Radiomics-based prediction of local control in patients with brain metastases following postoperative stereotactic radiotherapy

Josef A. Buchner[1], Florian Kofler[2,3,4,5], Michael Mayinger[6], Sebastian M. Christ[6], Thomas B. Brunner[7], Andrea Wittig[8], Bjoern Menze[5,9], Claus Zimmer[3], Bernhard Meyer[10], Matthias Guckenberger[6], Nicolaus Andratschke[6], Rami A. El Shafie[11,12,13], Jürgen Debus[11,12], Susanne Rogers[14], Oliver Riesterer[14], Katrin Schulze[15], Horst J. Feldmann[15], Oliver Blanck[16,20], Constantinos Zamboglou[17,18,19], Konstantinos Ferentinos[19], Angelika Bilger-Zähringer[17,18], Anca L. Grosu[17,18], Robert Wolff[20,21], Marie Piraud[2], Kerstin A. Eitz[1,22,23], Stephanie E. Combs[1,22,23], Denise Bernhardt[1,22], Daniel Rueckert[24], Benedikt Wiestler[3,4], Jan C. Peeken[1,22,23]

[1]Department of Radiation Oncology, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany
[2]Helmholtz AI, Helmholtz Zentrum Munich, Neuherberg, Germany
[3]Department of Diagnostic and Interventional Neuroradiology, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany
[4]TranslaTUM - Central Institute for Translational Cancer Research, Technical University of Munich, Munich, Germany
[5]Department of Informatics, Technical University of Munich, Munich, Germany
[6]Department of Radiation Oncology, University of Zurich, Zurich, Switzerland
[7]Department of Radiation Oncology, University Hospital Magdeburg, Magdeburg, Germany
[8]Department of Radiotherapy and Radiation Oncology, University Hospital Jena, Friedrich-Schiller University, Jena, Germany
[9]Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland
[10]Department of Neurosurgery, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany
[11]Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Germany
[12]Heidelberg Institute for Radiation Oncology (HIRO), National Center for Radiation Oncology (NCRO), Heidelberg, Germany
[13]Department of Radiation Oncology, University Medical Center Göttingen, Göttingen, Germany
[14]Radiation Oncology Center KSA-KSB, Kantonsspital Aarau, Aarau, Switzerland
[15]Department of Radiation Oncology, General Hospital Fulda, Fulda, Germany
[16]Department of Radiation Oncology, University Medical Center Schleswig Holstein, Kiel, Germany
[17]Department of Radiation Oncology, University of Freiburg - Medical Center, Freiburg, Germany
[18]German Cancer Consortium (DKTK), Partner Site Freiburg, Freiburg, Germany
[19]Department of Radiation Oncology, German Oncology Center, European University of Cyprus, Limassol, Cyprus
[20]Saphir Radiosurgery Center Frankfurt and Northern Germany, Kiel, Germany
[21]Department of Neurosurgery, University Hospital Frankfurt, Frankfurt, Germany
[22]Deutsches Konsortium für Translationale Krebsforschung (DKTK), Partner Site Munich, Munich, Germany
[23]Institute of Radiation Medicine (IRM), Department of Radiation Sciences (DRS), Helmholtz Center Munich, Munich, Germany
[24]Institute for Artificial Intelligence and Informatics in Medicine, Technical University of Munich, Munich, Germany

Abstract

Background

Surgical resection is the standard of care for patients with large or symptomatic brain metastases (BMs). Despite improved local control after adjuvant stereotactic radiotherapy, the local failure (LF) risk persists. Therefore, we aimed to develop and externally validate a pre-therapeutic radiomics-based prediction tool to identify patients at high LF risk.

Methods

Data were collected from *A Multicenter Analysis of Stereotactic Radiotherapy to the Resection Cavity of Brain Metastases* (AURORA) retrospective study (training cohort: 253 patients (two centers); external test cohort: 99 patients (five centers)). Radiomic features were extracted from the contrast-enhancing BM (T1-CE MRI sequence) and the surrounding edema (FLAIR sequence). Different combinations of radiomic and clinical features were compared. The final models were trained on the entire training cohort with the best parameters previously determined by internal 5-fold cross-validation and tested on the external test set.

Results

The best performance in the external test was achieved by an elastic net regression model trained with a combination of radiomic and clinical features with a concordance index (CI) of 0.77, outperforming any clinical model (best CI: 0.70). The model effectively stratified patients by LF risk in a Kaplan-Meier analysis ($p < 0.001$) and demonstrated an incremental net clinical benefit. At 24 months, we found LF in 9% and 74% of the low and high-risk groups, respectively.

Conclusions

A combination of clinical and radiomic features predicted freedom from LF better than any clinical feature set alone. Patients at high risk for LF may benefit from stricter follow-up routines or intensified therapy.

Keywords

- Machine learning
- Local failure prediction
- Radiomics
- Brain metastases
- Artificial intelligence

Key points

- Radiomics can predict the freedom from local failure in brain metastasis patients
- Clinical and MRI-based radiomic features combined performed better than either alone
- The proposed model significantly stratifies patients according to their risk

Importance of the Study

Local failure after treatment of brain metastases has a severe impact on patients, often resulting in additional therapy and loss of quality of life. This multicenter study investigated the possibility of predicting local failure of brain metastases after surgical resection and stereotactic radiotherapy using radiomic features extracted from the contrast-enhancing metastases and the surrounding FLAIR-hyperintense edema.

By interpreting this as a survival task rather than a classification task, we were able to predict the freedom from failure probability at different time points and appropriately account for the censoring present in clinical time-to-event data.

We found that synergistically combining clinical and imaging data performed better than either alone in the multicenter external test cohort, highlighting the potential of multimodal data analysis in this challenging task. Our results could improve the management of patients with brain metastases by tailoring follow-up and therapy to their individual risk of local failure.

Introduction

Brain metastases (BMs) are the most common malignant brain tumors, far outnumbering primary brain tumors such as gliomas[1]. Recent guidelines recommend surgery as a treatment for patients with symptomatic or large BMs[2]. To improve local control, stereotactic radiotherapy (SRT) should be applied to the resection cavity in patients with one to two resected BMs[2]. This way, local control rates of 70% to 90% at twelve months can be achieved[3].

Determining an individual patient's risk of local recurrence can benefit patients by tailoring follow-up treatment and care. For example, patients at high risk of local failure may benefit from SRT dose escalation, systemic therapy agents with penetration of the blood-brain barrier, and more frequent follow-up imaging after SRT to detect a potential failure early.

Recent publications have demonstrated the power of automated segmentation of BMs and their surrounding edema[4–6]. This cannot only help radiation oncologists by eliminating the time-consuming task of manual BM delineation but can also simplify other additional evaluations: Radiomics allows the extraction of large amounts of quantitative imaging features from a previously delineated image[7]. This enables professionals to analyze additional information that is not visible to the human eye and allows the creation of predictive mathematical models[8].

Such radiomics models can be used for multiple tasks such as tumor characterization, prediction of treatment response, and prognostic risk assessment[9–13].

Some radiomic features are sensitive to acquisition modes and reconstruction parameters[14]. Furthermore, MRI intensities are not standardized and depend on the manufacturer and model of the devices[15]. Moreover, patients and treatment characteristics can differ between medical institutions. Therefore, multicentric training and testing are needed to develop and validate generalizable models.

