Identifying core MRI sequences for reliable automatic brain metastasis segmentation

Josef A Buchner1, Jan C Peeken1,2,3*, Lucas Etzel1,2, Ivan Ezhov4,5, Michael Mayinger6, Sebastian M Christ6, Thomas B Brunner7, Andrea Wittig6, Björn Menze1,2,8, Claus Zimmer9, Bernhard Meyer11, Matthias Guckenberger8, Nicolaus Andratschke6, Rami A El Shafie12,13,14, Jürgen Debus12,13, Susanne Rogers15, Oliver Riesterer15, Katrin Schulze16, Horst J Feldmann16, Oliver Blanck17, Constantinos Zamboglou18,19, Konstantinos Ferentinos20, Angelika Bilger18,19, Anca L Grosu18,19, Robert Wolff21,22, Jan S Kirschke10, Kerstin A Eitz1,2,3, Stephanie E Combe1,2, Denise Bernhardt1,2, Daniel Rücker23, Marie Piraud24, Benedikt Wiestler5,10*, Florian Kofler24,4,5,10*

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

*shared authorships

Correspondence: Josef A. Buchner, E-Mail: j.buchner@tum.de
Mailing address: Klinik und Poliklinik für RadioÖnkologie und Strahlentherapie, Klinikum
rechts der Isar, Technische Universität München (TUM), Ismaninger Str. 22, 81675 München,
Germany

Keywords: brain metastases, MRI sequences, segmentation, CNN, deep learning, U-net,
T1, FLAIR

Short title: Core MRI sequences for automatic brain metastasis segmentation

Funding: This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German
Research Foundation - PE 3303/1-1 (JCP), WI 4936/4-1 (BW)).

Number of Tables: 3
Number of Figures: 3
Supplemental Materials: 1
Number of Supplemental Figures: 1
Number of Supplemental Tables: 3
Abstract:

Background: Many automatic approaches to brain tumor segmentation employ multiple magnetic resonance imaging (MRI) sequences. The goal of this project was to compare different combinations of input sequences to determine which MRI sequences are needed for effective automated brain metastasis (BM) segmentation.

Methods: We analyzed preoperative imaging (T1-weighted sequence ± contrast-enhancement (T1/T1-CE), T2-weighted sequence (T2), and T2 fluid-attenuated inversion recovery (T2-FLAIR) sequence) from 333 patients with BMs from six centers. A baseline 3D U-Net with all four sequences and six U-Nets with plausible sequence combinations (T1-CE, T1, T2-FLAIR, T1-CE+T2-FLAIR, T1-CE+T1+T2-FLAIR, T1-CE+T1) were trained on 239 patients from two centers and subsequently tested on an external cohort of 94 patients from four centers.

Results: The model based on T1-CE alone achieved the best segmentation performance for BM segmentation with a median Dice similarity coefficient (DSC) of 0.96. Models trained without T1-CE performed worse (T1-only: DSC = 0.70 and T2-FLAIR-only: DSC = 0.72). For edema segmentation, models that included both T1-CE and T2-FLAIR performed best (DSC = 0.93), while the remaining four models without simultaneous inclusion of these both sequences reached a median DSC of 0.81-0.89.

Conclusions: A T1-CE-only protocol suffices for the segmentation of BMs. The combination of T1-CE and T2-FLAIR is important for edema segmentation. Missing either T1-CE or T2-FLAIR decreases performance. These findings may improve imaging routines by omitting unnecessary sequences, thus allowing for faster procedures in daily clinical practice while enabling optimal neural network-based target definitions.
Introduction

Brain metastasis (BM) delineation is a time-consuming process in clinical practice and research alike. Automated BM segmentation algorithms can be used to assist in this task. They require only a fraction of the time an experienced clinician needs to perform delineation while achieving an overlap with the reference segmentation within the range of interrater variability [1,2].

We have previously developed a model for the simultaneous segmentation of both contrast-enhancing BMs and surrounding T2 fluid-attenuated inversion recovery (T2-FLAIR) hyperintense edema [1]. Like many other approaches to brain tumor segmentation, such as the BraTS challenge [3] or FeTS [4], our model uses four magnetic resonance imaging (MRI) sequences as input, namely T1, T1 with contrast enhancement (T1-CE), T2 and T2-FLAIR. Using fewer input sequences is clearly advantageous. In clinical practice, individual sequences may not be of the required quality, e.g., due to motion artifacts [5]. Furthermore, while a complete brain imaging protocol averages a scan time of about 21 minutes [6], an adapted protocol with only two sequences can decrease duration by about ten minutes. Also, shorter scan times are, in turn, known to reduce patient motion [5]. In addition, using fewer sequences reduces the amount of data that needs to be processed. This results in faster pre-processing times and leaner neural networks.

