JMI Journal of Medical Imaging

RESEARCH PAPER

Simulation of acquisition shifts in T2 weighted fluid-attenuated inversion recovery magnetic resonance images to stress test artificial intelligence segmentation networks

Christiane Posselt,[a] Mehmet Yigit Avci,[b] Mehmet Yigitsoy,[b] Patrick Schuenke,[c] Christoph Kolbitsch,[c] Tobias Schaeffter,[c,d] and Stefanie Remmele[a,*]

[a]University of Applied Sciences, Faculty of Electrical and Industrial Engineering, Landshut, Germany
[b]deepc GmbH, Munich, Germany
[c]Physikalisch-Technische Bundesanstalt (PTB), Braunschweig and Berlin, Germany
[d]Technical University of Berlin, Department of Medical Engineering, Berlin, Germany

ABSTRACT. **Purpose:** To provide a simulation framework for routine neuroimaging test data, which allows for "stress testing" of deep segmentation networks against acquisition shifts that commonly occur in clinical practice for T2 weighted (T2w) fluid-attenuated inversion recovery magnetic resonance imaging protocols.

Approach: The approach simulates "acquisition shift derivatives" of MR images based on MR signal equations. Experiments comprise the validation of the simulated images by real MR scans and example stress tests on state-of-the-art multiple sclerosis lesion segmentation networks to explore a generic model function to describe the F1 score in dependence of the contrast-affecting sequence parameters echo time (TE) and inversion time (TI).

Results: The differences between real and simulated images range up to 19% in gray and white matter for extreme parameter settings. For the segmentation networks under test, the F1 score dependency on TE and TI can be well described by quadratic model functions ($R^2 > 0.9$). The coefficients of the model functions indicate that changes of TE have more influence on the model performance than TI.

Conclusions: We show that these deviations are in the range of values as may be caused by erroneous or individual differences in relaxation times as described by literature. The coefficients of the F1 model function allow for a quantitative comparison of the influences of TE and TI. Limitations arise mainly from tissues with a low baseline signal (like cerebrospinal fluid) and when the protocol contains contrast-affecting measures that cannot be modeled due to missing information in the DICOM header.

Keywords: magnetic resonance image simulation; artificial intelligence validation; magnetic resonance imaging sequence; multiple sclerosis lesion segmentation; T2w fluid attenuated inversion recovery

Paper 23231GR received Aug. 10, 2023; revised Mar. 1, 2024; accepted Mar. 29, 2024; published Apr. 24, 2024.

*Address all correspondence to Stefanie Remmele, Stefanie.Remmele@haw-landshut.de

1 Introduction

In recent years, several machine learning and deep learning (DL) technologies have passed the approval process for a medical device to support radiologists in the diagnosis of medical images.[1] Still, the reliability of these new medical software devices and the underlying DL networks strongly depends on the training data and how well they represent the variety of real clinical image data (test images). Castro et al.[2] described different sources of "shifts" between training and test domains and among those, the "acquisition shift, resulting from the use of different scanners or imaging protocols, which is one of the most notorious and well-studied sources of dataset shift in medical imaging." This is a well-known challenge ever since researchers try to derive reproducible measurements of physiologic information based on heterogeneous medical image data, e.g., by image harmonization in Radiomics research.[3] In recent years, various studies have been dedicated to minimizing these shifts by domain adaptation methods. For instance, differences between a target and a source domain can be reduced by image preprocessing (e.g., normalizing intensities, or aligning images), by fine-tuning models on target domain data, or by translation of source into target domain images using generative adversarial networks (GANs) or transformers.[4] These methods have been shown to improve the robustness of artificial intelligence (AI) models but do not provide means for systematic testing and quantification of potential (residual) risks during application. Accordingly, several institutions underline the need for test procedures and published concepts for the evaluation of the robustness and transferability of a model to other data domains.[5–8] The ECLAIR guidelines,[8] for example, request "to check robustness to variability of acquisition parameters." This is especially important for magnetic resonance imaging (MRI), because MR acquisition protocols typically have a large number of sequence parameters, which affect the contrast, resolution, and SNR of the acquired images. On the one hand, this allows a wide range of clinical information to be presented by MR images, but on the other hand, it leads to a high heterogeneity between different radiology centers. MR acquisition protocols are often optimized individually at each site and sometimes even for different patients to take patient-specific features (e.g. weight and size) into account.[9] Thus, acquisition parameters may vary even for the same type of scan, hence resulting in different image contrasts. There are guidelines providing recommendations on appropriate MR protocols. Among those, e.g., the recently published MAGNIMS–CMSC–NAIMS consensus guidelines[10] prescribe the contrast weighting [i.e., T2w, T2w fluid-attenuated inversion recovery (FLAIR), and contrast-enhanced T1w] of the scans that need to be included in the "recommended core" of protocols for the examination of patients with multiple sclerosis (MS). Nevertheless, they lack specific information on contrast-affecting parameters, such as echo, repetition, and inversion time (TE, TR, and TI).

A multitude of visualization methods have been developed to identify the features within images that a neural network is most sensitive to.[11] Other methods quantify the uncertainty of a network during image processing.[12] However, there is no test procedure that predicts whether an AI product can be applied to the images of a particular radiology practice without loss of performance, e.g., given their customized imaging protocols. Further, it is currently not possible to determine which acquisition parameters can be changed without compromising the performance of an AI product.

The identification of the influencing factors that a system is most prone to is a well-known problem in the field of process improvement and quality management. It is generally solved by systematic testing based on the "design of experiment (DoE)" concept. DoE is a standardized statistical tool for quality control in Six Sigma processes to systematically evaluate the robustness of a process to its influencing factors (see Ref. 13, Chapter 5.4). It predicts the minimum number of experiments needed to quantify and compare the impact of all influencing factors and their interactions on a system's outcome or performance metric. Combined with dedicated analysis of the results, the dominating factors can be easily identified. However, to optimize the experimental design to the given problem, regression analysis needs to be performed to estimate the underlying model function that quantifies the dependence of the response variable (here: AI network performance) on the process' input (here: acquisition parameters), see Ref. 13, Chapter 5.3.3.6.

Therefore, the foremost objective of this work is to study the dependency of a network to the most relevant contrast-affecting acquisition parameters. In the above-mentioned neuroimaging T2w FLAIR scans for example, the TE and the TI have the strongest influence on the imaging contrast. But how can models be validated against the typical MR protocol variability of routine scans or even stress tested against rare but realistic maximum domain shifts if the related data are not available?

The benchmark dataset CLEVR-XAI aims to create a selective, controlled, and realistic test environment for the evaluation of explainable neural networks in non-medical applications.[14] Similar projects for medical applications have just started.[15] Using machine learning and neural networks for the simulation and synthesis of medical images is a field of intense research. Attempts have already been made to recreate MRI images through simulation and synthesis, e.g., using GANs or variational autoencoders (VAEs), phantoms, and dedicated multi-parametric MR sequences.[16] Other simulators use virtual phantoms, for example from Brainweb and Shepp–Logan, which represent the human brain[17,18] to generate images that represent a particular protocol. The limiting factors in all the above-mentioned approaches, however, are either the limited number of anatomies (Brainweb), the lack of anatomical realism (Shepp-Logan), the dependency on specific software (sequences), hardware (phantoms), or the ability to synthesize the result of arbitrary MRI sequences settings with only one model (GANs, etc.).

The secondary objective of this study is thus the combination of simulation and synthesis to generate artificial MRI data of arbitrary sequence character (i.e., "shift derivatives") from a set of real MR images. These data are finally used to stress test a model against variations of acquisition parameters.

For the sake of simplicity, the experiments in this study are focusing on the simulation of domain shift derivatives of T2w FLAIR scans for different TE and TI values to describe the performance of MS lesion segmentation networks in dependence of these scan parameters.

2 Methods

This work comprises two levels of methodology and experiments (see Table 1). First, the simulation of domain shift derivatives given a real baseline image dataset, and second, the use of these data to stress test state-of-the-art (SOTA) MS lesion segmentation networks against these shifts. Those networks are trained on data (Table 2) of heterogeneous contrast (e.g., from different field strengths and using different acquisition protocols). The stress tests intend to evaluate to what extent the networks are robust to changes of image contrast. The simulated data are validated by real MRI scans. The dependency of the models' performance (F1-score) against changes of the MRI protocol parameters (TI, TE) is modeled by second-order polynomial functions, recommended by the above-mentioned DoE guidelines to quantitatively compare the robustness of the networks against acquisition shifts, by the functions' coefficients.

The MS data used in this study consist of several open MRI benchmark datasets (see Table 2).

Table 1 Research questions, methodology, and experiments.

Research questions	General methodology	Experiments
How well can acquisition shift derivatives of a (real) MRI dataset be modeled?	1. Estimation of tissue properties (tissue segmentation, partial volume tissue fractions PV, relaxation parameters $\vec{\mathbf{p}}_{\mathrm{Relax}}$, texture map S_{Tex}). 2. Simulation of acquisition shifts based on MRI signal equation dependent on arbitrary sequence parameters $\vec{\mathbf{p}}_{\mathrm{Seq}} = (\mathrm{TI}, \mathrm{TE})$.	1:1 Comparison of simulated and real MRI scans in healthy volunteers by average MR signal values in gray matter, white matter, and CSF. Comparison of the heuristic estimates of T1 and T2 in tissue ROIs with those of relaxometry methods and literature.
Is a quadratic model function appropriate to describe the dependence of the F1 score of a segmentation network to acquisition shifts (i.e., changing sequence parameters)?	1. Generation of representative shift derivatives of a real MS dataset. 2. Measurement of F1 as a function of $\vec{\mathbf{p}}_{\mathrm{Seq}} = (\mathrm{TI}, \mathrm{TE},)$ in model tests.	Modeling of $F1(\vec{\mathbf{p}}_{\mathrm{Seq}})$ as a second order polynomial function, using R^2 as a metric to evaluate the model fit.