Several previous studies could demonstrate the general propensity of radiomics to predict local failure (LF) as a binary variable in patients receiving stereotactic radiotherapy without surgery in monocentric studies without external validation[16–18].

The aim of this project was to develop a pre-therapeutic radiomics-based machine learning model to predict freedom from local failure (FFLF) after surgical resection and SRT of BMs. All models were validated in an external multicenter international test cohort. The ability to stratify patients into specific risk groups and their net clinical benefit were assessed.

Methods

AURORA study

MR imaging and clinical data was collected as part of the "*A Multicenter Analysis of Stereotactic Radiotherapy to the Resection Cavity of Brain Metastases"* (AURORA) retrospective trial. The trial was supported by the *Radiosurgery and Stereotactic Radiotherapy Working Group* of the *German Society for Radiation Oncology* (DEGRO). The inclusion criteria were: known primary tumor with resected BM and SRT with a radiation dose of > 5 Gy per fraction. Exclusion criteria were: interval between surgery and RT > 100 days, premature discontinuation of RT, and any previous cranial radiation therapy (RT). Synchronous non-resected BMs had to be treated simultaneously with SRT. Ethical approval was obtained at each institution (main approval at the Technical University of Munich: 119/19 S-SR).
LF was determined by individual radiologic review or by histologic results after recurrence surgery. FFLF was calculated as the time difference between the end of SRT and LF. If no LF occurred, patients were right-censored after the last available imaging follow-up.

Dataset

In total, we collected data from 481 patients from seven centers. We analyzed four preoperative imaging sequences of each patient: a T1-weighted sequence with and without contrast enhancement (T1-CE and T1), a T2-weighted sequence (T2) as well as a T2 fluid-attenuated inversion recovery sequence (T2-FLAIR). Except for T1-CE, a missing sequence was allowed.
The required data were available for 352 patients. We split the patients into a training cohort with 253 patients from two centers and an external, multicenter, international test cohort with 99 patients from five centers.

Preprocessing

The DICOM (Digital Imaging and Communications in Medicine file format) images were converted to NIfTI (Neuroimaging Informatics Technology Initiative file format) using dcm2niix[19]. The MRI sequences were then further preprocessed using BraTS-Toolkit[20]. First, the sequences were co-registered using niftyreg[21] and these were then transformed into the T1-CE space. A brain mask was created using HD-BET[22] and applied to all sequences to extract only the brain without the surrounding skull. The skull-stripped sequences were transformed into the BraTS space using the SRI-24 atlas[23]. Overall, the preprocessing provided co-registered, skull-stripped sequences in a 1 millimeter isotropic resolution in BraTS space.
The missing sequences were then synthesized using a generative adversarial network (GAN). The GAN takes the three available sequences as input and generates the matching missing fourth sequence. We used a GAN which was originally developed for missing sequences in glioma imaging[24], but has been proven to work for metastasis imaging[4,5].

Segmentation

All contrast-enhancing metastases and their surrounding edema were individually segmented using the open-source software 3D-Slicer (version 4.13.0, stable release, https://www.slicer.org/)[25] by a medical doctoral student (JAB) after undergoing extensive training by a board-certified radiation oncologist (JCP) (7 years of experience). To ensure accuracy, all segmentations for the test cohort were reviewed and manually adjusted by JCP.
To test the feasibility of a fully automated workflow, segmentations generated by a previously trained neural network[4,5] were used as alternative segmentations and compared to the manual segmentations.

As around 25% of patients had multiple BMs, but usually only the largest is resected[26], we also determined the largest metastasis with a connected component analysis[27] in all patients with multiple BMs and used only that metastasis and its surrounding edema as segmentations for an additional analysis.

Radiomic feature extraction

Radiomic features were extracted with pyradiomics (version 3.0.1, https://github.com/AIM-Harvard/pyradiomics)[28] using the Python implementation. The metastasis segmentation was used to extract the T1-CE features, while the edema segmentation was used for the T2-FLAIR features. In total, we extracted 104 original features per segmentation (see Supplemental Table 1 for a list of features and extraction parameters).
Further analysis and modeling were performed in the programming language R 4.2.3[29]. To adhere to the *Image Biomarker Standardisation Initiative* (IBSI) standard[30], the kurtosis was adjusted by -3. We created nine feature sets in total. Three of these included only radiomic features. The *metastasis* and *edema* feature sets were created by extracting the features from the T1-CE sequence and T2-FLAIR sequence, respectively. Both feature sets were merged into a *combined* feature set. We also created three clinical feature sets with the following clinical features:

- *pre-OP* feature set: patient age at RT start, Karnofsky performance status (KPS), histology of the primary tumor, location of BM
- *post-OP* feature set: *pre-OP* + resection status
- *RT* feature set: *post-OP* + concurrent chemotherapy, concurrent immunotherapy and equivalent dose in two Gray fractions (EQD2)

As a seventh feature set, we combined all radiomic features (*combined*) with the *pre-OP* feature set to *comb+pre-OP*.
Multiple publications suggest the predictive value of the brain metastasis volume (BMV) for predicting LF[31–33]. Therefore, we created two additional feature sets by adding the cumulative BMV of each patient as an additional feature to the *pre-OP* set (*pre-OP+BMV*) and the *comb+pre-OP* set (*comb+pre-OP+BMV*).

Intraclass correlation

To identify radiomic features that were susceptible to small changes in segmentation, we generated additional segmentations of all patients in the training cohort using the previously mentioned neural network[4]. Intraclass correlation (ICC (3,1)) was calculated using the R package "irr"[34]. According to Koo *et al*., an ICC above 0.75 is considered "good"[35]. Consequently, this value was employed as a cut-off threshold. Of the 208 features, 173 (83%) had an ICC of > 0.75 and were selected for all further steps. Of the 35 excluded features, the majority (27) were extracted from the edema mask, while only eight excluded features were extracted from the metastasis mask.
All selected radiomic features were normalized by z-score standardization and by applying the Yeo-Johnson transformation[36] to transform the distribution of a variable into a Gaussian distribution.

Feature reduction

We applied a minimum redundancy - maximum relevance (MRMR) ensemble feature selection framework implemented in R[37] initially proposed by Ding *et al*.[38] as an efficient method for the selection of relevant and non-redundant features.
We created multiple smaller feature sets of the *metastasis*, *edema*, and *combined* feature sets with three, five, seven, nine, eleven, thirteen, and fifteen features each.
We used bootstrapping[39] to obtain more reliable results: Feature reduction was repeatedly applied to 1000 bootstrap samples for each set and each number of features. For our final set of features, we ranked the features based on the number of times they were selected.

The best number of features was later determined by nested cross-validation in the training set.

Batch harmonization

To account for differences created by 29 different MRI scanners in our multicenter dataset, we used batch harmonization implemented by neuroCombat[40]. In total, 10 batches were created according to the MRI model names by combining related models. According to Leithner *et al*.[41], ComBat harmonization without Empirical Bayes estimation provided slightly higher performance in similar machine learning tasks. Therefore, Empirical Bayes was deactivated. Besides the non-harmonized dataset, we created two harmonized datasets: one by only adjusting the means and the other by adjusting means and variances.