Although the administration of MRI contrast agents generally results in fewer and less severe adverse effects than the use of iodine-based computed tomography contrast agents, there are still some adverse reactions including rare, life-threatening anaphylactoid reactions [7]. Their use should therefore be carefully considered. Nevertheless, contrast-enhanced sequences are part of many imaging routines, such as in the radiation therapy planning of brain tumors [8]. Thus, BM segmentation algorithms that work without contrast would be of great use.

While some authors have built neural networks for BM segmentation using only T1-CE, they focused only on the BM itself without the surrounding T2-FLAIR hyperintense edema [9,10]. While edema segmentation currently has no relevance for the radiotherapy (RT) planning of BMs, it can be relevant for glioma [11]. Moreover, the delineation of edema may provide valuable information for downstream analysis with techniques such as radiomics [12] or neural network-based feature extraction.

This project aimed to compare neural networks with different combinations of input sequences for the segmentation of the contrast-enhancing metastasis and the surrounding T2-FLAIR hyperintense edema. All neural networks were tested in a multicenter international external test cohort composed of 94 patients from four different centers to investigate the contribution of different MRI sequences to the segmentation of contrast-enhancing BMs and their surrounding edema.
Methods

Automatic segmentation of brain metastases

In our previous work, we focused on how to improve the detection and segmentation of BMs [1]. This project aimed to quantify the contribution of individual MRI sequences to the quality of segmentation. In the following, we will refer to our previous publication and highlight the changes in our workflow.

AURORA study

Data were collected as part of the A Multicenter Analysis of Stereotactic Radiotherapy to the Resection Cavity of Brain Metastases (AURORA) retrospective study conducted by the Radiosurgery and Stereotactic Radiotherapy Working Group of the German Society for Radiation Oncology (DEGRO) [13]. Inclusion criteria were a resected BM with a known primary tumor and stereotactic RT with radiation dose > 5 Gy per fraction. Exclusion criteria were an interval between surgery and RT > 100 days, premature discontinuation of the RT, and any previous cranial RT. Synchronous unresected BMs were allowed but had to be treated concurrently with stereotactic RT [1]. Institutional ethical approval was obtained (main approval at the Technical University of Munich: 119/19 S-SR; 466/16 S). While the study focuses clinically on the postoperative situation, we analyzed only the preoperative imaging.

Dataset

We received data from 481 patients from seven centers in total (TUM: Klinikum rechts der Isar of the Technical University of Munich, USZ: University Hospital Zurich, FD: General Hospital Fulda, FFM: Saphir Radiochirurgie/University Hospital Frankfurt, FR: University Hospital Freiburg, HD: Heidelberg University Hospital, KSA: Kantonsspital Aarau). As an extension of the previous study, an additional center was included in the test group (FR).

We analyzed preoperative MRI scans only. For our established preprocessing workflow, we needed four MRI sequences from each patient: T1, T1-CE, T2, and T2-FLAIR. Unlike in our last workflow [1], only the T2 sequence was allowed to be missing because it was not available for a large fraction of the cohort. If other sequences besides the T2 sequence or multiple sequences were missing, the patient was excluded. Since all patients from FD had a missing T1 sequence, the whole center was excluded.

The required sequences were available in sufficient quality for a total of 333 patients (69% of total). We divided the patients into a training cohort of 239 patients from two centers (TUM and USZ) and a test cohort of 94 patients from four centers (FFM, FR, HD, and KSA).

Data preprocessing

We used the same established preprocessing workflow as previously [1]. In short, we used BraTS-Toolkit [14] to generate co-registered, skull-stripped sequences with an isotropic resolution of 1 millimeter in BraTS space.

A total of 117 T2 sequences were missing (106 (44%) in the training cohort and eleven (12%) in the test cohort). These were synthesized with a generative adversarial network (GAN) [15] by feeding the remaining three sequences into the GAN. The synthesized sequences passed visual inspection.