Table 2 Datasets used in this work. The first dataset (OpenMS* longitudinal) is utilized as baseline data in the simulation, since this is the only dataset, for which all contrast-affecting parameters (TE, TI, and TR) are provided.

	Data	Nr	Description	Source
Baseline and Test Data	OpenMS* (longitudinal)[19]	20	2D FLAIR image: TR = 11000 ms, TE = 140 ms, TI = 2800 ms, FA = 90 deg, sampling: $0.9 \times 0.9 \times 3$ mm^3	1.5 T Philips, University Medical Centre Ljubljana (UMCL)
Training Data[a]	OpenMS (cross-sectional)[20]	30	3D FLAIR image: TR = 5000 ms, TE = 392 ms, TI = 1800 ms, FA = 120 deg, sampling: $0.47 \times 0.47 \times 0.80$ mm^3	3 T Siemens Magnetom Trio, University Medical Center Ljubljana
	Lesion challenge 2015[21]	5	2D FLAIR image: TI = 835 ms, TE = 68 ms, sampling: $0.82 \times 0.82 \times 2.2$ mm^3	3 T Philips, Best, The Netherlands
	Lesion segmentation challenge 2008[22]	20	2D FLAIR image: sampling: $0.5 \times 0.5 \times 0.5$ mm^3	3 T Siemens
	MSSEG-2[23]	40	3D FLAIR image	1.5 T and 3 T GE, Philips, Siemens
	NAMIC[24]	4	3D FLAIR image: sampling: $1 \times 1 \times 1$ mm^3	3 T Siemens Magnetom Trio

[a] The data were split up randomly into 80% and 20% fractions for training and validation of the networks. See Section 2.2.3 for more detail.

2.1 Concept of Image Generation to Mimic Acquisition Shifts

Data simulation uses an *in-vivo* MRI scan (baseline data) and mimics changes in that baseline scan in response to changing sequence parameters. The concept of image generation is based on the following equation:

$$S(\vec{\mathbf{r}}) = \kappa \cdot \left(\left(\sum_{t=1}^{Nr\,\text{Tissues}} PV_t(\vec{\mathbf{r}}) \cdot s_{\text{FLAIR},t}(\vec{\mathbf{p}}_{\text{Tis},t}, \vec{\mathbf{p}}_{\text{Seq}}) \right) + S_{\text{Tex}}(\vec{\mathbf{r}}) \right)$$
$$= \left(\left(\sum_{t=1}^{Nr\,\text{Tissues}} PV_t(\vec{\mathbf{r}}) \cdot \kappa \cdot s_{\text{FLAIR},t}(\vec{\mathbf{p}}_{\text{Tis},t}, \vec{\mathbf{p}}_{\text{Seq}}) \right) + \kappa \cdot S_{\text{Tex}}(\vec{\mathbf{r}}) \right), \tag{1}$$

with $S(\vec{\mathbf{r}})$, being the simulated signal at pixel position $\vec{\mathbf{r}} = (x, y, z)$. The contribution $s_{\text{FLAIR},t}$ of each tissue t to the signal of a pixel or voxel is weighted with its local volume fraction $PV_t(\vec{\mathbf{r}})$. κ is the (typically unknown) digital imaging and communications in medicine (DICOM) scaling factor. The texture map $S_{\text{Tex}}(\vec{\mathbf{r}})$ is introduced to approximate all texture influences other than tissue, e.g., based on artifacts, field inhomogeneities, noise, etc. The entire image generation process therefore consists of two different steps (Fig. 1). The first step comprises the preliminary estimation of these tissue properties followed by the second step, the final image simulation according to Eq. (1).

$s_{\text{FLAIR},t}(\vec{\mathbf{p}}_{\text{Tis},t}, \vec{\mathbf{p}}_{\text{Seq}})$ is the signal as determined by the sequence and the tissue properties, i.e., the parameters $\vec{\mathbf{p}}_{\text{Tis},t} = (\rho_t, T1_t, T2_t)$ of the underlying tissue t [like the spin density ρ and relaxation parameters T1 and T2 of gray matter (GM), white matter (WM), cerebrospinal fluid (CSF), and lesion]. $s_{\text{FLAIR},t}$ is given by the T2w FLAIR signal equation in Eq. (2) as published in Ref. 25

$$s_{FLAIR,t}(\vec{\mathbf{p}}_{Tis,t}, \vec{\mathbf{p}}_{Seq}) = \rho_t \cdot \left(1 - 2 \cdot \exp\left(-\frac{TI}{T1_t}\right) + \exp\left(-\frac{(TR - TE_{\text{last}})}{T1_t}\right)\right) \cdot \exp\left(-\frac{TE}{T2_t}\right) \tag{2}$$

with $\vec{\mathbf{p}}_{\text{Seq}} = (\text{TE}, \text{TI}, \text{TR}, \dots)$, i.e., the sequence parameters.

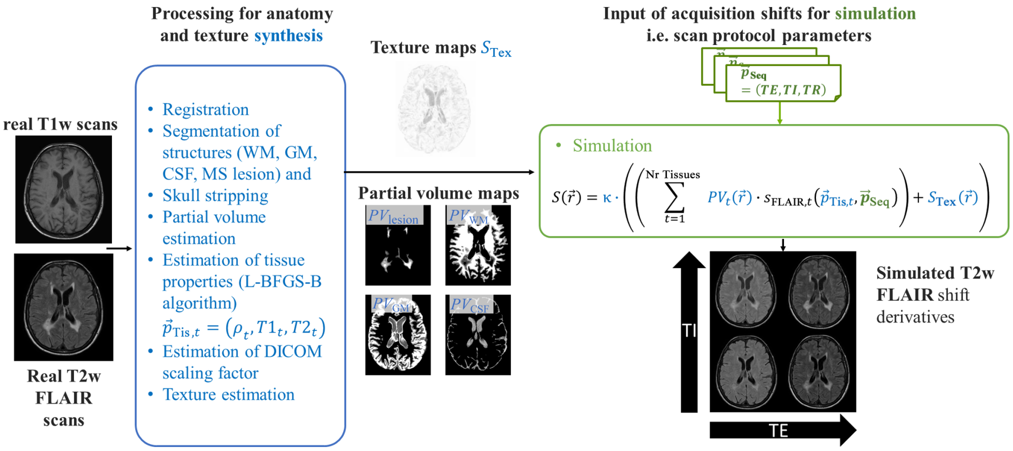
Processing for anatomy and texture synthesis
real T1w scans
Real T2w FLAIR scans
• Registration
• Segmentation of structures (WM, GM, CSF, MS lesion) and
• Skull stripping
• Partial volume estimation
• Estimation of tissue properties (L-BFGS-B algorithm)
$\vec{p}_{\mathrm{Tis},t} = (\rho_t, T1_t, T2_t)$
• Estimation of DICOM scaling factor
• Texture estimation
Texture maps S_{Tex}
Partial volume maps
PV_{lesion}
PV_{WM}
PV_{GM}
PV_{CSF}
Input of acquisition shifts for simulation i.e. scan protocol parameters
$\vec{p}_{\mathrm{Seq}} = (TE, TI, TR)$
• Simulation
$S(\vec{r}) = \kappa \cdot \left(\left(\sum_{t=1}^{\mathrm{Nr\ Tissues}} PV_t(\vec{r}) \cdot s_{\mathrm{FLAIR},t}(\vec{p}_{\mathrm{Tis},t}, \vec{p}_{\mathrm{Seq}}) \right) + S_{\mathrm{Tex}}(\vec{r}) \right)$
TI
TE
Simulated T2w FLAIR shift derivatives

Fig. 1 Acquisition shifts of a real baseline dataset are simulated based on the MRI signal equation of a T2w FLAIR sequence. The signal contribution of each tissue t is scaled by its volume fraction PV_t and enriched by a texture map S_{Tex}. All influences other than those of the sequence (anatomic structures, DICOM scaling or texture) are synthesized (blue box) from the real baseline scan prior to simulation (green box).

2.1.1 *Simulation and synthesis methods*

Equations (1) and (2) contain a number of tissue parameters that must be represented as realistic as possible for the data generation process but cannot be easily simulated (e.g., anatomical structures, lesion sizes, and locations). The idea behind the proposed generative approach is thus to combine image synthesis and simulation as follows.