Model creation, testing, and patient stratification

For model creation and evaluation, the R package MLR3[42] was used as a basis. Our prediction target was right-censored time-to-event data, where we used LF as the event and the FFLF or time-to-last imaging follow-up as the time variable for patients with and without event, respectively. We compared three different learners: random forest (RF), extreme gradient boosting (xgboost), and generalized linear models with elastic net regularization learner (ENR).

We implemented nested cross-validation to select the best mode of batch harmonization and the best number of features: For batch harmonization selection, all three datasets were compared while always using the *combined* feature set with nine features. Five iterations of five-fold nested cross-validation for dataset selection showed no significant difference between the sets with and without batch harmonization ($p = 0.3$, Kruskal-Wallis rank sum test). Therefore, all further analyses were performed on the base dataset without batch harmonization to avoid unnecessary and potentially distorting preprocessing steps. To select the ideal number of features in each feature set, the nested cross-validation was conducted without batch harmonization. The best average performance was achieved with seven, three, and seven features in the *metastasis*, *edema*, and *combined* sets, respectively. The *comb+pre-OP* set, which included the seven *combined* and four *clinical* features, therefore, had 11 features. The features are listed in Supplementary Table 2.

The parameter tuning was performed using repeated cross-validation. All tuning and selection steps were performed on the training set. To account for the class imbalance (around 1:5 event:no-event), synthetic minority over-sampling was implemented using SMOTE[43]. We used an implementation in R which is capable of handling numeric and categorical data. The number of samples in the minority class was increased by creating synthetic samples to reach a ratio of 1:2. We only used SMOTE on the training folds in each step of our (nested) cross-validation. This way we ensured that our models were only validated on real patients.

The final models were trained with the best parameters determined by the cross-validation on the whole training set while also using SMOTE to balance the classes. The models were then tested on our multicentric external test cohort.

The 33rd and 66th percentiles of the continuous risk ranks in the training cohort were used as cutoffs for patient stratification. These cutoffs were used to divide the test cohort into three groups according to their predicted continuous risk rank and compare their survival with Kaplan Meier analysis.

Metrics

To account for both timing and outcome, the learners' performance was quantified using the concordance index (CI)[44]. The 95% confidence intervals are based on 10,000 bootstrap samples. A decision curve analysis was performed to consider clinical consequences with a time endpoint of 24 months[45]. The threshold range was chosen as suggested by Vickers *et al.*[46] based on these considerations: Since LF is a severe event and its detection is critical, a

lower threshold of 5% seems appropriate. Especially in elderly and multimorbid patients, where additional imaging may be burdensome, an upper threshold of 30% is reasonable.

Results

An overview of patient characteristics of both patient cohorts is shown in Table 1. A total of 147 patients had missing sequences, the majority of which were missing T2 and T1 sequences (82% and 10%, respectively), which were not relevant for our further analyses. The general workflow, with example images of a test cohort patient, is shown in Figure 1.

Baseline clinical models

To create a baseline for comparison with our radiomic models, we first tested the predictive value of two established clinical indices with univariate Cox analysis: the Recursive Partitioning Analysis (RPA)[47] and the Graded Prognostic Assessment (GPA)[48] index. They reached a CI of 0.47 and 0.52 in the internal validation, respectively. In external testing, RPA again performed worse with a CI of 0.39 compared to GPA with a CI of 0.44.

Model performance

The performances in the internal validation, as well as in the multicentric external test cohort, are shown in Table 2. To determine the best overall learner, we ranked the performance across all feature sets and found that ENR ranked best, followed by RF and xgboost with mean ranks of 1.4, 1.6, and 2.9, respectively. Therefore, all further experiments were conducted with ENR. For completeness, the results obtained by RF and xgboost are shown in Supplementary Tables 3 and 4. The highest mean CI across all five folds and ten iterations of the cross-validation was achieved with the *comb+pre-OP* feature set (CI = 0.67).
The *comb+pre-OP* feature set also led to the highest performance in the external test cohort and achieved a CI of 0.77. While the *T1-CE* feature set achieved a CI of 0.76, *FLAIR* was only able to reach 0.50. The three clinical feature sets performed slightly worse than our radiomic feature sets or the combined feature sets: the *pre-OP*, *post-OP*, and *RT* feature sets reached a CI of 0.64, 0.63, and 0.63 in the internal validation, respectively. In external testing, they achieved a CI of 0.70, 0.65, and 0.70, respectively. While adding the BMV to the *pre-OP* feature set did not change the predictive performance, adding it to *comb+pre-OP* led to worse results with a CI of 0.72.
For reproducibility, we listed the beta values used by our best model (*comb+pre-OP* ENR) in Supplementary Table 5. The corresponding calibration curve to this model is shown in Figure 2 (right panel). Furthermore, we calculated the time-dependent area under the receiver operating characteristic curve (AUC) by transforming the crank to an event probability distribution. The proposed model reached a mean of 0.80. Supplementary Figure 1 shows the plotted time-dependent AUC.

Patient stratification

Using the cutoffs determined by the training cohort as described above, our *comb+pre-OP* ENR model was able to significantly stratify the patients into three risk groups with a low, medium, and high risk of local failure ($p = 0.0001$, Chi-squared Test). A Kaplan-Meier analysis with all three groups is shown in Supplementary Figure 2.
By combining the low- and medium-risk groups into one, we created dichotomous predictions. Kaplan-Meier analysis (Figure 2) illustrates the survival in each risk group. Decision curve analysis using these predictions showed a net benefit of our predictive model compared to treating all patients in the relevant threshold range (Figure 3).

The relevance of brain metastasis volume

The predictions of our *comb+pre-OP* ENR model did weakly correlate with the cumulative BMV or BMV of the largest BM (Spearman's rank correlation: $r = 0.246$ ($p = 0.014$) and 0.254 ($p = 0.011$), respectively).
While cumulative BMV alone was highly predictive in the test cohort, with a CI of 0.76 in a univariate Cox analysis, it only achieved a CI of 0.53 in internal validation. Using the BMV of

only the largest BM increased the internal validation and external testing performance to 0.55 and 0.77, respectively. There was no significant difference in the BMV between the training and test cohort (p = 0.64, Wilcoxon rank sum test).
Stratifying our test set into small and large BMs by dividing the set at the median volume resulted in groups with three and 13 events, respectively. Our best model scored a CI of 0.58 and 0.78 in the respective groups. Interestingly, the model significantly risk-stratified the patients in the small BMV group, but not in the high BMV group (corresponding Kaplan-Meier analysis are depicted in Supplementary Figures 3 and 4).
When repeating the training and testing with the radiomic features extracted only from the biggest BM, the ENR learner was able to reach a CI of 0.75 (*comb+clinical*, Table 3). The results obtained by the RF model even surpassed our previously best model by 0.01 (CI = 0.78, Supplementary Table 4).

End-to-end model using neural network-based automatic segmentations

To test the predictive value of neuronal network-based segmentations and therefore test the feasibility of a fully automated workflow, we conducted an additional parameter tuning and training run with radiomic features extracted from the automatically created segmentations. The results for our ENR learner are shown in Table 3. The best test results with this data were again obtained with the *comb+pre-OP* feature set (CI = 0.72). Overall, we observed an average decrease in performance by 0.06.

