Can deep learning replace gadolinium in neuro-oncology?

Samy Ammari, Alexandre Bône, Corinne Balleyguier, Eric Moulton, Émilie Chouzenoux, Andreas Volk, Yves Menu, François Bidault, François Nicolas, Philippe Robert, et al.

Title

Can deep learning replace gadolinium in neuro-oncology? A reader study

Authors

- Samy Ammari[1, 2, *], MD, MSc
- Alexandre Bône[3, *], PhD
- Corinne Balleyguier[1, 2], MD, PhD
- Eric Moulton[3], PhD
- Émilie Chouzenoux[4], PhD
- Andreas Volk[1, 2], PhD
- Yves Menu[1], MD, PhD
- François Bidault[1], MD, PhD
- François Nicolas[3], PhD
- Philippe Robert[3], PhD
- Marc-Michel Rohé[3], PhD
- Nathalie Lassau[1, 2], MD, PhD

(: equal contributions)*

Institutions

1. Imaging Department, Gustave Roussy Cancer Campus, Université Paris-Saclay, Villejuif, France
2. BioMaps (UMR1281), Université Paris-Saclay, CNRS, Inserm, CEA, Villejuif, France
3. Guerbet Research, Villepinte, France
4. Center for Visual Computing, CentraleSupélec, Inria, Université Paris-Saclay, Gif-sur-Yvette, France

Funding information

Research study funded by Guerbet. Among its portfolio of solutions for medical imaging, Guerbet commercializes Dotarem (gadoterate meglumine), a gadolinium-based contrast agent.

Abbreviations

- GBCA = gadolinium-based contrast agent,
- ADC = apparent diffusion coefficient,
- low-T1c = low-dose contrast-enhanced T1-weighted image (0.025mmol/kg injection of GBCA),
- ref-T1c = reference contrast-enhanced T1-weighted image (0.1mmol/kg in two successive injections of GBCA),
- vir-T1c = virtual contrast-enhanced T1-weighted image,
- SE = sensitivity,
- SP = specificity,
- FDR = false detection rate,
- PPV = positive predictive value,
- F1 = F1-score.

Abstract

Objectives

This study proposes and evaluates a deep learning method that predicts surrogate images for contrast-enhanced T1 from multiparametric MRI acquired using only a quarter of the standard 0.1mmol/kg dose of gadolinium-based contrast agent. In particular, the predicted images are quantitatively evaluated in terms of lesion detection performance.

Materials and methods

This monocentric retrospective study leveraged 200 multiparametric brain MRIs acquired between November 2019 and February 2020 at Gustave Roussy Cancer Campus (Villejuif, France). A total of 145 patients were included: 107 formed the training sample (age 55y±14, 58 women) and 38 the separate test sample (age 62±12, 22 women). Patients had glioma, brain metastases, meningioma, or no enhancing lesion. T1, T2-Flair, DWI, low-dose and standard-dose postcontrast T1 sequences were acquired. A deep network was trained to process the precontrast and low-dose sequences in order to predict "virtual" surrogate images for contrast-enhanced T1. Once trained, the deep learning method was evaluated on the test sample. The discrepancies between the predicted virtual images and the standard-dose MRIs were qualitatively and quantitatively evaluated using both automated voxel-wise metrics and a reader study, where two radiologists graded image qualities and marked all visible enhancing lesions.

Results

The automated analysis of the test brain MRIs computed a structural similarity index of 87.1% (±4.8) between the predicted virtual sequences and the reference contrast-enhanced T1 MRIs, a peak signal-to-noise ratio of 31.6dB (±2.0), and an area under the curve of 96.4% (±3.1). At Youden's operating point, the voxel-wise sensitivity and specificity were 96.4% and 94.8% respectively. The reader study found that virtual images were preferred to standard-dose MRI in terms of image quality (p=.008). A total of 91 reference lesions were identified in the 38 test T1 sequences enhanced with full dose of contrast agent. On average across readers, the brain lesion sensitivity of the virtual images was 83% for lesions larger than 10mm (n=42), and the associated false detection rate was 0.08 lesion/patient. The corresponding positive predictive value of detected lesions was 92%, and the F1-score 88%. Lesion detection performance however dropped when smaller lesions were included: average sensitivity was 67% for lesions larger than 5mm (n=74), and 56% with all lesions included regardless of their size. The false detection rate remained below 0.50 lesion/patient in all cases, and the positive predictive value above 73%. The composite F1 score was 63% at worst.

Conclusion

The proposed deep learning method for virtual contrast-enhanced T1 brain MRI prediction showed very high quantitative performance when evaluated with standard voxel-wise metrics. The reader study demonstrated that for lesions larger than 10mm, good detection performance could be maintained despite a 4-fold division in contrast agent usage, unveiling a promising avenue for reducing the gadolinium exposure of returning patients. Small lesions proved however difficult to handle for the deep network, showing that full-dose injections remain essential for accurate first-line diagnosis in neuro-oncology.

