prescribed dose and fraction combinations. We also show the full list of MRI devices used in the data acquisition and their presence in the training and test cohorts in Table A1.

Among the 352 patients, 52 experienced LF, resulting in a highly imbalanced event rate of 15%. The FFLF exhibited a long-tail distribution, with 31.5% of patients having follow-up times under six months. The longest LF-free duration was 111 months, while the earliest censoring occurred after two days. LF was first observed at three days and last observed at 101 months.

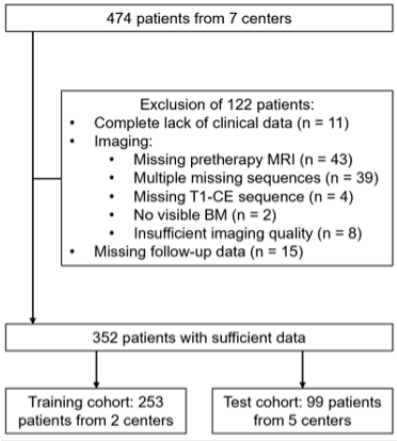

Fig. A1: Flowchart visualizing the total number of patients obtained in the trial, the inclusion criteria for our study, as well as the final partitions

Vendor	Scanner	Train	Test
GE	Discovery MR750w	✓	✓
	Optima MR360	✗	✓
	Optima MR450w	✗	✓
	Signa Excite	✗	✓
	Signa Explorer	✓	✗
	Signa Voyager	✓	✗
	Signa HDxt	✗	✓
	Signa PET MR	✗	✓
Philips	Achieva	✓	✓
	Achieva dStream	✓	✓
	Ingenia	✓	✓
	Intera	✓	✓
	Panorama HFO	✓	✓
Siemens	Aera	✓	✓
	Avanto	✓	✓
	Avanto fit	✓	✓
	Espree	✓	✓
	Harmony	✓	✗
	HarmonyExpert	✓	✓
	Magnetom Vida	✓	✗
	Skyra	✓	✓
	Prisma fit	✓	✗
	Symphony	✓	✓
	SymphonyTim	✓	✗
	TrioTim	✗	✓
	Verio	✓	✓
Toshiba	MRT200SP8	✗	✓

Table A1: List of vendors and MRI scanners used to acquire imaging data in the training and test cohorts.

Fraction × Dose (in Gray)	**Train** ($N = 253$)	**Test** ($N = 99$)
1 × 12 Gy	0 (0%)	5 (5.1%)
1 × 13 Gy	0 (0%)	1 (1.0%)
1 × 14 Gy	0 (0%)	6 (6.1%)
1 × 15 Gy	0 (0%)	1 (1.0%)
1 × 16 Gy	0 (0%)	6 (6.1%)
1 × 17 Gy	0 (0%)	3 (3.0%)
1 × 18 Gy	1 (0.4%)	6 (6.1%)
1 × 20 Gy	4 (1.6%)	1 (1.0%)
3 × 7 Gy	0 (0%)	1 (1.0%)
3 × 8 Gy	0 (0%)	3 (3.0%)
5 × 5 Gy	3 (1.2%)	8 (8.1%)
5 × 6 Gy	0 (0%)	16 (16%)
6 × 4 Gy	1 (0.4%)	0 (0%)
6 × 5 Gy	91 (36%)	22 (22%)
7 × 5 Gy	153 (60%)	14 (14%)
13 × 3 Gy	0 (0%)	5 (5.1%)
14 × 3 Gy	0 (0%)	1 (1.0%)

Table A2: Fractions and radiation doses in Gray (Gy) in this study. Five and 29 patients were treated with stereotactic radiosurgery (SRT) in the training and test cohorts. 248 and 70 were treated with fractioned SRT.

Chemotherapy Agents	**Train** ($N = 15$)	**Test** ($N = 3$)
5-Fluorouracil+Folinic Acid/Oxaliplatin	1	0
Bicalutamid/Leuprorelin	1	0
Capecitabine	1	0
Carboplatin/Gemcitabine	1	1
Cisplatin/Pemetrexed	5	2
Cisplatin/Vinorelbin	2	0
Enzalutamid	1	0
Vinorelbin	1	0
Unknown	2	0

Table A3: Chemotherapy (CTX) agents. In total, 18 patients were treated with concurrent CTX

Immunotherapy Agents	**Train** ($N = 10$)	**Test** ($N = 13$)
Alectinib	0	1
Axitinib	1	0
Cabozantinib	1	0
Crizotinib	0	1
Denosumab	0	1
Erlotinib	1	0
Ipilimumab/Nivolumab	1	1
Lapatinib	1	0
Nivolumab	1	0
Palbociclib/Anastrozol	0	1
Pembrolizumab	0	5
Tamoxifen	1	0
Trastuzumab	1	0
Trastuzumab/Pertuzumab	0	1
Vemurafenib	1	0
Vemurafenib/Trametinib	1	0
Vinorelbin/Trastuzumab/Pertuzumab	0	1
Unknown	0	1

Table A4: Immunotherapy (ITX) agents. In total, 23 patients were treated with concurrent ITX

Appendix B Supplementary Methods

B.1 Optimization

The momentum parameters of the AdamW optimizer were $\beta_1 = 0.9$ and $\beta_2 = 0.999$. A *weight decay* parameter of 10^{-3} was used as a regularization technique to avoid overfitting. The initial *learning rate* was set to $5 * 10^{-3}$ but divided by two if the loss value did not decrease for ten epochs.