Discussion

In this work, we were able to develop radiomics-based machine learning models that were able to predict FFLF better than clinical features alone. Our best model was trained with a combination of radiomic and clinical features and achieved a CI of 0.77 in a multicenter external test cohort outperforming any clinical predictive model. Our final model's predictions significantly stratified the test patients into two risk groups and achieved an incremental net clinical benefit.

When using automatically generated segmentations from a previously trained neural network, the models performed slightly worse, with an average performance loss of 0.06. Still, the *comb+pre-OP* ENR model was able to reach a respectable CI of 0.72 in external testing. This demonstrates that an end-to-end solution is possible without clinician intervention.

The results in the external test cohort were, on average, better by a CI of 0.04. This may be explained by the larger amount of data available for training: The models tested on the external cohort were trained on all training data, while for internal validation, only 80% of the data was used for training, while testing was performed on the remaining 20%.

Several studies have approached predicting the LF of BMs. Most of them interpreted the prediction as a classification task and therefore only predicted whether an event occurred at a predetermined time[16–18,49–58]. In contrast, we approached the task as a survival task and therefore predicted a combination of event and time in terms of FFLF.
Another study predicting event and time of local failure by Huang *et al.*[59] used Cox proportional hazards models and found that non-small cell lung cancer BMs with a higher zone percentage were more likely to respond favorably to Gamma Knife radiosurgery. In contrast to the treatment with surgery and adjuvant SRT in our study, the aforementioned studies focused on BMs treated with SRT, WBRT, and immune checkpoint inhibitors. Only one monocentric study with 67 patients by Mulford *et al.*[52] investigated the prediction of local recurrence after surgical resection and adjuvant stereotactic radiosurgery, and found that radiomic features provided more robust predictive models of local control rates than clinical features (AUC = 0.73 vs. 0.40). Unlike our study, they predicted local failure as a binary classification task.
Another unique feature of our study is the multicenter external test cohort with patients treated at five different centers in multiple countries. In contrast to our study, the aforementioned studies all tested their models on an internal validation set and were therefore not tested on such a wide variety of scanners and imaging protocols as our models were.

Contrary to findings in previous studies[60], the cumulative BMV and the BMV of the largest BM were not predictive in the internal validation, where they only reached a CI of 0.53 and 0.55, respectively. Since outcome and BMV appear to be independent in the training cohort, radiomic features representing BM size were not selected by our feature reduction algorithm. The only selected shape class feature in the best-performing feature set was metastasis flatness. Moreover, there was only a minor correlation (r = 0.25) between the predictions of the radiomic model and BMV. This shows that Radiomics can predict local failure based on features that do not directly represent BM size or volume.

Compared to approaches focusing on the use of neural networks, the use of classical machine learning has some advantages: Because only a small number of features are fed into the model, it becomes more comprehensible. Since it is known how the radiomics features are computed, it is possible to infer the clinical correlates. Neural networks, on the other hand, are more of intransparent black boxes, and it is difficult to understand exactly which characteristics of the tumor are predictive. In addition, neural networks often require the use of a graphics processing unit (GPU) to complete predictions in a reasonable amount

of time, while our models run on the central processing unit (CPU) and can, therefore, run on low-end hardware.

Nevertheless, this work has several limitations: Training the models with only a limited number of features extracted from the segmentations prevents them from taking other factors into account, such as the surrounding tissue. Furthermore, segmentations of consistent quality are necessary for reliable results. In this study, all segmentations were created by the same person. To reduce the influence of the personal segmentation style, only features with a high correlation between manual and automatic segmentations were used for further modeling. The sole use of automatically generated segmentations may help with this limitation.

Around one-quarter of our patients had multiple BMs. By using the cumulative BMV as a feature, we not only took the volume of the resected BM into account but also the volume of all additional BMs. In our additional analysis, where we only considered radiomic features extracted from the largest metastasis, the best result improved slightly, while the mean across all models decreased by 0.01 compared to using the combined segmentation of all BMs. From this, we can conclude that segmenting all BMs did not harm the prediction of local failure of the resected BM.

In addition, radiomic features were extracted from a total of twelve synthesized T2-FLAIR sequences (six in the training cohort and six in the test cohort). Excluding these patients from the training and test sets resulted in a slight increase in performance. The largest increase in performance was found in the combined feature set (CI = 0.72 from 0.69). Furthermore, the *T1-CE* model showed the second largest increase in performance, surpassing our previous best feature set (*comb+pre-OP*), which showed no change in performance. Since the new best model did not even include features extracted from the T2-FLAIR sequence, we can conclude that radiomic features extracted from the synthesized T2-FLAIR sequences did not noticeably affect the performance of our model and the increase in performance may be attributed to the exclusion of difficult cases.

Despite these limitations, we were able to develop a model to predict freedom from local progression of BMs after resection and adjuvant SRT. The model performed well in a multicenter external test cohort with a variety of MRI scanners and imaging and therapy protocols. This model may help to tailor treatment to a patient's individual risk of metastasis recurrence, thereby improving the overall management of BMs. We have published the model as an easy-to-use web app (https://jbuchner.shinyapps.io/shiny/), where the user can either upload the required MRI sequences and segmentations or input previously extracted radiomic features.

Funding

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, Project number 504320104 - PE 3303/1-1 (JCP), WI 4936/4-1 (BW), RU 1738/5-1 (DR)).

Conflict of Interest

AW: Consultant: Gilead and Hologic Medicor GmbH; Honoraria: ACCURAY International, Universitätsklinikum Leipzig AöR, and Sanofi-Aventis GmbH; Board: IKF GmbH am Krankenhaus Nordwest
BeM: Grants: BrainLab, Zeiss, Ulrich, and Spineart; Royalties: Medacta and Spineart; Consultant and Honoraria: Medacta, Brainlab, and Zeiss; Stock: Sonovum
MG: President-Elect of ESTRO
NA: Independent Contractor: SAKK - Swiss Association for Clinical Cancer Research; Board: AstraZeneca; Research funding: ViewRay Inc.; Stock: Moderna Inc. and Idorsia AG; Chair: EORTC and Global Harmonization Group
SR: Honoraria: Brainlab
OB: Grants: European Union's Horizon 2020 research and innovation programme; Board: working groups for Stereotactic Radiotherapy of the German Radiation Oncology and Medical Physics Societies, Section Editor of Strahlentherapie und Onkologie Journal
ALG: Research funding: Novocure
SEC: Honoraria and travel expenses: Roche, Bristol-Myers Squibb, Brainlab, AstraZeneca, Accuray, Dr. Sennewald, Daiichi Sankyo, Elekta, Medac, Icotec AG, HMG Systems Engineering, and Carl Zeiss Meditec AG
DB: Honoraria and travel expenses: Novocure
The remaining authors have no potential conflicts of interest to disclose.

Authorship

All authors were involved in the data curation and acquisition of resources.
Formal analysis, methodology, and software: JAB, JCP
Visualization: JAB
Writing - Original Draft: JAB, FK, BW, JCP
Writing - Review & Editing: JAB, FK, SMC, TBB, AW, BeM, SR, OR, OB, CoZ, ABZ, ALG, BW, JCP
Supervision: MP, SEC, BW, JCP
Project administration: KAE, SEC, DB, JCP
Funding acquisition: SEC, RU, BW, JCP
All authors approved the manuscript.

Data Availability

The trained model is available as a shiny web app. Training and test data are not publicly available.