1. Synthesis: anatomy and disease related signal contributions are derived from a real MR baseline dataset S_m. These data are used to estimate:
   - the partial volume maps $PV_t(\vec{\mathbf{r}})$ using a partial volume estimation method based on Ref. 26 (see next section). The approach requires an additional T1w scan, which is also included in the above-mentioned “recommended core” protocols of MS examinations. The approach further requires prior tissue segmentation.
   - the DICOM scaling factor κ of the baseline T2w FLAIR scan and
   - $S_{\mathrm{Tex}}(\vec{\mathbf{r}})$, to mimic other texture influences (e.g., noise and artifacts).
2. Simulation of all signal contributions that are affected by the sequence and the choice of parameters.
   - Simulation of acquisition shifts is performed through variation of $\vec{\mathbf{p}}_{\mathrm{Seq}}$ in $s_{\mathrm{FLAIR},t}(\vec{\mathbf{p}}_{\mathrm{Tis},t}, \vec{\mathbf{p}}_{\mathrm{Seq}})$ using Eq. (2). $T1$ and $T2$ are set to random values within a realistic range.

2.1.2 *Partial volume estimation*

For estimation of the partial volume fractions of each tissue, we apply the method described in Ref. 26. This approach requires that a signal rise or decline from one region to the other is unique for one kind of tissue-tissue interface. However, in case the brain contains lesions, a rise of signal when leaving the WM region may be attributed to either a WM-lesion or a WM-GM interface. The partial volume maps are thus generated in two steps, assuming that lesions are solely located in and surrounded by WM.[27] First, as required by the approach, segmentation masks are created. We used Synthseg[28] for segmentation of normal tissues, and expert lesion masks were provided through the datasets.[29] Second, the T1w scans are used to estimate the PV-maps $\mathrm{PV}_{\mathrm{WM1}}$, $\mathrm{PV}_{\mathrm{GM}}$, and $\mathrm{PV}_{\mathrm{CSF}}$ of normal tissue. Lesion pixels might be falsely assigned to the PV-map of GM, which can be easily corrected by setting the GM maps to 0 at all lesion pixels as given by segmentation. Third, WM and lesion ROIs are extracted from the FLAIR images and are fed through the PV-algorithm, to obtain another $\mathrm{PV}_{\mathrm{WM2}}$ and $\mathrm{PV}_{\mathrm{lesion}}$ map. The final $\mathrm{PV}_{\mathrm{WM}}$ is initialized with $\mathrm{PV}_{\mathrm{WM1}}$.

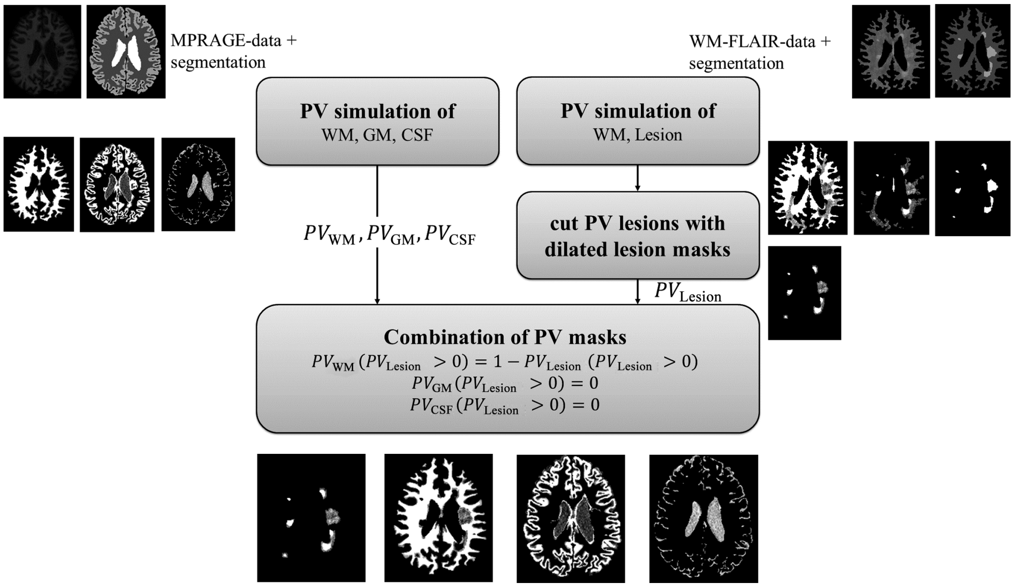
MPRAGE-data + segmentation
WM-FLAIR-data + segmentation
PV simulation of WM, GM, CSF
PV simulation of WM, Lesion
cut PV lesions with dilated lesion masks
$PV_{\mathrm{WM}}, PV_{\mathrm{GM}}, PV_{\mathrm{CSF}}$
PV_{Lesion}
Combination of PV masks
$PV_{\mathrm{WM}}(PV_{\mathrm{Lesion}} > 0) = 1 - PV_{\mathrm{Lesion}}(PV_{\mathrm{Lesion}} > 0)$
$PV_{\mathrm{GM}}(PV_{\mathrm{Lesion}} > 0) = 0$
$PV_{\mathrm{CSF}}(PV_{\mathrm{Lesion}} > 0) = 0$

Fig. 2 Partial volume (PV) maps for normal tissue are determined based on a T1w scan and the method described in Ref. 26. The PV map for the lesion is estimated using the same method and a WM-lesion segment of the T2w FLAIR scan, where lesions differentiate better from the WM background. Fusing all PV information yields the final PV maps.

Finally, in pixels, where $\mathrm{PV}_{\mathrm{lesion}} > 0$, the partial volume fraction in WM is then set to $\mathrm{PV}_{\mathrm{WM}} = 1 - \mathrm{PV}_{\mathrm{lesion}}$. All steps are summarized in Fig. 2.

2.1.3 *Estimation of the DICOM scaling factor κ and the texture map $S_{\mathrm{Tex}}(\vec{\mathbf{r}})$*

A simplified version of Eq. (1) describes the signal of those pixels of the real baseline image S_m that contain only one tissue fraction ($PV = 1$)

$$S_{m,t}(\vec{\mathbf{r}}) = \underbrace{\kappa \cdot 1 \cdot s_{\mathrm{FLAIR},t}(\vec{\mathbf{p}}_{\mathrm{Tis},t}, \vec{\mathbf{p}}_{\mathrm{Seq}})}_{\text{Signal term}} + \underbrace{\kappa \cdot S_{\mathrm{Tex}}(\vec{\mathbf{r}})}_{\text{Texture term}}. \quad \text{for all } \vec{\mathbf{r}}_{PVt=1}, \quad \text{where } PV_t(\vec{\mathbf{r}}) = 1. \tag{3}$$

Since both S_{Tex} and $\vec{\mathbf{p}}_{\mathrm{Tis},t}$ are unknown, the problem of computing S_{Tex} is overdetermined. We solve this by introducing the assumption that signal variations are primarily caused by noise and thus the average texture $\overline{S}_{\mathrm{Tex}}(\vec{\mathbf{r}}_{\mathrm{PV}t=1})$ in this region is 0. Eq. (3) can then be written as

$$\overline{S}_{m,t}(\vec{\mathbf{r}}_{\mathrm{PV}t=1}) = \kappa \cdot s_{\mathrm{FLAIR},t}(\vec{\mathbf{p}}_{\mathrm{Tis},t}, \vec{\mathbf{p}}_{\mathrm{Seq}}). \tag{4}$$

This allows for a preliminary estimation of the apparent tissue parameters $\tilde{\tilde{\vec{\mathbf{p}}}}_{\mathrm{Tis},t}$ from the ratio of average real and simulated signals for different tissues t [the ratio eliminates the unknown κ in Eq. (4)], or more precisely by comparing the real and simulated contrast metrics given in the following equations:

$$C_{\mathrm{sim},t_1,t_2} = \frac{s_{\mathrm{FLAIR},t_1}(\vec{\mathbf{p}}_{\mathrm{TisEst},t_1}, \vec{\mathbf{p}}_{\mathrm{Seq}}) - s_{\mathrm{FLAIR},t_2}(\vec{\mathbf{p}}_{\mathrm{TisEst},t_2}, \vec{\boldsymbol{p}}_{\mathbf{Seq}})}{s_{\mathrm{FLAIR},t_1}(\vec{\mathbf{p}}_{\mathrm{TisEst},t_1}, \vec{\mathbf{p}}_{\mathrm{Seq}}) + s_{\mathrm{FLAIR},t_2}(\vec{\mathbf{p}}_{\mathrm{TisEst},t_2}, \vec{\boldsymbol{p}}_{\mathbf{Seq}})}, \tag{5}$$

$$C_{m,t_1,t_2} = \frac{\overline{S}_{m,t_1}(\vec{\mathbf{r}}_{PVt_1=1}) - \overline{S}_{m,t_2}(\vec{\mathbf{r}}_{PVt_2=1})}{\overline{S}_{m,t_1}(\vec{\mathbf{r}}_{PVt_1=1}) + \overline{S}_{m,t_2}(\vec{\mathbf{r}}_{PVt_2=1})}. \tag{6}$$

The parameters of $\vec{\mathbf{p}}_{\mathrm{Tis},t}$ are optimized to minimize the cost function

$$\begin{aligned}&(C_{\mathrm{sim,GM,WM}} - C_{m,\mathrm{GM,WM}})^2 + (C_{\mathrm{sim,CSF,WM}} - C_{m,\mathrm{CSF,WM}})^2 \\ &+ (C_{\mathrm{sim,Lesion,WM}} - C_{m,\mathrm{Lesion,WM}})^2 \rightarrow \min.\end{aligned} \tag{7}$$

