Can Virtual Contrast Enhancement in Brain MRI Replace Gadolinium?

A Feasibility Study

Jens Kleesiek, PhD, MD,* Jan Nikolas Morshuis, BSc,† Fabian Isensee, MSc,‡ Katerina Deike-Hofmann, MD,* Daniel Paech, MD,* Philipp Kickingereder, MD,§ Ullrich Köthe, PhD,† Carsten Rother, PhD,† Michael Forsting,¶ Wolfgang Wick, MD,¶ Martin Bendszus, MD,§ Heinz-Peter Schlemmer, MD,* and Alexander Radbruch, MD*||

Objectives: Gadolinium-based contrast agents (GBCAs) have become an integral part in daily clinical decision making in the last 3 decades. However, there is a broad consensus that GBCAs should be exclusively used if no contrast-free magnetic resonance imaging (MRI) technique is available to reduce the amount of applied GBCAs in patients. In the current study, we investigate the possibility of predicting contrast enhancement from noncontrast multiparametric brain MRI scans using a deep-learning (DL) architecture.

Materials and Methods: A Bayesian DL architecture for the prediction of virtual contrast enhancement was developed using 10-channel multiparametric MRI data acquired before GBCA application. The model was quantitatively and qualitatively evaluated on 116 data sets from glioma patients and healthy subjects by comparing the virtual contrast enhancement maps to the ground truth contrast-enhanced T1-weighted imaging. Subjective evaluation by 2 raters using a 4-point Likert scale showed good to excellent (3–4) results for 91.5% of the enhancing and 92.3% of the nonenhancing gliomas. However, despite the good scores and ratings, there were visual deviations between the virtual contrast maps and the ground truth, including a more blurry, less nodular-like ring enhancement, in low contrast false-positive enhancements of nonenhancing gliomas, and a tendency to omit smaller vessels. These “features” were also exploited by 2 trained radiologists when performing a Turing test, allowing them to discriminate between real and virtual contrast-enhanced images in 80% and 90% of the cases, respectively.

Conclusions: The introduced model for virtual gadolinium enhancement demonstrates a very good quantitative and qualitative performance. Future systematic studies in larger patient collectives with varying neurological disorders need to evaluate if the introduced virtual contrast enhancement might reduce GBCA exposure in clinical practice.

Key Words: deep learning, Bayesian uncertainty, multiparametric MRI, gadolinium-based contrast agents, glioma

Received for publication March 9, 2019; and accepted for publication, after revision, April 15, 2019.

From the 1Division of Radiology, German Cancer Research Center, 1Visual Learning Lab, Heidelberg Collaboratory for Image Processing, Interdisciplinary Center for Scientific Computing, University of Heidelberg; 2Medical Image Computing, German Cancer Research Center; 3Department of Neuroradiology, University of Heidelberg Medical Center, Heidelberg; 4Department of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen, Essen; 5Neurology Clinic, University of Heidelberg Medical Center, Heidelberg, Germany. The authors report no conflicts of interest.

Supplemental digital contents are available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal’s Web site (www.investigativeradiology.com).

Correspondence to: Jens Kleesiek, PhD, MD, Division of Radiology, German Cancer Research Center, Im Neuenheimer Feld 223, 69120 Heidelberg, Germany. E-mail: j.kleesiek@dkfz.de.

Copyright © 2019 Wolters Kluwer Health, Inc. All rights reserved.

ISSN: 0020-9961/190000–0000
DOI: 10.1097/RLI.0000000000000583
acquired before contrast agent application. Using these data, we trained a 3-dimensional (3D) Bayesian neural network (BayesUNet) and evaluated the generated vcT1w maps in comparison to ground truth contrast-enhanced T1w scans (ceT1ws).