1. Johnson JD, Young B. Demographics of Brain Metastasis. *Neurosurg Clin N Am*. 1996;7(3):337-344. doi:10.1016/S1042-3680(18)30365-6

2. Vogelbaum MA, Brown PD, Messersmith H, et al. Treatment for Brain Metastases: ASCO-SNO-ASTRO Guideline. *Journal of Clinical Oncology*. 2022;40(5):492-516. doi:10.1200/JCO.21.02314

3. Minniti G, Niyazi M, Andratschke N, et al. Current status and recent advances in resection cavity irradiation of brain metastases. *Radiation Oncology*. 2021;16(1):1-14. doi:10.1186/S13014-021-01802-9

4. Buchner JA, Kofler F, Etzel L, et al. Development and external validation of an MRI-based neural network for brain metastasis segmentation in the AURORA multicenter study. *Radiotherapy and Oncology*. 2023;178:109425. doi:10.1016/J.RADONC.2022.11.014

5. Buchner JA, Peeken JC, Etzel L, et al. Identifying core MRI sequences for reliable automatic brain metastasis segmentation. *Radiotherapy and Oncology*. 2023;188:109901. doi:10.1016/J.RADONC.2023.109901

6. Pflüger I, Wald T, Isensee F, et al. Automated detection and quantification of brain metastases on clinical MRI data using artificial neural networks. *Neurooncol Adv*. 2022;4(1). doi:10.1093/noajnl/vdac138

7. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: Extracting more information from medical images using advanced feature analysis. *Eur J Cancer*. 2012;48(4):441-446. doi:10.1016/j.ejca.2011.11.036

8. Peeken JC, Wiestler B, Combs SE. Image-Guided Radiooncology: The Potential of Radiomics in Clinical Application. *Recent Results in Cancer Research*. 2020;216:773-794. doi:10.1007/978-3-030-42618-7_24

9. Lang DM, Peeken JC, Combs SE, Wilkens JJ, Bartzsch S. Deep Learning Based HPV Status Prediction for Oropharyngeal Cancer Patients. *Cancers (Basel)*. 2021;13(4):786. doi:10.3390/cancers13040786

10. Peeken JC, Neumann J, Asadpour R, et al. Prognostic assessment in high-grade soft-tissue sarcoma patients: A comparison of semantic image analysis and radiomics. *Cancers (Basel)*. 2021;13(8):1929. doi:10.3390/CANCERS13081929

11. Shahzadi I, Lattermann A, Linge A, et al. Do We Need Complex Image Features to Personalize Treatment of Patients with Locally Advanced Rectal Cancer? In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol 12907 LNCS. Springer Science and Business Media Deutschland GmbH; 2021:775-785. doi:10.1007/978-3-030-87234-2_73

12. Spohn SKB, Schmidt-Hegemann NS, Ruf J, et al. Development of PSMA-PET-guided CT-based radiomic signature to predict biochemical recurrence after salvage radiotherapy. *Eur J Nucl Med Mol Imaging*. 2023;50(8):2537-2547. doi:10.1007/S00259-023-06195-3

13. Leger S, Zwanenburg A, Leger K, et al. Comprehensive Analysis of Tumour Sub-Volumes for Radiomic Risk Modelling in Locally Advanced HNSCC. *Cancers (Basel)*. 2020;12(10):3047. doi:10.3390/cancers12103047

14. Abdollahi H, Chin E, Clark H, et al. Applications and limitations of radiomics. *Phys Med Biol*. 2016;61(13):R150. doi:10.1088/0031-9155/61/13/R150

15. Simmons A, Tofts PS, Barker GJ, Arridge SR. *Sources of Intensity Nonuniformity in Spin Echo Images at 1.5 T.*; 1994.

16. Mouraviev A, Detsky J, Sahgal A, et al. Use of radiomics for the prediction of local control of brain metastases after stereotactic radiosurgery. *Neuro Oncol*. 2020;22(6):797-805. doi:10.1093/NEUONC/NOAA007

17. Kawahara D, Tang X, Lee CK, Nagata Y, Watanabe Y. Predicting the Local Response of Metastatic Brain Tumor to Gamma Knife Radiosurgery by Radiomics With a Machine Learning Method. *Front Oncol*. 2021;10:3003. doi:10.3389/fonc.2020.569461

18. Karami E, Soliman H, Ruschin M, et al. Quantitative MRI Biomarkers of Stereotactic Radiotherapy Outcome in Brain Metastasis. *Sci Rep*. 2019;9(1):1-11. doi:10.1038/s41598-019-56185-5

19. Li X, Morgan PS, Ashburner J, Smith J, Rorden C. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *J Neurosci Methods*. 2016;264:47-56. doi:10.1016/J.JNEUMETH.2016.03.001

20. Kofler F, Berger C, Waldmannstetter D, et al. BraTS Toolkit: Translating BraTS Brain Tumor Segmentation Algorithms Into Clinical and Scientific Practice. *Front Neurosci*. 2020;14. doi:10.3389/fnins.2020.00125

21. Modat M, Cash DM, Daga P, Winston GP, Duncan JS, Ourselin S. Global image registration using a symmetric block-matching approach. *J Med Imaging (Bellingham)*. 2014;1(2):024003. doi:10.1117/1.JMI.1.2.024003

22. Isensee F, Schell M, Pflueger I, et al. Automated brain extraction of multisequence MRI using artificial neural networks. *Hum Brain Mapp*. 2019;40(17):4952-4964. doi:10.1002/hbm.24750

23. Rohlfing T, Zahr NM, Sullivan E V., Pfefferbaum A. The SRI24 multichannel atlas of normal adult human brain structure. *Hum Brain Mapp*. 2010;31(5):798-819. doi:10.1002/HBM.20906

24. Thomas MF, Kofler F, Grundl L, et al. Improving Automated Glioma Segmentation in Routine Clinical Use Through Artificial Intelligence-Based Replacement of Missing Sequences With Synthetic Magnetic Resonance Imaging Scans. *Invest Radiol*. 2022;57(3):187-193. doi:10.1097/RLI.0000000000000828

25. Kikinis R, Pieper SD, Vosburgh KG. 3D Slicer: A Platform for Subject-Specific Image Analysis, Visualization, and Clinical Support. *Intraoperative Imaging and Image-Guided Therapy*. Published online 2014:277-289. doi:10.1007/978-1-4614-7657-3_19

26. Hatiboglu MA, Wildrick DM, Sawaya R. The role of surgical resection in patients with brain metastases. *Ecancermedicalscience*. 2013;7(1). doi:10.3332/ECANCER.2013.308

27. Silversmith W, Kemnitz N. 2020 seung-lab/connected-components-3d. seung-lab. See https://github.com/seung-lab/connected-components-3d.