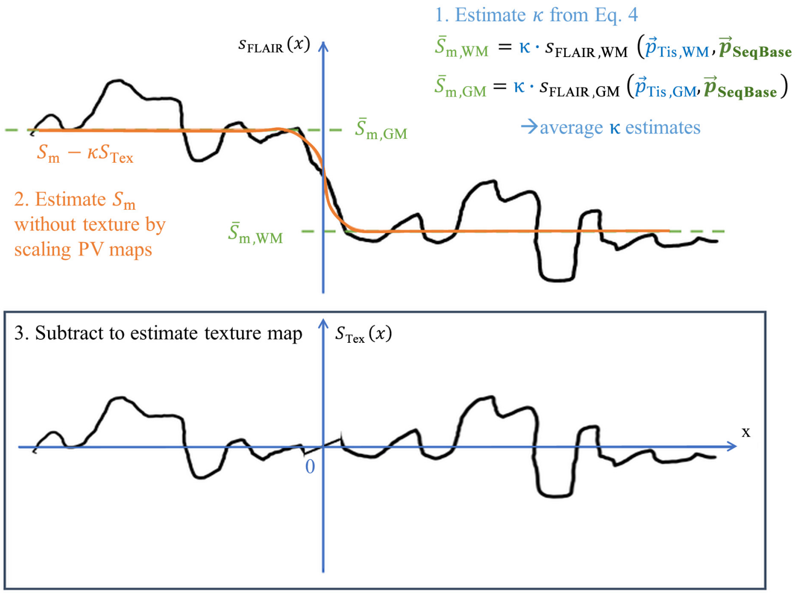
1. Estimate κ from Eq. 4
$\bar{S}_{\mathrm{m,WM}} = \kappa \cdot s_{\mathrm{FLAIR,WM}}\left(\vec{p}_{\mathrm{Tis,WM}}, \vec{\boldsymbol{p}}_{\mathbf{SeqBase}}\right)$
$\bar{S}_{\mathrm{m,GM}} = \kappa \cdot s_{\mathrm{FLAIR,GM}}\left(\vec{p}_{\mathrm{Tis,GM}}, \vec{\boldsymbol{p}}_{\mathbf{SeqBase}}\right)$
→average κ estimates
$s_{\mathrm{FLAIR}}(x)$
$\bar{S}_{\mathrm{m,GM}}$
$S_{\mathrm{m}} - \kappa S_{\mathrm{Tex}}$
2. Estimate S_{m} without texture by scaling PV maps
$\bar{S}_{\mathrm{m,WM}}$
3. Subtract to estimate texture map
$S_{\mathrm{Tex}}(x)$
0
x

Fig. 3 Method to estimate the texture of an MR image by subtracting the estimated signal in consideration of the partial volume effect.

Then, κ can be estimated using Eq. (4). Now, that all unknowns are determined, Eq. (1) is solved to determine the texture map $S_{\mathrm{Tex}}(\vec{\mathbf{r}})$ (see Fig. 3).

2.1.4 *Experiments - comparison of simulation and measurement*

MR images of 10 healthy volunteers were acquired to compare the simulations with real measurements. The examinations were approved by the ethics committee of the Physikalisch-Technische Bundesanstalt and are in accordance with the relevant guidelines and regulations. Written informed consent was obtained from all volunteers prior to the measurements. Data were acquired at 3T (Siemens Verio) using the following sequences: a magnetization prepared rapid gradient echo for the estimation of the PV-maps (3D, TR = 2300 ms, TI = 900 ms, TE = 3.2 ms, voxel size: $0.75 \times 0.75 \times 4.69\ \mathrm{mm}^3$) and five T2w FLAIR scans as a reference measurement for the simulated images (Multislice 2D, TR = 9000 ms, voxel size: $0.75 \times 0.75 \times 4.69\ \mathrm{mm}^3$) with TE and TI values as given in Table 3 to represent the extreme shift derivatives of the possible scan domain and its center (see Fig. 5). The "center" protocol serves as the baseline scan for the simulations of the "corner" protocols.

Reference T1 values were obtained from saturation-recovery measurements. Eleven T1-weighted images for different saturation delay times (TD = 0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 1.0 1.25, 1.5, 2.0, and 8.0 s) were acquired using a fully sampled single-shot centric-reordered GRE readout (TE/TR = 3.0/6.5 ms, flip angle = 6 deg, voxel size: $1.3 \times 1.3 \times 8.0\ \mathrm{mm}^3$) implemented in pulseq.[30] Final quantitative T1 values were generated using a non-linear least squares

Table 3 TE and TI of the five T2w FLAIR acquisition protocols.

TE/ms	TI/ms
112	2500
84	2200
84	2900
150	2200
150	2900

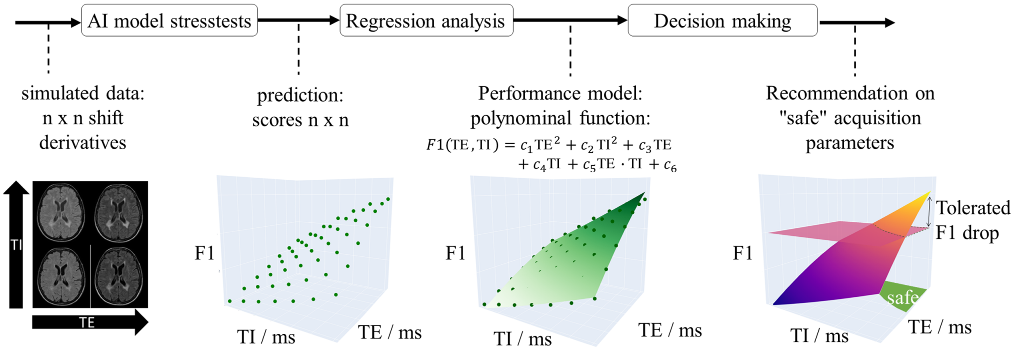
AI model stresstests
Regression analysis
Decision making
simulated data: n x n shift derivatives
prediction: scores n x n
Performance model: polynominal function:
$F1(\mathrm{TE},\mathrm{TI}) = c_1\mathrm{TE}^2 + c_2\mathrm{TI}^2 + c_3\mathrm{TE} + c_4\mathrm{TI} + c_5\mathrm{TE}\cdot\mathrm{TI} + c_6$
Recommendation on "safe" acquisition parameters
TI
TE
F1
TI / ms
TE / ms
F1
TI / ms
TE / ms
F1
Tolerated F1 drop
safe
TI / ms
TE / ms

Fig. 4 The AI model undergoes testing using generated images that represent varying acquisition shifts. Regression analysis delivers a model function for F1 to provide the user with an assessment of the model's limitations.

curve fitting algorithm[31] assuming a simple mono-exponential magnetization recovery. T2 reference values were derived from the two different TEs ($\mathrm{TE}_1 = 84$ ms and $\mathrm{TE}_2 = 150$ ms) of the FLAIR scans S_m using the following equation:

$$T_2 = \frac{\mathrm{TE}_2 - \mathrm{TE}_1}{ln\left(\frac{S_m(\mathrm{TE}_1)}{S_m(\mathrm{TE}_2)}\right)}. \tag{8}$$

The T2 estimates obtained with TI = 2900 ms and 2200 ms are averaged to deliver the final reference T2 values. The relaxometry estimates described in Sec. 2.1.3 are compared to these reference values and to values given by literature.[32–35] Finally, the five real and simulated scans are compared by the theoretical percentage signal deviation per ms relaxometry errors $\Delta T1$ and $\Delta T2$ approximated by error propagation as

$$ds_{\mathrm{T1}}(\mathrm{T1}) = \frac{\frac{ds_{\mathrm{FLAIR}}(T1,T2)}{dT1}}{s_{\mathrm{FLAIR}}}\cdot 100\%, \quad ds_{T2}(T2) = \frac{\frac{ds_{\mathrm{FLAIR}}(T1,T2)}{dT2}}{s_{\mathrm{FLAIR}}}\cdot 100\%, \tag{9}$$

and in dependence of T1 and T2 to confirm that signal differences are related to relaxometry imperfections. The stress test pipeline is summarized in Fig. 4 and comprises two steps as described in the following sections.

2.2 Model Stress Tests to Determine the Influence of Acquisition Shifts

2.2.1 *Generation of test data*

With the methods described in 2.1, derivatives of the baseline data can be generated that represent arbitrary acquisition shifts of a baseline scan ("shift derivatives"). Typical variations of scan protocols (minimum and maximum TE and TI values) were estimated using literature and real scans. The outcome of that investigation is published in Ref. 36 and is depicted in Fig. 5.

7 × 7 test datasets were generated that represent seven different TE values and seven different TI values, since these are the most contrast-affecting parameters in T2w FLAIR sequences.