Key words

Gadolinium, contrast agents, multiparametric MRI, low-dose, deep learning, image prediction, lesion detection, neuro-oncology, reader study.

Introduction

It is estimated that half of central nervous system MRI examinations are performed with intravenous gadolinium-based contrast agents (GBCAs) in order to detect, characterize, or monitor brain tumors with optimal accuracy (1). For more than 30 years, GBCAs have been widely considered to be among the safest drugs ever introduced (2). However, since the identification of the role of gadolinium in nephrogenic systemic fibrosis (3,4) and the observation of its deposition in the brain (5), numerous studies have evaluated the cumulative effects of GBCA injections (6–10). Linear GBCAs have been shown to accumulate in the nucleus dentatus and globus pallidus, unlike macrocyclic agents which are more stable chemically (11,12). Despite the lack of evidence that gadolinium retention causes disease or disorders in subjects with normal kidney function, researchers are still investigating multiple leads (13–15) and a precautionary principle led the EMA to restrict the usage of linear agents in the European market since 2017 (16). In the United States, the FDA has urged clinicians to minimize GBCA usage whenever feasible (17).

To meet this ambition without degrading diagnostic accuracy, three approaches are developed by the imaging, chemistry, and artificial intelligence research communities. First, contrast sequences may be entirely replaced in certain cases by zero-contrast sequences (18) such as amide proton transfer imaging (19–21), magnetic resonance spectroscopy (22), arterial spin labeling (23), or intravoxel incoherent motion imaging (24–27). Second, alternative contrast agents may either compensate for lower gadolinium dosage thanks to their higher relaxivity (28–30) or eliminate completely gadolinium from their chemical composition by using manganese (31,32) or iron (33,34) instead. Third, deep learning algorithms may automatically enhance the level of contrast of low-dose MRI acquisitions (35) or predict virtual contrast-enhanced images by analyzing multiple zero-contrast sequences (36), possibly completed by a low-dose acquisition (37).

While promising, these artificial intelligence methods have only been validated using either simple voxel-wise metrics (e.g. with the structural similarity index (38)) or subjective quality assessment grades (e.g. using 5-point Likert scales). It is therefore difficult to anticipate their true clinical impact as surrogates for standard contrast-enhanced imaging, and in turn their potential for reducing GBCA usage in neuro-oncology. Our aim was to address this limitation by quantitatively evaluating the lesion detection performance, understood notably in terms of sensitivity and false detection rate, of virtual contrast-enhanced images predicted from a combination of zero-dose and quarter-dose MRI sequences by a state-of-the-art deep learning method, taking full-dose MRI as the reference standard.

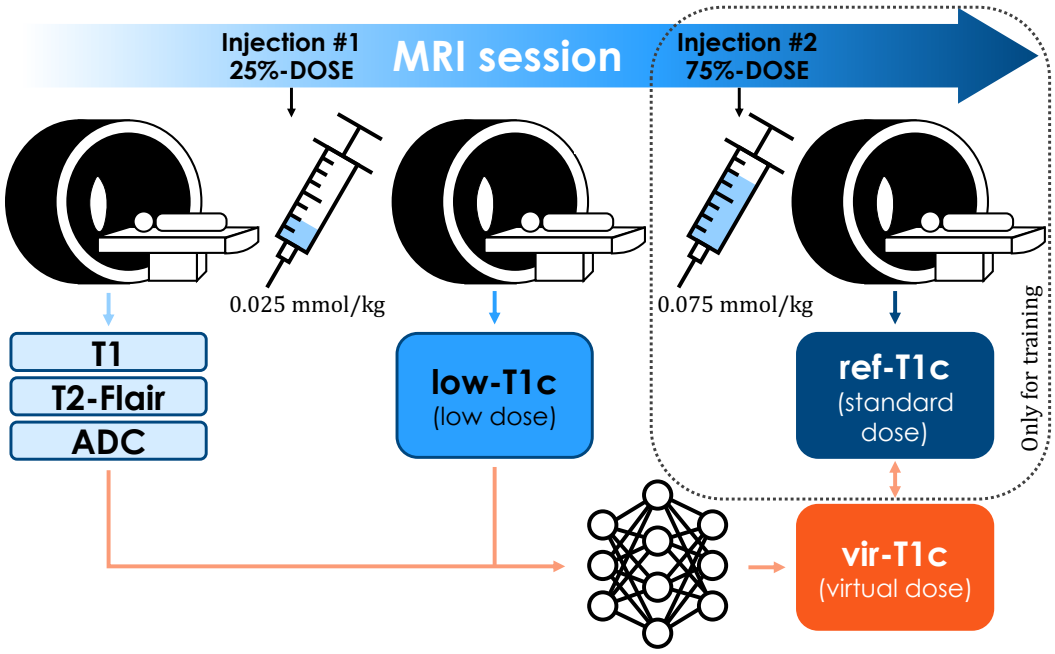

Figure 1: Proposed image prediction method based on deep learning, quantitatively evaluated in this study in terms of lesion detection performance. The deep learning algorithm is trained to predict, from T1, T2-Flair, apparent diffusion coefficient (ADC) and 25%-dose contrast-enhanced T1 sequences, surrogate images for T1 MRI enhanced with a standard 0.1mmol/kg dose of gadolinium-based contrast agent, administered in two successive injections. Low-dose, virtual and reference images are respectively abbreviated low-T1c, vir-T1c and ref-T1c.