28. Van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*. 2017;77(21):e104-e107. doi:10.1158/0008-5472.CAN-17-0339

29. R Core Team. R: A Language and Environment for Statistical Computing. Published online 2022. https://www.R-project.org/

30. Zwanenburg A, Vallières M, Abdalah MA, et al. The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*. 2020;295(2):328-338. doi:10.1148/RADIOL.2020191145

31. Mahajan A, Ahmed S, McAleer MF, et al. Post-operative stereotactic radiosurgery versus observation for completely resected brain metastases: a single-centre, randomised, controlled, phase 3 trial. *Lancet Oncol*. 2017;18(8):1040-1048. doi:10.1016/S1470-2045(17)30414-X

32. Hughes RT, Black PJ, Page BR, et al. Local control of brain metastases after stereotactic radiosurgery: the impact of whole brain radiotherapy and treatment paradigm. *J Radiosurg SBRT*. 2016;4(2):89. Accessed May 9, 2023. /pmc/articles/PMC5658880/

33. de Azevedo Santos TR, Tundisi CF, Ramos H, et al. Local control after radiosurgery for brain metastases: Predictive factors and implications for clinical decision. *Radiation Oncology*. 2015;10(1):1-9. doi:10.1186/S13014-015-0367-Y

34. Gamer M, Lemon J, <puspendra.pusp22@gmail.com> IFPS. irr: Various Coefficients of Interrater Reliability and Agreement. Published online 2019. https://www.r-project.org

35. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J Chiropr Med*. 2016;15(2):155. doi:10.1016/J.JCM.2016.02.012

36. Yeo INK, Johnson RA. A new family of power transformations to improve normality or symmetry. *Biometrika*. 2000;87(4):954-959. doi:10.1093/BIOMET/87.4.954

37. De Jay N, Papillon-Cavanagh S, Olsen C, El-Hachem N, Bontempi G, Haibe-Kains B. mRMRe: an R package for parallelized mRMR ensemble feature selection. *Bioinformatics*. 2013;29(18):2365-2368. doi:10.1093/BIOINFORMATICS/BTT383

38. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol*. 2005;3(2):185-205. doi:10.1142/S0219720005001004

39. Efron B. Bootstrap Methods: Another Look at the Jackknife. *https://doi.org/101214/aos/1176344552*. 1979;7(1):1-26. doi:10.1214/AOS/1176344552

40. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118-127. doi:10.1093/BIOSTATISTICS/KXJ037

41. Leithner D, Schöder H, Haug A, et al. Impact of ComBat Harmonization on PET Radiomics-Based Tissue Classification: A Dual-Center PET/MRI and PET/CT Study. *J Nucl Med*. 2022;63(10):1611-1616. doi:10.2967/jnumed.121.263102

42. Lang M, Binder M, Richter J, et al. mlr3: A modern object-oriented machine learning framework in R. *J Open Source Softw*. 2019;4(44):1903. doi:10.21105/JOSS.01903

43. Chawla N V., Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*. 2002;16:321-357. doi:10.1613/JAIR.953

44. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the Yield of Medical Tests. *JAMA*. 1982;247(18):2543-2546. doi:10.1001/JAMA.1982.03320430047030

45. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6):565-574. doi:10.1177/0272989X06295361

46. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*. 2016;352. doi:10.1136/BMJ.I6

47. Gaspar L, Scott C, Rotman M, et al. Recursive partitioning analysis (RPA) of prognostic factors in three Radiation Therapy Oncology Group (RTOG) brain metastases trials. *Int J Radiat Oncol Biol Phys*. 1997;37(4):745-751. doi:10.1016/S0360-3016(96)00619-0

48. Sperduto PW, Berkey B, Gaspar LE, Mehta M, Curran W. A new prognostic index and comparison to three other indices for patients with brain metastases: an analysis of 1,960 patients in the RTOG database. *Int J Radiat Oncol Biol Phys*. 2008;70(2):510-514. doi:10.1016/J.IJROBP.2007.06.074

49. Jalalifar SA, Soliman H, Sahgal A, Sadeghi-Naini A. Predicting the outcome of radiotherapy in brain metastasis by integrating the clinical and MRI-based deep learning features. *Med Phys*. 2022;49(11):7167-7178. doi:10.1002/MP.15814

50. Jaberipour M, Soliman H, Sahgal A, Sadeghi-Naini A. A priori prediction of local failure in brain metastasis after hypo-fractionated stereotactic radiotherapy using quantitative MRI and machine learning. *Scientific Reports 2021 11:1*. 2021;11(1):1-10. doi:10.1038/s41598-021-01024-9

51. Du P, Liu X, Shen L, et al. Prediction of treatment response in patients with brain metastasis receiving stereotactic radiosurgery based on pre-treatment multimodal MRI radiomics and clinical risk factors: A machine learning model. *Front Oncol*. 2023;13. doi:10.3389/FONC.2023.1114194

52. Mulford K, Chen C, Dusenbery K, et al. A radiomics-based model for predicting local control of resected brain metastases receiving adjuvant SRS. *Clin Transl Radiat Oncol*. 2021;29:27-32. doi:10.1016/J.CTRO.2021.05.001

53. Du P, Liu X, Xiang R, et al. Development and validation of a radiomics-based prediction pipeline for the response to stereotactic radiosurgery therapy in brain metastases. *Eur Radiol*. 2023;1:1-11. doi:10.1007/S00330-023-09930-4

54. Devries DA, Tang T, Alqaidy G, et al. Dual-center validation of using magnetic resonance imaging radiomics to predict stereotactic radiosurgery outcomes. *Neurooncol Adv*. 2023;5(1):1-14. doi:10.1093/NOAJNL/VDAD064

55. Zhao S, Hou D, Zheng X, et al. MRI radiomic signature predicts intracranial progression-free survival in patients with brain metastases of ALK-positive non-small cell lung cancer. *Transl Lung Cancer Res*. 2021;10(1):368-380. doi:10.21037/TLCR-20-361

56. Cha YJ, Jang W Il, Kim MS, et al. Prediction of Response to Stereotactic Radiosurgery for Brain Metastases Using Convolutional Neural Networks. *Anticancer Res*. 2018;38(9):5437-5445. doi:10.21873/ANTICANRES.12875

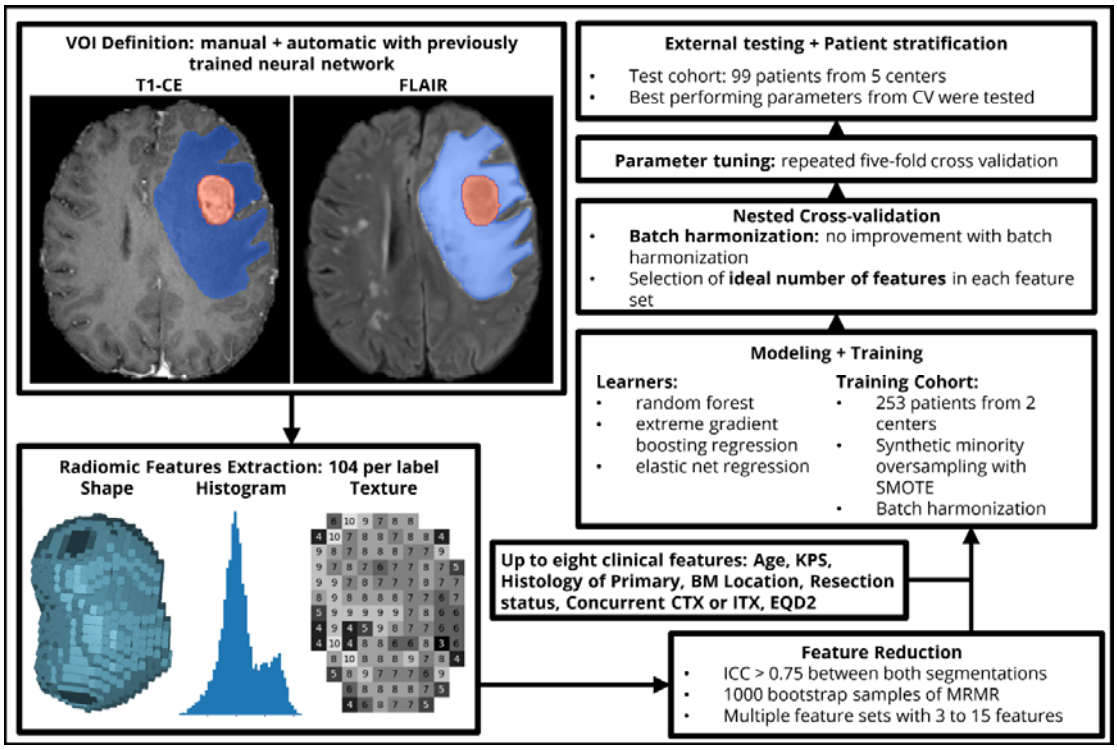
57. Wang H, Xue J, Qu T, et al. Predicting local failure of brain metastases after stereotactic radiosurgery with radiomics on planning MR images and dose maps. *Med Phys*. 2021;48(9):5522-5530. doi:10.1002/MP.15110