2.2.2 *Modeling the network performance in dependence of sequence parameters*

The lesion F1 score of a lesion segmentation network can be determined for all of these data comparing the network prediction with the lesion ground truth segmentation masks. Averaging all lesion F1 scores finally delivers F1 as a function of TE and TI. We use a response surface method (quadratic model, cubic terms neglected) to describe the dependence of F1 on arbitrary values of the influencing factors TE and TI and their interactions as recommended by Ref. 13. Accordingly, the quadratic model in Eq. (10) is fitted to these F1 measurements

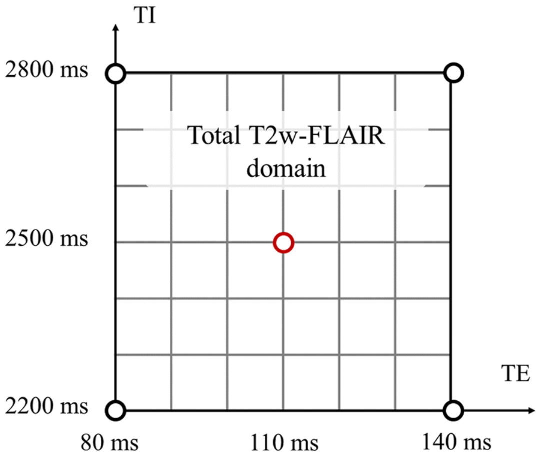

Fig. 5 Minimum and maximum values of TI and TE as determined by literature research and real scans. These values limit the real-world scan domain. Test data are generated by simulation to represent all possible data within this domain on a regular grid. The corners and the center (red circle) determine the MRI protocols for reference measurements used to validate the simulated data.

$$F1(\mathrm{TE},\mathrm{TI}) = c1 \cdot \mathrm{TE}^2 + \boldsymbol{c}2 \cdot \mathrm{TI}^2 + c3 \cdot (\mathrm{TI} \cdot \mathrm{TE})^2 + c4 \cdot \mathrm{TE} + c5 \cdot \mathrm{TI} + c6 \cdot \mathrm{TI} \cdot \mathrm{TE} + c7. \tag{10}$$

The coefficients $c1$ to $c6$ can each be understood as a measure of the relevance of the influencing factors TE and TI (main factors) and their interactions TE · TI.

2.2.3 *Experiments - stress testing SOTA models against acquisition shifts*

To validate the model function described in Eq. (10), two SOTA models are trained on data with heterogeneous contrast as described in Table 2. First, the nnU-Net framework is used, which utilizes a U-Net architecture and automatically configures its hyperparameters and configuration.[37] The first model is a 3D full-resolution nnU-Net, which is chosen by nnU-Net's auto-configured framework as the best-performing model among 2D and low-resolution 3D counterparts. Training is done by nnU-Net's self-configured automatic framework, where fivefold cross-validation is employed with 80% for training and 20% for validation, and the best-performing fold is chosen as the final model. The second model is a SegResNet model, which uses ResNet-like blocks and skip connections without the variational autoencoder part.[38] The network is trained with $64 \times 64 \times 64$ cropped blocks for 1000 epochs with an Adam optimizer and learning rate of 0.001 with Pytorch and MONAI tools. The training data are randomly split into fractions of 80% for training and 20% for validation.

The "longitudinal" OpenMS dataset is the only open benchmark dataset for which all contrast-affecting parameters (TE, TI, TR) are provided (Table 2). All data are skull stripped using the FSL brain extraction tool (FSL BET)[39] prior to all processing steps. The average F1 is determined and modeled as a function of TE and TI as described in 2.2.2. R^2 is used to evaluate the appropriateness of the model function in Eq. (10).

3 Results

3.1 Comparison of Simulation and Measurement

Figure 6 shows the variation of the estimated and reference relaxation measurements in comparison to the literature ranges. The estimated and measured relaxation times mostly lie within the literature range. As further underlined by the mean relaxometry values in Table 4, the high T1 value and the low FLAIR signal hampers relaxometry in CSF. The literature does not report on CSF T2 measurements at 3T. T2 is independent of the field strength but even at 1.5 T, to our knowledge, the Brainweb catalogue is the only literature source reporting a T2 value for CSF

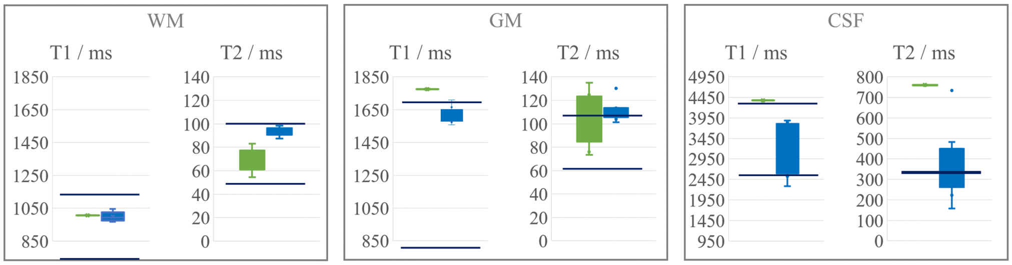

Fig. 6 Comparison of the ranges of estimated (green) and reference relaxation measurements (blue) at 3T. Blue lines show the literature ranges.[32–35]

Table 4 Mean values for T1 and T2 in normal tissue. All values are given in ms.

	$T1_{\mathrm{wm}}$	$T2_{\mathrm{wm}}$	$T1_{\mathrm{gm}}$	$T2_{\mathrm{gm}}$	$T1_{\mathrm{csf}}$	$T2_{\mathrm{csf}}$
Estimated	1007 ± 1	69 ± 9	1776 ± 3	102 ± 20	4376 ± 4	760 ± 379
Measured	999 ± 27	94 ± 3	1616 ± 46	111 ± 8	3176 ± 568	379 ± 157

(329 ms), although the values presented in that catalogue (in WM and GM) tend to be lower than most other values at 1.5 T.[40]

Visually, the images obtained by the simulations and measurements agree well (Fig. 7). Small scaling errors of the nulled CSF signal result in high relative signal deviations. In addition, Table 5 lists the relative error between real and simulated images in different manually drawn ROIs.

The deviation between the simulated and the measured MR signals in WM is higher than in GM. The theoretical error propagation of the relaxometry estimates on the simulated signal is depicted in Fig. 8.

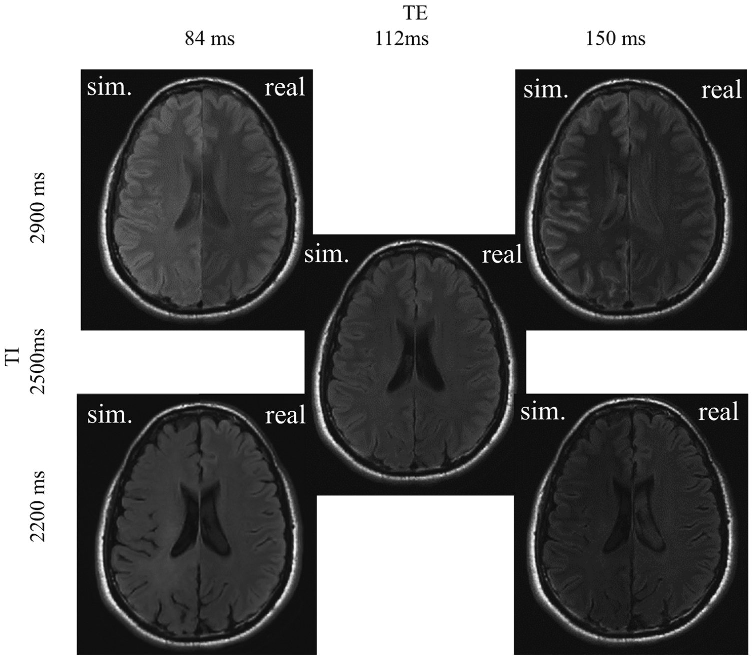

Fig. 7 MRI simulations (left side of the MRIs) and their real counterparts (right side of the MRIs) for all five protocols of one example volunteer. The simulation results are embedded in the skull segment to adjust the scaling of the images.

Table 5 Comparison of the mean signals of WM, GM, CSF, and skull of simulation and reference MRI with relative percentage error.

TE / ms	150	150	112	84	84
TI / ms	2900	2200	2500	2900	2200
WM / %	18	19	0	13	12
	±6	±6	0	±6	±6
GM / %	7	9	0	8	8
	±3	±7	±0	±6	±5
CSF / %	75	36	0	22	58
	±30	±9	±0	±13	±10

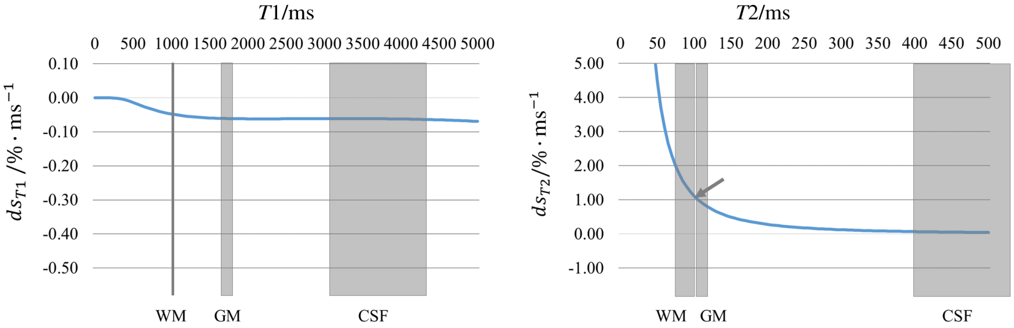

Fig. 8 Percentage signal simulation errors per ms relaxometry value as described by error propagation in dependence on the tissue's T1 or T2 values; e.g., (see arrows) the overestimation of T2 by 1 ms results in about 1% signal simulation error of WM and GM signals (here: given average protocol parameters). The absolute errors increase with T1 and decrease with T2.