Annotation

All images were segmented by a doctoral student (JAB) using the open-source software 3D Slicer (version 4.13.0, stable release, https://www.slicer.org/) [16]. Two separate, non-overlapping labels were segmented: The metastasis label, consisting of the contrast-enhancing metastasis and necrosis, and the T2-FLAIR hyperintense edema label. The segmentations of the test set patients were reviewed by a senior radiation oncologist (JCP).
Sequence combinations

To reduce the number of models to be trained, we did not train with every possible combination of input sequences, but instead only analyzed clinically plausible combinations by following these considerations: To identify the exact outline of the BM, T1-CE is required [17]. To quantify the added benefit of administering contrast agents, a comparison between T1 and T1-CE may provide further insight. If the main interest is edema, T2-FLAIR may be sufficient. Additional sequences may further improve the quality of segmentations. We did not train a T2-only model to prevent neural networks from receiving only synthetic data from some patients without original data as input. The model trained with all four sequences is referred to as baseline for the remainder of this manuscript. Overall, we trained models with the following sequence combinations:

- T1-CE + T1 + T2 + T2-FLAIR (baseline)
- T1-CE only
- T1 only
- T2-FLAIR only
- T1-CE + T2-FLAIR
- T1-CE + T1
- T1-CE + T1 + T2-FLAIR

Neural Network

We kept all training parameters the same as in our previous study [1]. We implemented spatial flips, Gaussian noise, and random affine transformations to augment our training data. As loss function, we chose an equally weighted Dice + Binary Cross Entropy (BCE) loss, as used by Isensee et al. [18]. We trained all networks for a total of 500 epochs. The best model was chosen based on the lowest overall loss in the training set.

All models were trained on a workstation equipped with an Intel 9940X CPU combined with two NVIDIA RTX 8000 GPUs using CUDA version 11.4 in conjunction with Pytorch version 1.13.1 [19] and MONAI version 1.1.0 [20].

Metrics

We calculated the Dice similarity coefficient (DSC) with the Python package pymia [21]. To assess the BM detection performance, we used a pipeline created by Pan et al. [22] to determine the F1-score (F1), sensitivity, and precision. The performance of multiple models was compared with the Kruskal-Wallis rank sum test. Numeric and categorical data in the patient cohorts was compared with the Kruskal-Wallis rank sum test and Pearson's Chi-squared test, respectively.
Results

The mean number of BMs, patient demographics, and the number of patients with synthesized T2 sequences in each center are shown in Table 1.

Table 2 summarizes our model evaluation results. Regarding metastasis segmentation, all models that included T1-CE in their selected sequences showed similar performance, with a small but significant difference (median DSC = 0.93-0.96, p < 0.001). In contrast, the models trained only on T2-FLAIR and only on T1 reached a significantly lower median DSC for the metastasis of 0.72 (IQR = 0.53-0.84) and 0.70 (IQR = 0.45-0.81). The models trained only on T1-CE or T1-CE and T1 performed even better than baseline with a median DSC of 0.96 and 0.95, respectively.

For edema segmentation, all models which included both T1-CE and T2-FLAIR (baseline, T1-CE + T2-FLAIR, T1-CE + T1 + T2-FLAIR) performed best with a median DSC of 0.93. The remaining three models with only one of these two sequences (T1-CE-only, T2-FLAIR-only, T1-CE + T1) reached a median DSC of 0.87-0.89. Again, the T1-only model performed worst with a median DSC of 0.81 (IQR = 0.66-0.87). A segmentation of the metastasis and edema generated by our T1-CE-only model is shown in Figure 1.

When evaluating the metastasis and edema labels as a combined whole lesion label, the T2-FLAIR-only model exhibited only minimally worse performance than the T1-CE + T2-FLAIR model with median DSCs of 0.93 and 0.95, respectively. This demonstrates that the boundary between metastasis and edema rather than the outline of the whole lesion poses a challenge to the T2-FLAIR-only model. The segmentation metrics for the whole lesion label for all models are summarized in Supplementary Table 1. Qualitative inspection of the segmentations supports this thesis (see Supplementary Figure 1).

To check the generalizability of the models, the performance in the individual centers of the test set was compared. As an example, the performance of our T1-CE + T2-FLAIR model for the metastasis and edema labels is shown in Figure 2 for each center separately. The only significant difference (p = 0.045) was found in the T1-CE-only model for metastasis segmentation. The median DSC ranged from 0.94 (FFM and FR) to 0.96 (HD and KSA). Excluding the eleven patients with synthetic T2 showed largely similar results: If there was any change at all, it was a slight increase in DSC of 0.01-0.02. In the baseline model, the only model that included the T2 sequences among the selected sequences, there was no change in median DSC for metastasis and an increase from 0.93 to 0.94 and 0.95 to 0.96 for edema and whole lesion, respectively. The segmentation performance of all 83 patients with four available sequences is shown in Supplementary Table 2.