B.2 Augmentations

All of the augmentations on Table B5 were applied with a probability of 0.5 without restricting the total number of augmentations per image.

Augmentation	**Probability**	**Parameters**
Random Flip	0.5	**Axis**
		0,1
Random Contrast Adjustment	0.5	**Gamma Range**
		[0.5,1.5]
Random Gaussian Smoothing	0.5	**Sigma Range**
		[0.25,1.5]

Table B5: Data augmentations and corresponding parameters used during training.

B.3 Loss Function

The Cox Partial Log-Likelihood [24] is computed with the following formula:

$$\mathcal{L}_{Cox}(\beta, X) = -\sum_{i \in U} (\beta^T X_i - log \sum_{j \in \Omega_i} e^{\beta^T X_j}) \tag{B1}$$

Here $r_i = \beta^T X_i$ is the predicted hazard ratio, U is the set of patients with LF, and Ω_i is the "at-risk" patients concerning patient i, $\Omega_i = \{j | t_j > t_i\}$.

In the ideal case, the formula for the loss function would consider all predictions simultaneously, allowing for a global optimization of the model across the entire dataset. However, because our predictions were generated by propagating whole images through large and complex networks, it became computationally infeasible to process every patient in the dataset simultaneously due to the high demand for memory and processing power. To address this limitation, we computed the loss function locally on smaller batches of data, updating the network incrementally with each batch. The loss function learns to rank the predicted hazard ratios anti-concordant to the follow-up times while considering the censoring information.

B.4 Explainability

The *beyond attention* method [32] generates the network's explanations as:

$$\bar{A}^{(l)} = I + \mathbb{E}_h[\nabla A^{(l)} \odot R^{(n_l)}]^+ \; ; \; l = 1, ..., L \tag{B2}$$

$$C = \bar{A}^{(1)} \cdot \bar{A}^{(2)} \cdot ... \cdot \bar{A}^{(L)} \tag{B3}$$

where $R^{(n_l)}$ is the propagated relevance of attention layer $A^{(l)}$, $\odot$ is the Hadamard product and $\mathbb{E}_h$ is the mean operation across the *heads* dimension. Only positive values of the gradients-relevance multiplication are considered to resemble positive relevance.

Appendix C Supplementary Experiments

C.1 Ablation Study: Different Sizes of ViT

The number of model parameters highly impacts neural network training, especially regarding the model's generalizability, a key factor in preventing overfitting the training set. The network size is also important for optimizing the utilization of hardware resources. In Table C6, we present the results of our selection of transformer size, comparing it to commonly used ones in the literature [21]. ViT-Small and ViT-Basic depicted examples of underfitting and overfitting to the dataset, respectively, while our *medium-sized* ViT model demonstrated superior performance. We also explain the performance of TransRP [31] as overfitting due to it being the largest network overall. Moreover, training ViT-*Medium* was also 15% faster compared to ViT-Basic .

Transfomer Encoder	**#Parameters**	**Train C-Index**	**Val C-Index**	**Test C-Index**
ViT-Small	26.1M	0.8824 ± 0.042	0.7010 ± 0.044	0.6911 ± 0.026
ViT-Basic	94.7M	0.9145 ± 0.039	0.6773 ± 0.099	0.6976 ± 0.055
TransRP	118M	0.9084 ± 0.057	0.6815 ± 0.062	0.7138 ± 0.049
ViT-*Medium* (ours)	57.3M	0.8945 ± 0.031	0.7042 ± 0.055	$\mathbf{0.7311} \pm 0.048$

Table C6: Comparison of results for different sizes of ViT and TransRP. All the networks were trained with the final configurations for input-cropping, batching, etc. No register tokens were used.

C.2 Effect of Cropping

We investigated the effect of cropping the MRI volumes to the VOI guided by the pre-trained segmentation network [19]. For a pure analysis, we trained end-to-end models with a single type of image processing (i.e., fully convolutional ResNet or transformer, and no hybrid). Row three of Table C7 shows that even with expanded bounding boxes, cropping had a strong impact on model performance, as it supported the focus on the target area with an 85% size reduction. It was also greatly beneficial for the

Model	Added Feature	Test C-Index
Vanilla ResNet34	-	0.5672 ± 0.035
Vanilla ViT		0.5912 ± 0.041
ResNet34	+ *Clinical Feautes*	0.6493 ± 0.039
ViT		0.6691 ± 0.031
ResNet34	+ *Segmentation-guided Image Cropping*	0.6705 ± 0.039
ViT		0.7139 ± 0.044
ResNet34	+ *Segmentation as Extra Input*	0.6796 ± 0.041
ViT		0.7267 ± 0.042
ResNet34	+ *Stratified Batching*	0.6879 ± 0.032
ViT		0.7311 ± 0.048
ResNet34	+ *Register Tokens*	n.a.
ViT		0.7336 ± 0.029

Table C7: Extension of Table 2 with ResNet34 results. Each row represents the addition of the given feature to the configuration from the previous row. It showcases the incremental improvement through our methodological enhancements, independent of the used model.

efficient use of computation resources, as the training duration for ViTs dropped from 60 minutes to 20 per 100 epochs.

C.3 Segmentation Maps as Input

For further assistance from the segmentation results, we used the predicted edema and tumor contours as an additional input channel. Row four of Table C7 demonstrates that models benefited from this additional guidance.

C.4 Improvement by Stratified Batching

When trained with randomly sampled batches, the prediction performance drops for all models. This proves that our stratified batching method yields more stable and powerful training.