3.2 Results of Stress Testing SOTA Models Against Acquisition Shift

Testing the models with the real baseline data and their simulated counterpart (TE = 140 ms and TI = 2800 ms) yields F1 scores, which differ in the fourth decimal place (OpenMS data: SegResNet: 0.4398 ± 0.2242; nnU–Net: 0.6105 ± 0.1500, see Fig. 9). The coefficient of determination R^2 of the model fit (second-order polynomial) is 0.991 for the SegResNet results and 0.982 for the nnU-Net results.

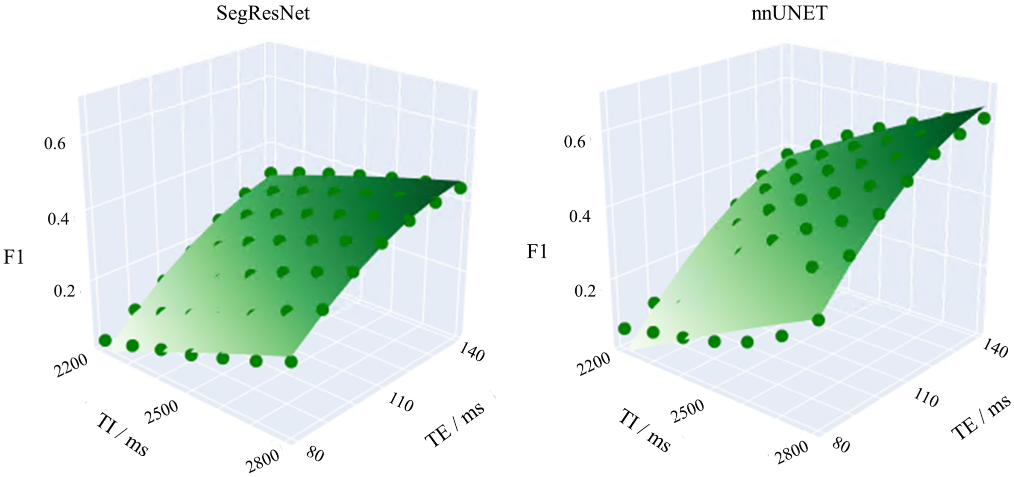

Fig. 9 The surface plots show the behavior of the AI models in dependence of the data shifts. Points: F1 scores of the predictions, surface: model fit, i.e., the F1 trend as a function of the acquisition parameters TE and TI.

Table 6 Coefficients $c1$ to $c7$ as given by the model fit (see Eq. 10). Units are given in ms^{-1} and ms^{-2} for linear, quadratic, and combined terms, respectively. The highest coefficients are those scaling the influencing factor TE.

	Intersection	TE	TI	TE^2	TI^2	$\text{TE} \cdot \text{TI}$
nnU-Net	−2.95	$2.24 \cdot 10^{-2}$	$9.42 \cdot 10^{-4}$	$-6.48 \cdot 10^{-5}$	$-7.59 \cdot 10^{-8}$	$-8.14 \cdot 10^{-7}$
SegResNet	−2.57	$2.14 \cdot 10^{-2}$	$8.25 \cdot 10^{-4}$	$-5.15 \cdot 10^{-5}$	$-5.37 \cdot 10^{-8}$	$-2.18 \cdot 10^{-6}$

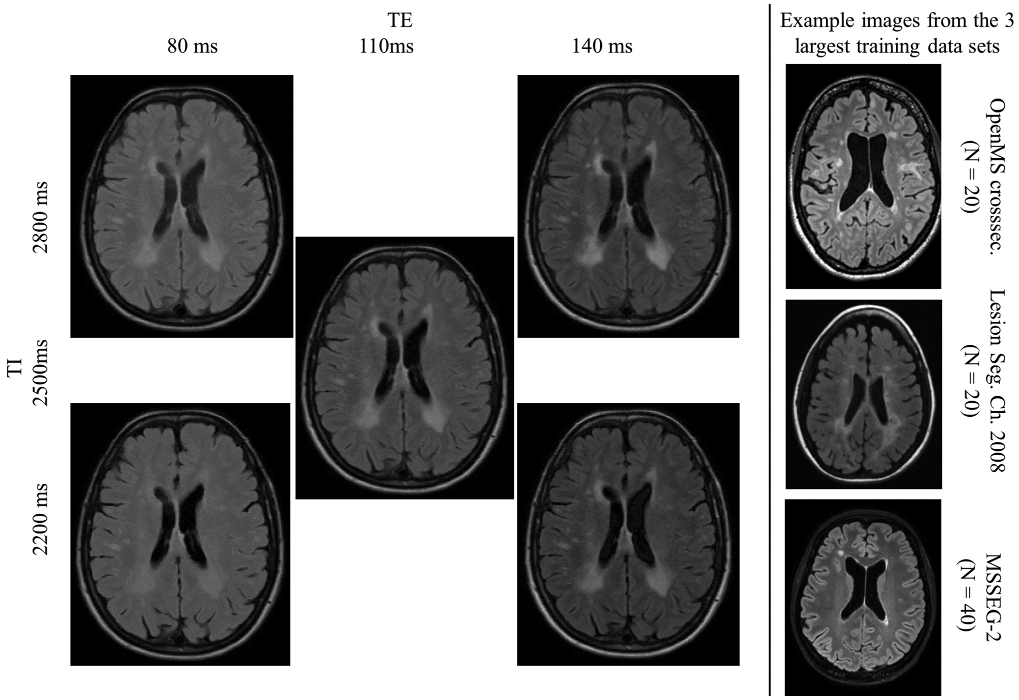
TE
80 ms
110ms
140 ms
TI
2800 ms
2500ms
2200 ms
Example images from the 3 largest training data sets
OpenMS crosssec. (N = 20)
Lesion Seg. Ch. 2008 (N = 20)
MSSEG-2 (N = 40)

Fig. 10 Left: MRI simulations for five protocols (Fig. 5) of one patient of the OpenMS dataset. The simulation results are embedded in the skull segment to adjust the scaling of the images. Right: Example images of the most relevant training datasets. The lesions differentiate well from the WM background, comparable to the shift derivatives with higher lesion-WM contrast (at high TE and TI).

The coefficients for Eq. (10) in Table 6 show that TE has the highest influence on both segmentation networks.

In the simulated images of Fig. 10, the lesion-to-WM contrast decreases for lower TE and TI values. This is accompanied by a decline of the F1-score, i.e., the models' ability to differentiate between the lesion and white matter decreases with lower contrast.

4 Discussion and Conclusion

The image generation method simulates acquisition shift derivatives of a real baseline scan for arbitrary sequence parameters. It was designed to be applicable to common clinical neuroimaging studies that normally contain T2w FLAIR and T1w images. It does not require extra sequences but only knowledge of the scan parameters of the baseline T2w FLAIR data.

4.1 Comparison of Simulation and Measurements

At the extreme points of the experimental design, the simulation shows a 19% deviation to the measured values in white matter and lower deviation in gray matter. This can most likely be explained by the inaccuracies of the relaxometry method used in this work. Using the error propagation as a rough guess, the misestimation of 19% could be explained by a 19 ms deviation of T2, which is likely to be realistic considering the reference measurements and the range of literature reference values. Even those reference relaxometry methods suffer from inaccuracies caused by inflow or sequence imperfections, in particular when estimating the T1 and T2 of flowing tissue like blood or CSF.[41] One could improve the validation by including T1 and T2 mapping sequences in the same resolution and spatial coverage. Common relaxometry sequences in neuroimaging rely on multiple 3D spoiled gradient recalled echo or inversion recovery sequences for T1 mapping and multi-echo or balanced steady-state free precession sequences at variable flip angles for T2 mapping.[41,42] The imaging study in this work was already time-consuming due to the five times repetition of the lengthy T2w FLAIR protocol and the T1 weighted scan. Therefore, there was just limited time for a rough dual echo T2 estimation and for the addition of a time-efficient single-slice T1-mapping protocol (acquisition time ∼30 s) to examine the T1 estimates in one slice, and thus values were compared ROI-wise. Still, the T1 and T2 values estimated here mostly lie in the range of literature values, and differences in the reference measurements are also comparable to the range of literature values. A one-to-one comparison of real and simulated images is challenging as it requires the exact knowledge of the relaxation times of that particular patient. Precise relaxometry is neither the aim of this work nor is it necessary for the simulation of test data. The relaxometry parameters in Eqs. (1) and (2) are set to arbitrary values to deliver a representative cohort of anatomies. Relaxometry imperfections hamper accurate validation of the simulated values, yet, they manifest only in a misestimation of the DICOM scaling factor κ and thus in under- or overestimation of the texture amplitude. Unfortunately, for MRI sequences this scaling factor is not part of the DICOM header as it is for the Hounsfield units in CT imaging. Irregularities of the texture amplitude, on the other hand, might be balanced by normalizing the texture amplitude over the entire dataset. Furthermore, the texture amplitude could be also included as another influencing factor in the stress test analysis in addition to the sequence parameters—e.g., as a measure of noise or artifact level. In contrast to using other AI-based generative approaches like GANs, VAEs, or diffusion models,[16,43–46] the underlying signal equation allows for the generation of arbitrary but distinct shift derivatives from just one dataset.