Table 3 summarizes the BM detection performance. Only mean values are given since the performance was calculated on a per-patient basis and the median performance across all patients was often 1. Patients in our test cohort had 1.4 BMs on average. While all models including T1-CE among their selected sequences showed a high sensitivity of at least 0.96, the T2-FLAIR-only and T1-only models reached only 0.91 and 0.83, respectively. T1-CE-only and T1-CE + T1 detected BMs with a mean precision of 0.97 and 0.92, respectively. In contrast, all models including T2-FLAIR segmented a high number of false positives and therefore achieved a mean precision of only 0.58-0.77. The T1-only model also reached a similar precision of 0.70. As the model with the highest mean number of false positives (1.5), the T2-FLAIR-only model segmented a mean of 2.7 BMs per patient. On the other hand, the T1-CE-only model only labeled 0.06 false positives per patient on average. See Figure 3 for an example patient with five false positives in total segmented by the T2-FLAIR-only model. The F1 showed similar behavior: T1-CE-only and T1-CE + T1 achieved a mean F1 of 0.97 and 0.93 while the remaining models achieved a mean score between 0.64 and 0.82.
Automatic segmentation from preprocessed files took less than 20 seconds on consumer-grade hardware (NVIDIA RTX 3090), regardless of the model selected.
Discussion

We compared neural networks with different clinically plausible combinations of input sequences to determine the influence of individual sequences on metastasis and edema segmentation performance and BM detection performance. For segmenting the metastasis, the sole presence of T1-CE is important, and a T1-CE-only protocol appears to be sufficient with a median DSC of 0.96. In contrast, for the segmentation of the edema, the combination of T1-CE and T2-FLAIR is crucial, while employing only T2-FLAIR as input leads to worse results. T1-only performed worst in all segmentation tasks. Thus, we consider the administration of contrast agents necessary for BM segmentation.

As with metastasis segmentation, the presence of T1-CE was most important for the sensitivity in detecting BMs, and additional sequences did not improve sensitivity. T1-CE-only and T1-CE + T1 detected BMs with the best precision (0.97 and 0.92, respectively). Contrary to expectation, the addition of T2-FLAIR did not improve BM detection performance but instead resulted in more false positives. Together with this article, we are publishing a flexible segmentation tool that chooses the appropriate neural network depending on the availability of sequences.

The mean DSC for both the metastasis and the edema label is on average 0.05 points lower than the median DSC. This shows that while most labels are of very good quality, some outliers reduce the mean. The consistently high performance across the four centers of our test set shows that our models generalize well.

The poor edema segmentation performance of our T2-FLAIR-only model might be explained by the way the model generates labels: As in our previous study [1], the model generates an output for the metastasis and the whole lesion. The edema label is then calculated by subtracting the metastasis label from the whole lesion label to ensure gapless segmentation. Therefore, poor metastasis segmentation will also result in a low DSC in the edema segmentation.

This new segmentation method has some advantages over our previous workflow: Previously, only one sequence was allowed to be missing or corrupted, which was then synthesized using a generative adversarial network (GAN) [15]. While this allows our previous network to be used on patients with only three available MRI sequences, it adds complexity to the preprocessing workflow. In addition, examinations with multiple missing sequences cannot be segmented with the previous workflow. Furthermore, having to acquire fewer sequences for objective metastasis and edema segmentation benefits both patients and physicians.

To our knowledge, no other publication has performed a comparable in-depth analysis of the contribution of individual MRI sequences to the segmentation performance of metastasis and edema labels. However, for example, Pflüger et al. also created a “slim” version of their neural network using only the T1-CE and T2-FLAIR sequences as input in addition to their standard model [23]. They observed a slight but significant decrease in performance for the contrast-enhancing metastasis when using fewer input sequences (median DSC: 0.90 down to 0.89).

Charron et al. compared several databases of single MRI sequences (T1-CE, T1, T2-FLAIR) and combinations of them for the detection and segmentation of BMs [24]. They found that when using a single sequence, T1-CE performed best. When two sequences were used, T1-CE + T2-FLAIR resulted in better sensitivity and fewer false positives. The simultaneous use of all three sequences resulted in the best DSC (0.79) and the lowest number of false positives per patient (7.8). These results are only partially comparable to our results because they focused only on metastasis detection and segmentation. In addition, all data were collected from the same center and there was no external or multicentric test cohort. The difference in mean DSC of their best model (T1-CE + T1+ T2-FLAIR) compared to our similar model (0.79
vs. our 0.90) can be partially explained by the higher number and smaller size of BMs in their dataset. The high proportion of edge voxels in small metastases may make segmentation more difficult.