MATERIALS AND METHODS

Clinical Data Description

We included imaging data from a total of 82 patients in this study. Some patients were examined several times (up to 5 times), usually while undergoing therapy, leading to a total of 116 data sets used for training and evaluating the model. We separated the images into 3 groups: normal, that is, without pathological findings (n = 30); nonenhancing (n = 39; 16 male, 23 female scans; mean age at examination, 42.5 ± 10.0 years); or enhancing (n = 47; 34 male, 13 female scans; mean age at examination, 55.0 ± 13.8 years) tumors. The tumor entities comprised astrocytoma, oligoastrocytoma, anaplastic astrocytoma, and glioblastoma multiforme. Therapies and interventions included biopsy, radiotherapy, surgery, chemotherapy, and their combinations. It might be noted that we purposefully used a heterogeneous real-world data set for evaluating our method.

Ethics Statement

Based on institutional guidelines and approval by the local medical ethics committee (Faculty of Clinical Medicine, University of Heidelberg), a retrospective analysis of the data was performed. Only pseudonymized imaging data were included in the current study. Apart from aggregated data describing sex, age, treatment, and the composition of tumor entities, no further clinical data were used. The performed data analysis was in accordance with the Declaration of Helsinki.6

Imaging Data

The multiparametric input data comprise native T1w (nT1w), T2w, FLAIR, DWI, as well as SWI. Diffusion-weighted imaging was acquired using 2 different b values (0, 1200 mm²/s), and based on those, the apparent diffusion coefficient (ADC) map was computed. From the SWI, 3 additional sequences were derived: magnitude (SWI MAG), phase (SWI Phase), and a minimum intensity projection (SWI mIP). Thus, in total, 10 channels were used for the model input. For training, we computed the T1w subtraction map (subT1w) using the contrast-enhanced T1w (ceT1w) image, which was obtained using intravenous gadoterate meglumine (Gd-DOTA, Dotarem) with a standard dose of 0.1 mmol/kg body weight.

The data acquisition was performed on 2 different 3T MR Systems (Magnetom Verio and Magnetom Trio; Siemens HealthCare, Erlangen, Germany) with the following specifications: nT1w and ceT1w magnetization prepared rapid acquisition gradient echo sequence with TE/TR = 3.4/1740 milliseconds and a voxel resolution of 1.0 × 0.8 × 0.8 mm; T2w spin echo imaging with TE/TR = 85/5860 milliseconds and a pixel size of 0.63 × 0.63 mm, slice thickness of 5 mm; FLAIR imaging with TE/TR = 133/8500 milliseconds and a pixel size of 0.9 × 0.9 mm, slice thickness of 5 mm; DWI with TE/ TR = 90/5300 milliseconds and a pixel size of 1.8 × 1.8 mm, slice thickness of 5 mm; SWI with TE/TR = 19.7/27 milliseconds and a pixel size of 0.7 × 0.7 mm, slice thickness of 5 mm.

Data Preprocessing

All sequences of the multiparametric data were coregistered to the nT1w images. A rigid 6-DOF registration was performed using the BRAINSFit algorithm.25 All data were resampled to 1 × 1 × 1 mm using spline interpolation (volume size, 160 × 256 × 192 voxels). The registration accuracy was visually confirmed. For each data set, a brain mask was created based on the nT1w image and applied to all channels. The intensity values of the voxels within the brain mask were normalized to zero mean and unit variance. Outliers were clipped and the intensities of the voxels were then rescaled to the range [0, 1].

Subtraction maps (subT1w) were created by subtracting the nT1w from the ceT1w data. Values smaller than 0 were clipped and outliers were removed. The subtraction was rescaled to be in the range [−1, 1].

Data Augmentation

During training, random reflections along the x, y, and z axes were performed with a probability of 0.5. At each step, 4 minibatches of patch-size 10 × 96 × 96 × 96 (x, y, z) were randomly selected from the augmented data and fed into the network.