4.2 Stress Test Results

The stress test results between the two networks differ, either due to their architectures or different data splits used for training and validation. However, in both cases, the F1(TE, TI) measurements seem to be well described by the quadratic function. The metric varies only smoothly so that cubic terms can be neglected. TE seems to be the most influencing factor for all models, which is in line with the nature of the contrast weighting of the sequence (T2w FLAIR).

Furthermore, the lesion F1 values are comparable to that of real data (72%[47]) at least in or close to the baseline representation. The performance decreases towards the extreme points of the experimental grid (particularly for low TE values), where the lesion-WM contrast decreases. As one can see in Fig. 10 (example training images), the lesion-WM contrast of the training images was generally higher than in the low-TE simulations, which might explain the performance drop towards low TE values. In previous work, using fully simulated data, we showed that the maximum of the response surface plot and its shape are dependent on the contrast distribution of training and test data.[36] The stress test result can thus be a measure of model analysis and optimization. One has to bear in mind that these extreme points are mathematical constraints, given by the minimum and maximum combinations of TE and TI of real sequences. The boundary of the experimental grid does not represent the boundary of the typical scan domain. The latter does not necessarily contain the combination of extreme values of both TE and TI at the same time. Those extreme data simulations are thus not part of the training data therefore causing severe drops in the F1 value.

The high F1 scores for the two "high-TE corners" (Fig. 9) can also be explained by the high lesion contrast for these protocols. In contrast, the low lesion contrast yielded by low TE and TI values comes with low F1 scores, respectively. Another contribution of this work is thus a proof-of-concept for the description of the performance metric of an AI model in dependence of its influencing factors. The modeling yields a quantitative comparison of the relevance of all influencing factors. This concept of surface response modeling is based on well-established experimental designs and could be easily transferred to other common metrics[48] (e.g., confusion matrix and derivatives or even uncertainty estimates[49]) or other models (e.g., classification models). Now, that the model function was confirmed, the number of experiments could be reduced significantly in future studies to reduce the computational effort. For the optimal "positioning" of these sample points on the "domain grid" for meaningful sampling of the surface response curve, state-of-the-art guidelines in the field of experimental design offer several recommendations depending on the number of influencing factors.[13]

4.3 Limitations

One important limitation is the small number of test datasets used in this study. Thus, the absolute results of the stress tests might not be representative for a larger cohort of patients and lesions. They can only serve as a sample domain grid to confirm an appropriate model function and to demonstrate the proof of concept. Unfortunately, all open MS data are provided in NIfTI format and the OpenMS data are the only data that come at least with the information on TE, TI, and TR and thus all sequence parameters needed in the simulation. In real-world applications, one can assume that manufacturers of models have access to the entire DICOM header that also includes tags for TE, TI, TR, and many more. Thus, in theory, more acquisition shifts caused by other sequence parameters could be incorporated as influencing factors in the stress tests. However, since the number of sampling points on the domain grid quickly rises with every additional influencing factor, a prior prioritization is crucial.

An intrinsic limitation of the T2w FLAIR and T1w sequences is that the CSF signal is very low or even nulled hampering partial volume estimation and relaxometry in this tissue. Accordingly, the differences between the simulations and measurements become most apparent in CSF compared to the other tissues, limiting the validation of the approach in CSF. Future work should investigate if tissue and relaxometry estimation can be improved by additionally incorporating the contrast of conventional T2w sequences in the first step of the image generation pipeline, as in these images CSF shows up brightly. All three scans (T2w, T2w FLAIR, and the post Gd T1w scan) constitute the "recommended core" in current MS scanning guidelines.[10]

Another limitation is the assumption that the average texture contribution to the signal is zero. This is not true in the case of artifacts resulting from inhomogeneities of B0, B1, or the receive coil sensitivity profile.[50,51] The method is further only applicable to baseline images, of which the contrast can be fully described by the parameters accessible in the DICOM header; e.g., the parameter $\mathrm{TE}_{\mathrm{last}}$ in Eq. (2) is approximated by $2 \cdot \mathrm{TE}$, since it is not part of the DICOM header. In the real volunteer scans, the true value for $\mathrm{TE}_{\mathrm{last}}$ was 30% higher. In these experiments, changing the parameter to the correct value did not have any influence on the outcome of the comparison (due to the long TR value). Still, there might be other measures of contrast manipulation in T2w FLAIR studies that are not accessible by the DICOM tags and that prevent an accurate estimation of the DICOM scaling factor and thus the texture amplitude (e.g., modulated RF pulses to prevent the signal from decaying in long echo trains, acceleration techniques and dedicated *k*-space ordering, particularly common in 3D sequences,[25,52–55] blood inflow,[56] etc). Future work should elaborate to what extent these influences and their impact can be modeled and incorporated either in the simulation, e.g., by random guesses or in the stress tests represented by additional influencing factors.

Despite these limitations, the image simulation and stress test methodology presented in this work allows for investigation of the robustness of AI models in response to arbitrary data shifts. Due to the lack of a gold standard, the metrological proof of the F1 response to parameter changes is not possible and absolute predictions about these values remain uncertain. However, influencing parameters in the MR sequence can be compared with each other by the surface model coefficients and—given a tolerated performance drop—"safe" parameters settings can be at least roughly assessed (Fig. 4). Using the simulation algorithm as an alternative augmentation method also allows for introducing *a priori* knowledge on MR signal variations into the AI-model development process.

Disclosures

The co-authors Mehmet Yigit Avci und Mehmet Yigitsoy are employed by the deepc GmbH, Munich.

Code and Data Availability

The MS data utilized in this study are listed Table 2. The data policy of the clinical study does not allow free access to the volunteer MRI data. Due to the collaboration agreement with the industrial partner, the code cannot be made available.

Acknowledgments

This project was funded by the Zentrales Innovationsprogramm Mittelstand (ZIM) of the German Federal Ministry for Economic Affairs and Climate Action (BMWK) (Grant No. KK5050201 LB0). This work has further been partly supported by Collaborative Research Centre “Matrix in Vision” funded by German Research Foundation (DFG) (Grant No. CRC-1340).

References

1. “ACR List of FDA cleared AI medical products,” AI Central, ACR Data Science Institute, American College of Radiology, 2022, https://aicentral.acrdsi.org/ (accessed 21 Feb. 2022).
2. D. C. Castro, I. Walker, and B. Glocker, “Causality matters in medical imaging,” *Nat. Commun.* **11**(1), 3673 (2020).
3. E. Stamoulou et al., “Harmonization strategies in multicenter MRI-based radiomics,” *J. Imaging* **8**(11), 303 (2022).
4. H. Guan and M. Liu, “Domain adaptation for medical image analysis: a survey,” *IEEE Trans. Biomed. Eng.* **69**(3), 1173–1185 (2022).
5. Focus group on Artificial Intelligence for Health of the WHO, “Whitepaper for the ITU/WHO Focus Group on Artificial Intelligence for Health,” https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/FG-AI4H_Whitepaper.pdf (accessed 4 April 2024).
6. S. Reddy et al., “Evaluation framework to guide implementation of AI systems into healthcare settings,” *BMJ Health Care Inf.* **28**(1), e100444 (2021).
7. L. Oala et al., “Machine learning for health: algorithm auditing & quality control,” *J. Med. Syst.* **45**(12), 105 (2021).
8. P. Omoumi et al., “To buy or not to buy—evaluating commercial AI solutions in radiology (the ECLAIR guidelines),” *Eur. Radiol.* **31**(6), 3786–3796 (2021).
9. J. Denck et al., “Automated protocoling for MRI exams—challenges and solutions,” *J. Digital Imaging* **35**(5), 1293–1302 (2022).
10. M. P. Wattjes et al., “2021 MAGNIMS–CMSC–NAIMS consensus recommendations on the use of MRI in patients with multiple sclerosis,” *Lancet Neurol.* **20**(8), 653–670 (2021).
11. P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, “Explainable AI: a review of machine learning interpretability methods,” *Entropy* **23**(1), 18 (2020).
12. B. McCrindle et al., “A radiology-focused review of predictive uncertainty for AI interpretability in computer-assisted segmentation,” *Radiol. Artif. Intell.* **3**(6), e210031 (2021).
13. W. F. Guthrie, *NIST/SEMATECH e-Handbook of Statistical Methods (NIST Handbook 151)*, National Institute of Standards and Technology (2020).
14. L. Arras, A. Osman, and W. Samek, “CLEVR-XAI: a benchmark dataset for the ground truth evaluation of neural network explanations,” *Inf. Fusion* **81**, 14–40 (2022).
15. Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institute (HHI), “Syreal-Synthesizing realistic variations in data for reliable medical machine learning at scale,” https://www.hhi.fraunhofer.de/en/departments/ai/projects/syreal.html (accessed 9 Feb. 2022).
16. A. F. Frangi, S. A. Tsaftaris, and J. L. Prince, “Simulation and synthesis in medical imaging,” *IEEE Trans. Med. Imaging* **37**(3), 673–679 (2018).
17. R. K.-S. Kwan, A. C. Evans, and G. B. Pike, “MRISIM: Tissue MR parameters,” https://brainweb.bic.mni.mcgill.ca/brainweb/tissue_mr_parameters.txt (accessed 9 April 2024).
18. L. A. Shepp and B. F. Logan, “The Fourier reconstruction of a head section,” *IEEE Trans. Nucl. Sci.* **21**(3), 21–43 (1974).
19. Ž. Lesjak et al., “Validation of white-matter lesion change detection methods on a novel publicly available MRI image database,” *Neuroinformatics* **14**(4), 403–420 (2016).
20. Ž. Lesjak et al., “A novel public MR image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus,” *Neuroinformatics* **16**(1), 51–63 (2018).
21. A. Carass et al., “Longitudinal multiple sclerosis lesion segmentation: resource and challenge,” *NeuroImage* **148**, 77–102 (2017).
22. M. Styner et al., “3D segmentation in the clinic: a grand challenge II: MS lesion segmentation,” *MIDAS J.* (2008).
23. O. Commowick et al., “MSSEG-2 challenge proceedings: multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure,” presented at *MICCAI 2021 - 24th Int. Conf. Med. Image Comput. and Computer Assist. Interv.*, p. 126 (2021).