58. Jiang Z, Wang B, Han X, et al. Multimodality MRI-based radiomics approach to predict the posttreatment response of lung cancer brain metastases to gamma knife radiosurgery. *Eur Radiol*. 2022;32(4):2266-2276. doi:10.1007/S00330-021-08368-W

59. Huang CY, Lee CC, Yang HC, et al. Radiomics as prognostic factor in brain metastases treated with Gamma Knife radiosurgery. *J Neurooncol*. 2020;146(3):439-449. doi:10.1007/S11060-019-03343-4

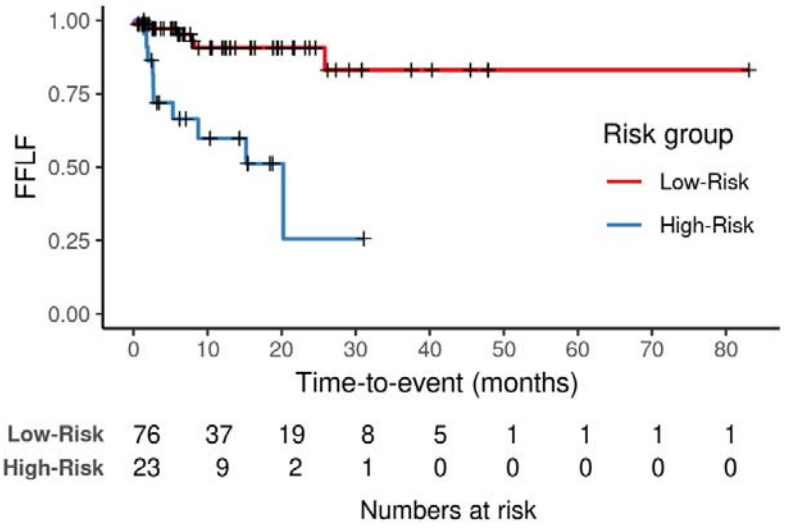
60. Baschnagel AM, Meyer KD, Chen PY, et al. Tumor volume as a predictor of survival and local control in patients with brain metastases treated with Gamma Knife surgery: Clinical article. *J Neurosurg*. 2013;119(5):1139-1144. doi:10.3171/2013.7.JNS13431

Figure 1: Summarized overview of our workflow



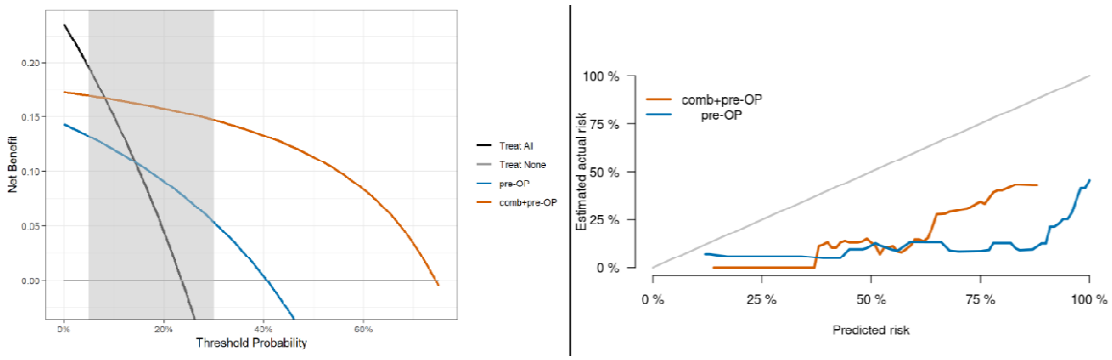
After manual and automatic definition of the volume of interest (VOI), we extracted 104 original features from each metastasis and edema segmentation. We reduced the number of features in each set with MRMR. Furthermore, we added up to eight clinical features and combined all features into multiple different feature sets. The optimal number of features in each set was determined with a nested cross-validation. The optimal parameters for our selected learners were chosen based on a 5-fold cross-validation. The best parameters for each learner-feature-combination were tested in the external test cohort.

Figure 2: Kaplan Meier analysis



We created dichotomous predictions of the *comb+pre-OP* ENR model by using the 66th percentiles of the continuous risk ranks in the training cohort as cutoffs for patient stratification. We found a significant difference in freedom from local failure (FFLF) between the predicted low- and high-risk groups ($p < 0.001$) in the multicenter external test cohort. After 24 months, we found a FFLF of 91% and 26% in the groups, respectively.

Figure 3: Decision curve analysis (left) and calibration curve (right)



Using the same groups as in Figure 2, we found a net benefit of our predictive model compared to treating all patients in the relevant threshold range from five to 30% through decision curve analysis (left). A decision model shows a clinical benefit if the respective curve shows larger net benefit values than reference strategies. The combination of radiomic features derived from the *metastasis*, *edema*, and *clinical* features (*comb+pre-OP*) resulted in a higher net benefit compared to using only the clinical *Pre-OP* features and treating all patients or none. The calibration curve on the right was created by transforming the continuous risk rank predicted by the best *comb+pre-OP* ENR model (in orange) and by the clinical *pre-OP* ENR model (in blue) to event probabilities at 24 months. Although both models seem to overestimate the actual risk of our patients, the *comb+pre-OP* model predicts the risk closer to the actual risk.