Deep Learning Architecture

Model Description

We trained a fully convolutional 3D BayesUNet with dropout to predict a virtual contrast enhancement map (vcSub) as well as an uncertainty map. The architecture is similar to a model proposed by Nair et al.24 A sketch of the model and its parameters is shown in Figure 1. Multiparametric imaging data with 10 channels (see aforementioned data) was used as input to the model. All 10 channels were acquired before GBCA administration. The training signal was given by the subT1w maps. The model output vcSub is akin to a subtraction map. It was added to the nT1w to obtain a vcT1w image.

Training of the Model

During training, the model parameters were optimized by minimizing a loss function that compares the model’s prediction vcSub with the ground truth subT1w. We used a method similar to Kendall et al25 to determine the uncertainty of the predictions. Instead of using a Gaussian likelihood to model the uncertainty, we used a Laplacian likelihood. The loss function is therefore given by

\[ L(θ) = \frac{1}{D} \sum_{i=1}^{D} \exp(-s_i) |y_i - \hat{y}_i| + s_i, \]

where \( \hat{y}_i \) is the value of voxel \( i \) in the ground truth, \( D \) is the number of predicted voxels, and \( [y_i, \hat{y}_i] \) are the output parameter of the model for each voxel \( i \), \( y_i \) is the current estimate of vcSub, and \( s_i \) is used for calculating the uncertainty:

\[ u_i = 2 \cdot \exp(2 \cdot s_i). \]

In regions in which no confident prediction of vcSub is possible, possibly due to lack of information in the input data, the uncertainty tends to be high. This can be exploited for judging the reliability of the virtual contrast maps.

The loss of the minibatches was used for backpropagation. Parameters were updated using AMSGrad, which displays a better behavior than the more common choice ADAM in situations where informative minibatches are rare.26

We performed a stratified 10-fold cross-validation. For each training iteration, the data set was split into a training set and a test set with ~90% and ~10% of the data, respectively. To control for overfitting during training, 4 randomly selected data sets from the training set were not used for parameter optimization and instead left out for validation. Longitudinal data from a single patient was either only in the training set or only in the test set. This guarantees that the predictions are made on previously unseen data that are uncorrelated to the data in the training set.

For each cross-validation iteration, the parameters of the model were optimized until convergence. Each training cycle took on average 147 ± 26 epochs, where each epoch consisted of 400 steps. After convergence the data previously used for validation were added to the training set and training was continued for 10 epochs.
CUDA 9.0, and cuDNN 7.1.2 on Ubuntu Linux

For the differing ratings, the raters agreed on a consensus. This

Virtual Contrast Enhancement in Radiology

ively. The scans of

www.investigativeradiology.com

3

SWI) used in our experiments. The network has 2 output channels, one for the virtual contrast enhancement map (upper right) using a leaky

Volume 00, Number 00, Month 2019

This paper can be cited using the date of access and the unique DOI number which can be found in the footnotes.

Copyright © 2019 Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

irr.

To assess the subjective visual quality, a board-certified neuroradiologist (A.R. with 10 years of experience) and a radiologist (J.K. with

5 years of experience) rated the predicted virtual contrast enhancement maps. For this purpose, we took 20 images, 10 for diagnosis of

enhancing and nonenhancing tumors, ratings were performed combining general image quality and degree of visual conformity of the tumor

region, using a 4-point Likert scale covering 1 (none), 2 (moderate), 3 (good), and 4 (excellent) ratings. To determine the interrater agreement,

the intraclass correlation coefficient was computed using the R package irr.

For the differing ratings, the raters agreed on a consensus. This consensus rating was used for correlation with the quantitative scores. The

consensus ratings were also dichotomized into low (1–2) and high (3–4) ratings, and the percentage of high ratings was calculated for

enhancing and nonenhancing tumors, respectively. The scans of the normal group were evaluated for their plausibility and screened for

systematic failures.