This work has several limitations: The preprocessing workflow we have been using [14] is designed to work with the established four sequences. For our new models to be viable for use in research, preprocessing pipelines must be created that can work with a reduced or variable number of input sequences. The reference annotations were all created by the same person. Thus, the trained neural network adapted the personal segmentation style of our original rater. Even though the segmentations of the test set were checked by an additional rater, a dataset created by multiple raters may hold even greater validity. Because we focused on the imaging of patients who later underwent surgery, many BMs were often larger than metastases that are primarily treated with RT. Especially when trying to detect smaller BMs, T2-FLAIR may be more important than our experiments suggest.

Despite these limitations, we were able to show that neural networks can segment contrast-enhancing BMs as well as their surrounding edemas with a reduced number of input sequences. For the segmentation of BMs, T1-CE-only appears to provide sufficient segmentation quality. For situations, where the edema segmentation is of relevance, such as glioma RT planning, the combination of T1-CE and T2-FLAIR seems to be particularly suitable, as it offers high segmentation performance for both tumor and edema combined with reduced image acquisition time. These findings can help to adapt RT planning MRI protocols and shorten them, thus speeding up procedures in daily clinical practice. Our tool has been uploaded to GitHub and can be accessed via the following link: https://github.com/HelmholtzAI-Consultants-Munich/AURORA.


A segmentation of the *metastasis* (in blue) and the *edema* (in red) is shown. Even though the network only received the T1-CE sequence as input (shown in axial, sagittal, and coronal orientation in the top left, top right, and bottom left, respectively), the edema has also been correctly segmented as illustrated by the axial view of the T2-FLAIR (bottom right).
The segmentation performance of our T1-CE + T2-FLAIR model was stable across all centers as shown by the consistently high Dice similarity coefficient (DSC). There were no significant differences in metastasis and edema segmentation performance (p = 0.14 and 0.4, Kruskal-Wallis rank sum test). While most segmentations were of good quality with a DSC of around 0.9, there were some outliers. This explains the difference between mean and median DSC.
The segmentation by our T2-FLAIR-only model is shown for a patient with an especially high number of false positives (five in total). On the left, the T1-CE sequence is displayed in the axial and coronal planes. On the right, the same slices of the T2-FLAIR sequence are shown. While the metastasis (in blue) and edema (in red) have been correctly identified and labeled, the model also labeled several false positives.
Table 1: Mean number of brain metastases, patient demographics, and number of synthesized T2 sequences

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Training-Cohort</th>
<th>Test-Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall, N = 239$^1$</td>
<td>TUM, N = 162$^1$</td>
</tr>
<tr>
<td>BMs</td>
<td>1.4 ± 0.7</td>
<td>1.3 ± 0.7</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
</tr>
<tr>
<td>f</td>
<td>117 (49%)</td>
<td>79 (49%)</td>
</tr>
<tr>
<td>m</td>
<td>122 (51%)</td>
<td>83 (51%)</td>
</tr>
<tr>
<td>Age at RT start</td>
<td>62 (53, 71)</td>
<td>62 (53, 71)</td>
</tr>
<tr>
<td>Primary Diagnosis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSCLC</td>
<td>84 (35%)</td>
<td>35 (22%)</td>
</tr>
<tr>
<td>SCLC</td>
<td>1 (0.4%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Melanoma</td>
<td>41 (17%)</td>
<td>23 (14%)</td>
</tr>
<tr>
<td>RCC</td>
<td>11 (4.6%)</td>
<td>9 (5.6%)</td>
</tr>
<tr>
<td>Breast</td>
<td>34 (14%)</td>
<td>33 (20%)</td>
</tr>
<tr>
<td>GI</td>
<td>24 (10%)</td>
<td>24 (15%)</td>
</tr>
<tr>
<td>other</td>
<td>44 (18%)</td>
<td>38 (23%)</td>
</tr>
<tr>
<td>Synthesized T2</td>
<td><strong>106 (44%)</strong></td>
<td>103 (64%)</td>
</tr>
</tbody>
</table>