Table 1: Cohort demographics

	Training-Cohort			Test-Cohort					
Characteristic	**Overall, N = 253[1]**	**TUM, N = 167[1]**	**USZ, N = 86[1]**	**Overall, N = 99[1]**	**FD, N = 5[1]**	**FFM, N = 11[1]**	**FR, N = 18[1]**	**HD, N = 44[1]**	**KSA, N = 21[1]**
Age at RT start	**62 (53, 71)**	62 (53, 71)	62 (54, 69)	**61 (54, 67)**	63 (55, 64)	57 (52, 66)	58 (50, 66)	61 (54, 65)	63 (59, 70)
KPS	**80 (70, 90)**	80 (70, 90)	90 (80, 90)	**90 (80, 90)**	80 (80, 80)	90 (90, 90)	90 (82, 100)	80 (78, 90)	90 (90, 100)
Location									
Frontal	**86 (34%)**	67 (40%)	19 (22%)	**33 (33%)**	1 (20%)	4 (36%)	5 (28%)	14 (32%)	9 (43%)
Temporal	**32 (13%)**	18 (11%)	14 (16%)	**7 (7.1%)**	2 (40%)	0 (0%)	1 (5.6%)	2 (4.5%)	2 (9.5%)
Parietal	**47 (19%)**	28 (17%)	19 (22%)	**20 (20%)**	2 (40%)	1 (9.1%)	1 (5.6%)	13 (30%)	3 (14%)
Occipital	**27 (11%)**	12 (7.2%)	15 (17%)	**12 (12%)**	0 (0%)	2 (18%)	3 (17%)	5 (11%)	2 (9.5%)
Cerebellar	**56 (22%)**	39 (23%)	17 (20%)	**24 (24%)**	0 (0%)	4 (36%)	5 (28%)	10 (23%)	5 (24%)
Other	**5 (2.0%)**	3 (1.8%)	2 (2.3%)	**3 (3.0%)**	0 (0%)	0 (0%)	3 (17%)	0 (0%)	0 (0%)
Primary Diagnosis									
NSCLC	**89 (35%)**	37 (22%)	52 (60%)	**39 (39%)**	3 (60%)	6 (55%)	2 (11%)	19 (43%)	9 (43%)
Melanoma	**47 (19%)**	24 (14%)	23 (27%)	**9 (9.1%)**	1 (20%)	1 (9.1%)	1 (5.6%)	2 (4.5%)	4 (19%)
RCC	**11 (4.3%)**	9 (5.4%)	2 (2.3%)	**8 (8.1%)**	0 (0%)	1 (9.1%)	2 (11%)	3 (6.8%)	2 (9.5%)
Breast	**34 (13%)**	33 (20%)	1 (1.2%)	**19 (19%)**	0 (0%)	3 (27%)	5 (28%)	9 (20%)	2 (9.5%)
GI	**26 (10%)**	26 (16%)	0 (0%)	**11 (11%)**	0 (0%)	0 (0%)	4 (22%)	5 (11%)	2 (9.5%)
Other	**46 (18%)**	38 (23%)	8 (9.3%)	**13 (13%)**	1 (20%)	0 (0%)	4 (22%)	6 (14%)	2 (9.5%)
Residual areas	**66 (26%)**	66 (40%)	0 (0%)	**21 (21%)**	1 (20%)	2 (18%)	1 (5.6%)	11 (25%)	6 (29%)
Concurrent CTX	**15 (5.9%)**	8 (4.8%)	7 (8.1%)	**3 (3.0%)**	0 (0%)	2 (18%)	0 (0%)	1 (2.3%)	0 (0%)
Concurrent ITX	**10 (4.0%)**	6 (3.6%)	4 (4.7%)	**13 (13%)**	0 (0%)	3 (27%)	0 (0%)	9 (20%)	1 (4.8%)
EQD2	**43.75 (37.50, 43.75)**	43.75 (43.75, 43.75)	37.50 (37.50, 37.50)	**37.5 (34.7, 42.0)**	37.5 (37.5, 40.0)	34.7 (28.9, 36.0)	37.5 (37.5, 42.3)	38.3 (34.7, 43.8)	40.0 (31.2, 40.0)
Total brain tumor burden (ml)	**11 (5, 21)**	11 (5, 20)	12 (7, 23)	**13 (5, 24)**	41 (23, 48)	17 (10, 21)	14 (5, 28)	9 (4, 15)	14 (6, 33)

[1]Median (IQR); n (%)

We split our patients into two cohorts: a training cohort (TUM: Klinikum rechts der Isar of the Technical University of Munich, USZ: University Hospital Zurich) and a multicenter external test cohort (FD: General Hospital Fulda, FFM: Saphir Radiochirurgie/University Hospital Frankfurt, FR: University Hospital Freiburg, HD: Heidelberg University Hospital, KSA: Kantonsspital Aarau).
We differentiated between six different histologies: non-small cell lung carcinoma (NSCLC, further differentiated into adenocarcinoma, non-adenocarcinoma, and not further specified), melanoma, renal cell carcinoma (RCC), breast cancer, gastrointestinal cancer (GI), and others.
There was no significant difference in age, location of the BM, primary diagnosis, residual area after resection, concurrent CTX, and total brain tumor burden between both cohorts. Significant differences were found in the Karnofsky performance status

(KPS, $p < 0.001$), concurrent ITX ($p = 0.002$), and the equivalent dose in 2Gy fractions (EQD2, $p < 0.001$).

Table 2: Performance in internal validation and external testing

Group	Learner	pre-OP	pre-OP + BMV	Post-OP	RT	T1-CE	FLAIR	comb	comb + pre-OP	comb + pre-OP + BMV
5-fold CV	ENR	0.64	0.63	0.63	0.63	0.65	0.47	0.62	**0.67**	0.67
	RF	0.63	0.63	0.63	0.63	0.61	0.58	0.64	0.66	0.66
	xgboost	0.54	0.56	0.53	0.56	0.58	0.55	0.62	0.65	0.64
external test cohort	ENR	0.70 (0.53-0.83)	0.70 (0.54-0.83)	0.65 (0.51-0.82)	0.70 (0.56-0.83)	0.76 (0.63-0.84)	0.50 (NA-NA)	0.69 (0.55-0.80)	**0.77** (0.61-0.87)	0.72 (0.57-0.82)

Parameter tuning and internal validation were performed with ten iterations of a five-fold cross-validation. The 95% confidence intervals (in parenthesis) are based on 10000 bootstrap samples. The combination of ENR learner and *comb+pre-OP* feature set performed best with a mean CI of 0.67. Adding BMV did not improve performance. By ranking the performance of the models across all feature sets, we identified ENR as the best learner and, therefore, tested this learner on the external test cohort. Again, the best performance was seen with the *comb+pre-OP* feature set (CI = 0.77). Best performance is printed in bold for the internal and external cohort.

Table 3: Performance in the test set with automated U-Net segmentations and segmentations of only the largest metastasis

Group	Learner	T1-CE	FLAIR	comb	comb + pre-OP	comb + pre-OP + BMV
Manual Segmentation	ENR	0.76 (0.63-0.84)	0.50 (NA-NA)	0.69 (0.55-0.80)	**0.77** (0.61-0.87)	0.72 (0.57-0.82)
U-Net Segmentation	ENR	0.68 (0.54-0.79)	0.43 (0.27-0.58)	0.64 (0.47-0.74)	0.72 (0.54-0.82)	0.67 (0.52-0.79)
largest BM	ENR	0.75 (0.62-0.85)	0.50 (NA-NA)	0.65 (0.55-0.80)	0.72 (0.60-0.86)	0.73 (0.58-0.84)

In addition to using our manual segmentations, we also trained and tested our proposed model on automatically generated U-Net segmentations and segmentations of only the largest BM. Since the clinical feature sets are independent of the segmentation method, they were not added to this analysis. Compared to the manual segmentations, the results were, on average 0.06 and 0.02 points worse, respectively.