Experiment to Determine the Influence of MRI Sequences on Prediction

To determine the influence of the individual MRI sequences on the prediction, we retrained the BayesUNet leaving out channels of rel-

ated groups, thereby replacing the voxel values with 0.5 for the respective channels. The related groups were as follows: nT1w; T2w, FLAIR;

DWI b = 0; DWI b = 1200, ADC; and SWI, SWI MAG, SWI mIP, SWI Phase. Based on comparison of the scores (AUC, SSIM, and PSNR) of the model with omitted channels to the scores obtained from the model trained with all available input data, we inferred the importance of the groups for prediction, admitting that this is only an indirect measure. The score evaluation was performed only for the tumor region.

Turing Test—Can Radiologists Tell the Difference Between Real and Virtual Contrast-Enhanced Images?

We asked 2 radiology residents, in their third (K.D.H.) and fourth year (D.P.), if they can tell the difference between real ceT1w and predicted vcT1w images. For this purpose, we took 20 images, 10 randomly chosen from each category. The readers did not know about the proportionate composition of the test set. The images were presented in random order in a viewer displaying the axial plane only. The blinded readers were allowed to scroll and to perform

Applying the DL Model

For prediction, a complete 3D multiparametric MRI data set is passed 10 times through the model. Through the randomness intro-

duced by the dropout layers, 10 Monte Carlo samples of the prediction are obtained. The mean of these samples represents the vcSub map. The variance of the samples, though also interesting, was not further investi-

gated in the current study. For the prediction of the uncertainty map, an additional forward-pass is performed without dropout. Due to the Monte Carlo sampling, prediction of a single brain volume takes approx-

imately 30 seconds. This could be improved using multiple graphics processing unit (GPUs) or using GPUs with bigger RAM.

For training and testing of the neural network, we used the PyTorch framework, CUDA 9.0, and cuDNN 7.1.2 on Ubuntu Linux servers with NVIDIA GTX 1080 Ti or Titan Xp GPUs.

Evaluation and Statistical Analysis

Quantitative scores were computed by comparing the model output vcSub to the ground truth subtraction maps subT1w. We computed the structural similarity index (SSIM), the peak signal-to-noise ratio

(PSNR), and the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. The PSNR measures the voxelwise differ-

ence, AUC captures the sensitivity and specificity, and SSIM compares nonlocal structural similarity.

To evaluate the scores separately for tumor and nontumor re-

gions, we segmented the enhancing and nonenhancing gliomas using the architecture from Isensee et al.

For the creation of the ROC curve, the ground truth subtraction maps subT1w were binarized. This was done using an Otsu filter, which selected the threshold value dividing the voxels into 2 classes. The AUC curve was then created by comparing the continuous prediction to the binarized ground truth using Scikit-learn.

To assess the subjective visual quality, a board-certified neuroradiologist (A.R. with 10 years of experience) and a radiologist (J.K. with

5 years of experience) rated the predicted virtual contrast enhancement maps vcT1w by comparing them to the ground truth ceT1w scans, tak-

ing into account both subtraction maps (model output vcSub and subT1w) as well as the uncertainty maps. For each data set of the

enhancing and nonenhancing tumors, ratings were performed combining general image quality and degree of visual conformity of the tumor

region, using a 4-point Likert scale covering 1 (none), 2 (moderate), 3 (good), and 4 (excellent) ratings. To determine the interrater agreement, the intraclass correlation coefficient was computed using the R package irr. For the differing ratings, the raters agreed on a consensus. This consensus rating was used for correlation with the quantitative scores. The consensus ratings were also dichotomized into low (1–2) and high (3–4) ratings, and the percentage of high ratings was calculated for enhancing and nonenhancing tumors, respectively. The scans of the normal group were evaluated for their plausibility and screened for systematic failures.

Experiment to Determine the Influence of MRI Sequences on Prediction

To determine the influence of the individual MRI sequences on the prediction, we retrained the BayesUNet leaving out channels of rel-

ated groups, thereby replacing the voxel values with 0.5 for the respective channels. The related groups were as follows: nT1w; T2w, FLAIR;

DWI b = 0; DWI b = 1200, ADC; and SWI, SWI MAG, SWI mIP, SWI Phase. Based on comparison of the scores (AUC, SSIM, and PSNR) of the model with omitted channels to the scores obtained from the model trained with all available input data, we inferred the importance of the groups for prediction, admitting that this is only an indirect measure. The score evaluation was performed only for the tumor region.