$^1$Mean ± SD; n (%); Median (IQR)

We split our dataset in a training cohort (TUM: Klinikum rechts der Isar of the Technical University of Munich, USZ: University Hospital Zurich) and a multicenter external test cohort (FFM: Saphir Radiochirurgie/University Hospital Frankfurt, FR: University Hospital Freiburg, HD: Heidelberg University Hospital, KSA: Kantonsspital Aarau). There were no significant differences in sex and age at the start of the radiotherapy ($p = 0.8$ and 0.7, Pearson’s Chi-squared test and Kruskal-Wallis rank sum test) between the six centers of our dataset. All seven histologies (non-small cell lung carcinoma (NSCLC), small-cell lung carcinoma (SCLC), melanoma, renal cell carcinoma (RCC), breast cancer, gastrointestinal cancer (GI), and others) were present in both the training and the test cohort.
We report the mean and median Dice similarity coefficient (DSC) of each model for the metastasis and edema labels. All models including T1-CE perform similarly for metastasis segmentation with a median DSC of 0.93-0.96 with the T1-CE-only model performing best. Adding more sequences does not improve the segmentation quality. The simultaneous presence of T1-CE and T2-FLAIR is important for edema segmentation.

<table>
<thead>
<tr>
<th>Model</th>
<th>Metastasis</th>
<th>Edema</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.90</td>
<td>0.94 (0.89, 0.95)</td>
</tr>
<tr>
<td>T2-FLAIR</td>
<td>0.66</td>
<td>0.72 (0.53, 0.84)</td>
</tr>
<tr>
<td>T1-CE + T2-FLAIR</td>
<td>0.87</td>
<td>0.94 (0.91, 0.96)</td>
</tr>
<tr>
<td>T1-CE</td>
<td>0.91</td>
<td>0.96 (0.93, 0.97)</td>
</tr>
<tr>
<td>T1-CE + T1</td>
<td>0.92</td>
<td>0.95 (0.93, 0.96)</td>
</tr>
<tr>
<td>T1-CE + T1 + T2-FLAIR</td>
<td>0.90</td>
<td>0.93 (0.88, 0.95)</td>
</tr>
<tr>
<td>T1</td>
<td>0.60</td>
<td>0.70 (0.45, 0.81)</td>
</tr>
</tbody>
</table>

\(^1\)Mean | Median (IQR)
### Table 3: Metastasis detection performance

<table>
<thead>
<tr>
<th>Model</th>
<th>F1(^1)</th>
<th>Sensitivity(^1)</th>
<th>Precision(^1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.78 ± 0.23</td>
<td>0.97 ± 0.11</td>
<td>0.72 ± 0.31</td>
</tr>
<tr>
<td>T2-FLAIR</td>
<td>0.64 ± 0.25</td>
<td>0.91 ± 0.21</td>
<td>0.58 ± 0.31</td>
</tr>
<tr>
<td>T1-CE + T2-FLAIR</td>
<td>0.76 ± 0.24</td>
<td>0.97 ± 0.11</td>
<td>0.69 ± 0.31</td>
</tr>
<tr>
<td>T1-CE</td>
<td>0.97 ± 0.10</td>
<td>0.98 ± 0.08</td>
<td>0.97 ± 0.12</td>
</tr>
<tr>
<td>T1-CE + T1</td>
<td>0.93 ± 0.14</td>
<td>0.98 ± 0.09</td>
<td>0.92 ± 0.19</td>
</tr>
<tr>
<td>T1-CE + T1 + T2-FLAIR</td>
<td>0.82 ± 0.21</td>
<td>0.96 ± 0.13</td>
<td>0.77 ± 0.28</td>
</tr>
<tr>
<td>T1</td>
<td>0.70 ± 0.30</td>
<td>0.83 ± 0.33</td>
<td>0.70 ± 0.32</td>
</tr>
</tbody>
</table>

\(^1\)Mean ± SD

The presence of T1-CE was most important for the detection of BMs, as shown by the high sensitivity of 0.96-0.98. Adding the T2-FLAIR sequence resulted in more false positives and therefore lower precision. The T2-FLAIR-only model segmented the most false positives (1.5 on average) and consequently only achieved a precision of 0.58. Since the F1 score considers true positives as well as false positives and false negatives, it showed similar behavior: The T1-CE-only model scored the highest (F1 = 0.97), the addition of T2-FLAIR led to a lower F1 of 0.76, and the T2-FLAIR-only model reached the lowest score of 0.64.