24. E. Roura et al., “Automated detection of lupus white matter lesions in MRI,” *Front. Neuroinf.* **10**, 33 (2016).
25. J. N. Rydberg et al., “Contrast optimization of fluid-attenuated inversion recovery (FLAIR) imaging,” *Magn. Reson. Med.* **34**(6), 868–877 (1995).
26. J. Tohka, A. Zijdenbos, and A. Evans, “Fast and robust parameter estimation for statistical partial volume models in brain MRI,” *Neuroimage* **23**(1), 84–97 (2004).
27. J. Linn, M. Wiesmann, and H. Brückmann, “Infektiöse und entzündlich-demyelinisierende Erkrankungen,” in *Atlas Klinische Neuroradiologie des Gehirns*, J. Linn, M. Wiesmann, and H. Brückmann, Eds., pp. 311–384, Springer, Berlin, Heidelberg (2011).
28. B. Billot et al., “SynthSeg: segmentation of brain MRI scans of any contrast and resolution without retraining,” *Med. Image Anal.* **86**, 102789 (2023).
29. B. Billot et al., “Partial volume segmentation of brain MRI scans of any resolution and contrast,” *Lect. Notes Comput. Sci.* **12267**, 177–187 (2020).
30. K. J. Layton et al., “Pulseq: a rapid and hardware-independent pulse sequence prototyping framework,” *Magn. Reson. Med.* **77**(4), 1544–1552 (2017).
31. M. Newville et al., lmfit/lmfit-py: 1.2.1, Zenodo (2023).
32. R. E. Gabr et al., “Patient-specific 3D FLAIR for enhanced visualization of brain white matter lesions in multiple sclerosis,” *J. Magn. Reson. Imaging* **46**(2), 557–564 (2017).
33. A. Parry et al., “White matter and lesion T1 relaxation times increase in parallel and correlate with disability in multiple sclerosis,” *J. Neurol.* **249**(9), 1279–1286 (2002).
34. G. J. Stanisz et al., “T1, T2 relaxation and magnetization transfer in tissue at 3T,” *Magn. Reson. Med.* **54**(3), 507–512 (2005).
35. A. Parry et al., “MRI Brain T1 relaxation time changes in MS patients increase over time in both the white matter and the cortex,” *J. Neuroimaging* **13**(3), 234–239 (2003).
36. C. Posselt et al., “Novel concept for systematic testing of AI models for MRI acquisition shifts with simulated data,” *Proc. SPIE* **12467**, 124671B (2023).
37. F. Isensee et al., “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation,” *Nat. Methods* **18**(2), 203–211 (2021).
38. A. Myronenko, “3D MRI brain tumor segmentation using autoencoder regularization,” *Lecture Notes in Comput. Sci.* **11384**, 311–320 (2018).
39. S. M. Smith, “Fast robust automated brain extraction,” *Hum. Brain Mapp.* **17**(3), 143–155 (2002).
40. McConnell Brain Imaging Centre, “MRISIM: tissue MR parameters,” 1996, https://brainweb.bic.mni.mcgill.ca/brainweb/tissue_mr_parameters.txt (accessed 10 Aug. 2023).
41. S. C. L. Deoni, “Quantitative relaxometry of the brain,” *Top. Magn. Reson. Imaging* **21**(2), 101–113 (2010).
42. M. Tranfa et al., “Quantitative MRI in multiple sclerosis: from theory to application,” *Am. J. Neuroradiol.* **43**(12), 1688–1695 (2022).
43. S. Kazeminia et al., “GANs for medical image analysis,” *Artif. Intell. Med.* **109**, 101938 (2020).
44. T. Wang et al., “A review on medical imaging synthesis using deep learning and its clinical applications,” *J. Appl. Clin. Med. Phys.* **22**(1), 11–36 (2021).
45. G. Müller-Franzes et al., “A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis,” *Sci. Rep.* **13**, 12098 (2023).
46. L. X. Nguyen et al., “A new chapter for medical image generation: the stable diffusion method,” in *Int. Conf. Inf. Networking (ICOIN)*, pp. 483–486 (2023).
47. P. Schmidt et al., “Automated segmentation of changes in FLAIR-hyperintense white matter lesions in multiple sclerosis on serial magnetic resonance imaging,” *Neuroimage Clin.* **23**, 101849 (2019).
48. L. Maier-Hein et al., “Metrics reloaded: recommendations for image analysis validation,” *Nat. Methods* **21**, 195–212 (2024).
49. T. Nair et al., “Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation,” *Med. Image Anal.* **59**, 101557 (2020).
50. U. Vovk, F. Pernus, and B. Likar, “A review of methods for correction of intensity inhomogeneity in MRI,” *IEEE Trans. Med. Imaging* **26**(3), 405–421 (2007).
51. O. Dietrich, M. F. Reiser, and S. O. Schoenberg, “Artifacts in 3-T MRI: physical background and reduction strategies,” *Eur. J. Radiol.* **65**(1), 29–35 (2008).
52. S. J. P. Meara and G. J. Barker, “Evolution of the longitudinal magnetization for pulse sequences using a fast spin-echo readout: application to fluid-attenuated inversion-recovery and double inversion-recovery sequences,” *Magn. Reson. Med.* **54**(1), 241–245 (2005).
53. R. F. Busse et al., “Fast spin echo sequences with very long echo trains: design of variable refocusing flip angle schedules and generation of clinicalT2 contrast,” *Magn. Reson. Med.* **55**(5), 1030–1037 (2006).
54. R. F. Busse et al., “Effects of refocusing flip angle modulation and view ordering in 3D fast spin echo,” *Magn. Reson. Med.* **60**(3), 640–649 (2008).
55. J. P. Mugler, III, “Optimized three-dimensional fast-spin-echo MRI,” *J. Magn. Reson. Imaging* **39**(4), 745–767 (2014).

56. S. Naganawa et al., "Comparison of flow artifacts between 2D-FLAIR and 3D-FLAIR sequences at 3 T," *Eur. Radiol.* **14**(10), 1901–1908 (2004).

Christiane Posselt worked as a research assistant on the NeuroTEST project at the University of Applied Sciences in Landshut. Her main focus was the exploration of the simulation and stress test methods used in this work. She holds a master's degree in electrical engineering from the University of Applied Sciences in Landshut, Germany.

Mehmet Yigit Avci received his bachelor's degree in electrical and electronics engineering from Bogazici University, Istanbul. He is currently a master's student at Technical University of Munich with a specialization in biomedical computing. His research interests are medical imaging and machine learning.

Mehmet Yigitsoy: Biography is not available.

Patrick Schuenke studied physics at the University of Heidelberg in Germany and completed his PhD in physics at the German Cancer Research Center (DKFZ) in 2017. Afterwards, he worked as a postdoctoral researcher at the Leibniz-Forschungsinstitut für Molekulare Pharmakologie in Berlin. In 2020, he transitioned to the Physikalisch-Technische Bundesanstalt, where he currently focuses on advancing quantitative MRI techniques and developing open-source MRI software.

Christoph Kolbitsch is head of the research group "Quantitative MRI" at the Physikalisch-Technische Bundesanstalt in Berlin, Germany. He received his PhD from King's College London working on motion compensation for high-resolution cardiac MRI. His group is mainly working on advanced image reconstruction techniques combining detailed physical models of the imaging process with the flexibility of deep learning. He is also an advocate for open-source image reconstruction software: https://github.com/PTB-MR/mrpro.

Tobias Schaeffter is the head of division of Medical Physics and metrological IT at the Physikalisch-Technische Bundesanstalt in Berlin, Germany. He is a professor in Biomedical Imaging at TU-Berlin and the Einstein Centre Digital Future. He studied electrical engineering at TU-Berlin and did his PhD at University Bremen. From 1996 to 2006, he worked as a principal scientist at the Philips Research Laboratories in Hamburg, Germany. From 2006 to 2015, he was professor of imaging sciences at King's College London.

Stefanie Remmele is a professor of medical technologies at the University of Applied Sciences in Landshut, Germany since 2012. Prior to that, she conducted research on quantitative MR methods at Philips Research in Hamburg, Germany. Her current research interests include image simulation and synthesis in radiology and image-guided therapy, and she is the head of the Research Group of Medical Technologies at the University in Landshut.