Turing Test—Can Radiologists Tell the Difference Between Real and Virtual Contrast-Enhanced Images?

We asked 2 radiology residents, in their third (K.D.H.) and fourth year (D.P.), if they can tell the difference between real ceT1w and predicted vcT1w images. For this purpose, we took 20 images, 10 randomly chosen from each category. The readers did not know about the proportionate composition of the test set. The images were presented in random order in a viewer displaying the axial plane only. The blinded readers were allowed to scroll and to perform
window leveling. The real images were mildly smoothed using a Gaussian kernel-width of 0.5 voxels.

**RESULTS**

### Qualitative and Visual Assessment

Examples of representative cases are shown in Figures 2 and 3, and an animation of real and virtual contrast enhancement next to each other is shown for an entire image volume in a Video (Supplemental Digital Content 1, http://links.lww.com/RLI/A443). When comparing the predicted maps side-by-side to the ground truth ceT1w images, we often yield good to excellent results, reflected by 91.5% of high ratings for the enhancing and 92.3% for the nonenhancing gliomas in the subjective evaluation of the human raters. The average rating of the 2 raters was 3.4 for enhancing and 3.7 for nonenhancing tumors. The intraclass correlation coefficient of the 2 radiologists was 0.63 (95% confidence interval, 0.479–0.739), signaling good agreement.

Still, sometimes enhancing regions were missed or not as boldly captured as in the ground truth data. This is exemplarily shown in Figure 3C for the right frontal lesion. Interestingly, the uncertainty maps usually prevent the reader from missing these regions during an assessment. We further made the observation that the ring enhancement often did not display a nodular-like pattern, as found in the ground truth data. Generally, the virtual contrast enhancement appeared more blurry. Figure 3D displays a correct prediction for a nonenhancing tumor with a focal region of high uncertainty in the insular cortex. In a few cases, nonenhancing tumors displayed a low-contrast false-positive virtual enhancement, as seen in Figure 3E in the right medial temporal lobe. Again, the uncertainty map points the reader to this region.

### Quantitative Evaluation

The average ROC curve over all subjects reaches a specificity of 91.21% and a sensitivity of 91.81% (Fig. 4). The mean AUC for the whole brain is 0.969. Additional scores, PSNR and SSIM, are detailed in Table 1. We compared different regions, that is, the gross tumor area, the whole brain, and the nontumor area, that represents the difference. The scores for the evaluation of the tumor subregions reveal an AUC decrease of −0.11 for the enhancing and −0.13 for the nonenhancing tumors, an SSIM decrease of −0.12 for the enhancing and −0.09 for the nonenhancing tumors. For the PSNR, we observed a decrease of −2.14 dB for the enhancing tumors, but an increase of +6.67 dB for the nonenhancing tumors.

Correlating the scores obtained via visual assessment with the quantitative scores identifies the highest correlation for the AUC, indicated by a Spearman rank correlation coefficient of 0.39 ($P < 0.001$).

### Experiment to Determine the Influence of the MRI Sequence on Prediction

Retraining the BayesUNet by omitting sequence groups revealed the T2w image group (mean AUC decrease of −0.20 for the tumor regions) followed by the DWI group (mean AUC decrease of −0.15 for the tumor regions) to be the most influential sequences for prediction. The AUC results obtained by omitting groups were significantly different from the AUC values obtained by taking all MRI sequences into account, confirmed by a 2-tailed $t$-test ($P < 10^{-6}$).

### Turing Test—Can Radiologists Tell the Difference Between Real and Virtual Contrast-Enhanced Images?

The radiology residents could correctly discriminate between the real ceT1w and the virtual vcT1w images in 80% and 90% of the cases, respectively. Although the results are clear, they were struggling with the individual decisions. They reported that their decision was mostly based on presence or absence of small subcortical vessel structures and a blurriness in the images. The appearance of the tumor enhancement was realistic, and they sometimes had to rely on their gut feeling. Interestingly, in the few errors they made, real images were mistaken to be virtual as well as the other way around.

### DISCUSSION

Using unenhanced multiparametric MRI data, comprising anatomical (nT1w, T2w, FLAIR) and functional scans (DWI, SWI), we demonstrate that a DL architecture is able to predict postcontrast T1w images with high sensitivity and specificity. We refer to the predictions...
as virtual contrast enhancement. Next to a qualitative evaluation, additional quantitative measures, comparing the predicted vcT1w to the ground truth ceT1w scans, confirm our results. Comparison between the scores for the tumor area and the whole brain reveals that enhancing behavior of normal tissue seems to be easier to predict than enhancing behavior of nonenhancing tumor regions (Table 1). One reason could be that anatomical structures, that is, vessels, and their appearance in the different MR sequences are learned by the model.

We demonstrate that enhancing and nonenhancing gliomas show the “correct” behavior in the majority of cases, that is, they are enhancing and nonenhancing in the predicted virtually enhanced maps, respectively. Moreover, when swiftly looking only at vcT1w maps of normal

![FIGURE 3. Example predictions of the virtual contrast enhancement vcSub for enhancing (A–C) and nonenhancing (D, E) gliomas. For details, please see results. The T2w axial slice is shown in the left column, ceT1w and subT1w indicate the ground truth contrast-enhanced scans and the ground truth subtractions, respectively. The right column displays the uncertainty of the BayesUNet.](image-url)
brains, they are almost indistinguishable to their ground truth counterparts acquired with GBCA (Fig. 2), albeit sometimes missing smaller vessel structures. This might be due to a smoothing process during the convolution operations and needs to be investigated further.

Our Turing test showed that trained radiologists can discriminate between real and virtual contrast-enhanced images in a majority of the cases. Their decision was mostly based on small vessel structures and image smoothness and not on the appearance of the enhancing tumor regions. Recently, a generative adversarial network was used to introduce or remove tumors in thoracic computed tomography scans. Despite knowing that the images had been altered, the radiologists could not reliably discriminate between real and altered images. Interestingly, the authors reported that a rater “picked up on a particular pattern,” showing resemblance to our observation of grounding the decision rather in the general appearance of an image than making the decision based on the actual appearance of the tumor.

The difficulty of predicting small vessels also has been encountered by Gong and colleagues. Notably, there are several decisive differences when comparing our work to their approach. To explore if we can completely remove the need for GBCA administration, we rely on information readily available in a comprehensive multiparametric MRI protocol. Instead, they learned a denoising of a low-dose GBCA image solely based on T1w imaging. Next, we take 3D information into account, whereas they used 2-dimensional images and pointed out that using a 3D DL model will most likely result in performance gains.

Analyzing the BayesUNet, we determined the T2w and DWI sequence families to be the most important contributors for the prediction. These imaging sequences tend to capture disturbances in the extra, as well as intracellular compartment, such as perifocal edema (hypointense in T2w-imaging, T2w-shine through in DWI), high tumor cellularity (hypointense in T2w-imaging, hyperintense in diffusion-weighted imaging with lower ADC values), and tumor necrosis (with inhomogeneous T2w signal intensities and, for example, loss of diffusion anisotropy in diffusion-weighted imaging). Such disturbances and the associated MRI signal alterations are very likely to increase the probability of blood-brain barrier breakdowns detected by ceT1w-imaging. This may explain why the DWI and T2w sequence families are important for predicting contrast enhancement. Hints underpinning this hypothesis are supported by studies finding that the T2w/FLAIR mismatch sign correlates with molecular markers or that pseudoprogression of glioblastoma might be discriminated from real progression based on diffusion imaging, which is more than the human visible information might be present in the canonical imaging sequences. However, the contribution of individual MR sequences needs to be investigated in more detail in future studies.

The correlation between the qualitative human ratings and the quantitative scores shows that we do not have a single score fully capturing human radiological perception. The score with the highest correlation to human perception was the AUC score. In general, the task of finding the perfect score that can be useful for automatic quantitative assessments is of utmost importance for model selection. Next to being

### TABLE 1. Quantitative Evaluation

<table>
<thead>
<tr>
<th>Score</th>
<th>Category</th>
<th>Region</th>
<th>BayesUNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>Nonenhancing</td>
<td>Tumor</td>
<td>0.845 ± 0.106</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nontumor</td>
<td>0.978 ± 0.007</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Whole brain</td>
<td>0.978 ± 0.007</td>
</tr>
<tr>
<td>Enhancing</td>
<td>Tumor</td>
<td>0.847 ± 0.105</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nontumor</td>
<td>0.970 ± 0.010</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Whole brain</td>
<td>0.956 ± 0.024</td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>Whole brain</td>
<td>0.978 ± 0.005</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>Whole brain</td>
<td>0.969 ± 0.019</td>
<td></td>
</tr>
<tr>
<td>PSNR [dB]</td>
<td>Nonenhancing</td>
<td>Tumor</td>
<td>30.00 ± 4.357</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nontumor</td>
<td>23.256 ± 0.979</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Whole brain</td>
<td>23.329 ± 0.879</td>
</tr>
<tr>
<td>Enhancing</td>
<td>Tumor</td>
<td>20.145 ± 4.702</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nontumor</td>
<td>22.734 ± 0.946</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Whole brain</td>
<td>22.582 ± 1.311</td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>Whole brain</td>
<td>23.570 ± 0.553</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>Whole brain</td>
<td>22.967 ± 1.162</td>
<td></td>
</tr>
<tr>
<td>SSIM</td>
<td>Nonenhancing</td>
<td>Tumor</td>
<td>0.784 ± 0.069</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nontumor</td>
<td>0.880 ± 0.036</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Whole brain</td>
<td>0.878 ± 0.038</td>
</tr>
<tr>
<td>Enhancing</td>
<td>Tumor</td>
<td>0.738 ± 0.078</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nontumor</td>
<td>0.875 ± 0.026</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Whole brain</td>
<td>0.862 ± 0.029</td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>Whole brain</td>
<td>0.878 ± 0.016</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>Whole brain</td>
<td>0.872 ± 0.031</td>
<td></td>
</tr>
</tbody>
</table>

AUC indicates area under the curve; PSNR, peak signal-to-noise ratio; SSIM, structural similarity index.
cautions when only looking at scores, we also have to keep in mind that using scores as loss functions for the training of ML architectures that do not capture the radiological perception might lead to unsatisfactory results.

In our experiments, we observed that the BayesUNet performed an automatic motion correction. Due to long scanning times and resulting patient exhaustion, the last recorded sequence, in our setting the ceT1w, often suffers from motion artifacts. As we compute the virtual contrast enhancement and add it to the native T1w image, we are able to obtain a “motion corrected” image. Of course, it has to be noted that motion artifacts in any of the native sequences used as input to our network will possibly impair the results. This is due to introduced offsets as well as intensity fluctuations and the same reasoning as for a suboptimal registration applies (see below).

For generating synthetic images, more powerful architectures are likely available. Especially, generative adversarial networks are well known for their capabilities in image synthesis and generation. Yet, generative models have to be taken with caution, as it cannot be guaranteed that they do not “confabulate,” that is, include fake or hidden information in the generated image. Therefore, we decided to rely on a more classical architecture that learns a mapping from the nonenhanced images to the contrast-enhanced T1w data. Simultaneously, we obtain uncertainties using this network, which allows us to monitor the prediction process. These uncertainties show high values for the contrast-enhanced regions, as defined by the ground truth subT1w. Because the model might not be able to infer the exact intensity values within these regions, the uncertainty is high. Future research will address to what extent the prediction of virtual contrast enhancement can be realized and if all information is readily available within canonical MR images.

As the method was only trained with limited data from 2 scanners of the same vendor, it is very unlikely that it is readily applicable to other clinical data sets. To decide if the obtained results generalize, the proposed architecture needs to be trained and evaluated with larger data sets from different vendors and from multiple sites. The most promising approach will be to enhance the training data to become a more diverse data set. Also transfer learning is a viable option to explore. Transfer learning refers to the ability of retraining ML models with only little data, displaying different characteristics than the data used for originally training the model. For instance, it has been demonstrated for the prediction of fluorescent labels from transmitted-light microscope images that a model can be successfully retrained using only scarce data from an independent data source.

Another limitation that should be noted refers to errors introduced by upsampling and coregistration of the images. Especially offsets due to registration errors can lead to incorrect predictions. One way to address this issue could be to employ synthetic MRI techniques that provide not just quantitative information but also images of multiple clinically relevant contrasts. As the images are acquired using a single MR sequence, it is guaranteed that the different quantitative imaging contrasts are perfectly aligned. Furthermore, it was reported that a related technique, T1-relaxometry, might allow identification of contrast-enhancing regions without the use of GBCA more accurately. Thus, combining our approach with quantitative MR may be a promising next step.

Importantly, our current study should be regarded as a feasibility study, and we do not claim that our technique is ready for clinical use, yet. It has been shown for many applications that contrast-enhanced T1w is crucial for an early diagnosis, and often very subtle and small changes of contrast enhancement, such as identification of micrometastases in the brain, substantially impact patient treatment. For example, it has been recently shown by Deike-Hofmann et al that ceT1w yields the highest diagnostic accuracy for identification of micro brain metastases in melanoma. Notably, the tumor lesions investigated in the current project are considerably larger than the small metastases investigated by Deike-Hofmann and colleagues. Other body regions than the brain will also present challenges, especially because registration might be more difficult. Movement and breathing artifacts as well as an impaired DWI quality in some body regions are examples of potential problems. In addition, a larger region has to be covered, which usually comes at the expense of resolution and might also adversely affect the prediction of virtual contrast enhancement maps.

There are also physical limitations, and not all information necessary for a perfect prediction might be available in native sequences. Finally, GBCA enhancement is due to a disruption of the blood-brain barrier and subsequent extravasation of the gadolinium in the extravascular space. As resolution improves, for example, with ultrahigh field scanners, we might come closer to the goal of molecular metabolic imaging. These advances will arguably also benefit the prediction of the virtual contrast enhancement.

There are many clinical examples that would profit from a reduced dose or complete omission of GBCAs, let alone the cost to the health care system associated with the application of contrast agents. Although GBCAs generally have an excellent safety profile, allergic reactions as well as gadolinium deposition occur and responsible medical conduct dictates that the amount of GBCAs administered should be minimized according to a strict risk versus reward assessment. A promising model that might be tested in future studies would be a stepwise reduction of injected GBCAs. T1w images acquired with a reduced dose of GBCAs could be used as an additional channel to create an improved virtual contrast enhancement.

Future studies with larger patient numbers and different pathologies should assess if the usage of virtual contrast enhancement might provide clinically relevant decision support to help reduce the amount of injected GBCAs.

**SUMMARY AND OUTLOOK**

We proposed a DL architecture for predicting virtual contrast enhancement solely based on native multiparametric MRI data. The model demonstrates a very good quantitative and qualitative performance and showed that the prediction of gadolinium enhancement might be feasible in the near future. Further studies in larger patient cohorts with varying neurological diseases need to assess clinical practicability of this novel approach.

**ACKNOWLEDGMENTS**

The authors thank Kai Ueltzhöfer, Andreas Bartsch, Armin Biller, and Jacob M. Murray for their valuable discussions and comments on the manuscript. The authors also thank the NVIDIA corporation for supporting their research by donating a Titan Xp GPU.

**REFERENCES**