Materials and methods

Data acquisition

Study sample

Imaging data was collected from 200 consecutive MRI exams performed at Gustave Roussy Cancer Campus (Villejuif, France) between November 2019 and February 2020 to explore for primary brain tumors or brain metastases. Analysis was retrospective, and approval was granted by the institutional review board (registration number 2021-18), in accordance with GDPR provisions. Prior to study initiation, the 200 MRI exams were split into training and test sets with a ratio 3:1, stratified with respect to acquisition dates, and ensuring that no patient could belong to both sets. In order to prevent any form of data leakage, the test set was strictly held out until all pre-processing and deep learning hyperparameters were chosen. Out of the 200 collected exams, 145 corresponding to as many different patients were finally included: the flow diagram (Figure 1) details the exclusion mechanisms. In the training and test samples confounded, the mean age was 57 (standard deviation 14), and 55% (80/145) were women. In terms of underlying conditions, 21% (30/145) had glioma, 46% (66/145) had brain metastases, 1% (1/145) had meningioma, and 33% (48/145) did not present any enhancing lesion. Table 1 details how these overall demographics and baseline characteristics are distributed between the training and test samples.

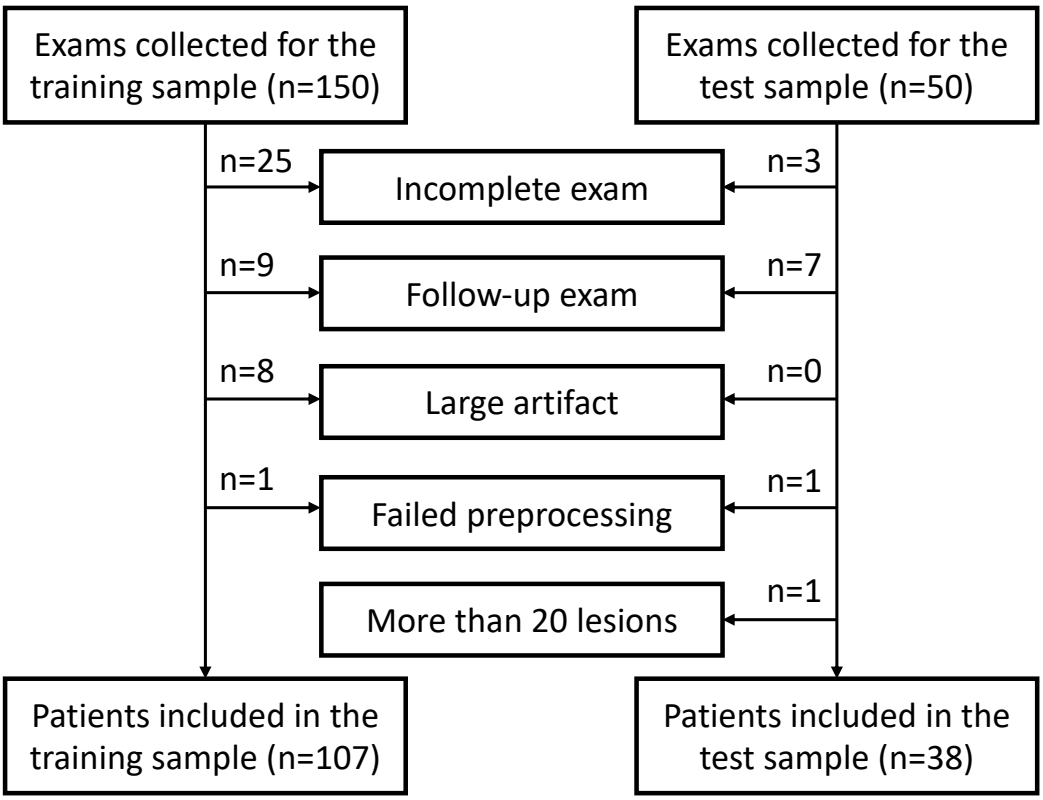

Figure 2: Exclusion flowchart for the training and test samples construction.

Variable	Training sample	Test sample	*p-value*
Nb. of included patients	107	38	
Women	54% (58/107)	58% (22/38)	*p=0.71*
Age	55y (±14)	62y (±12)	*p=0.006*
Weight	70kg (±18)	71kg (±18)	*p=0.31*
Indication			*p=0.51*
Glioma	21% (22/107)	21% (8/38)	
Brain metastases	49% (52/107)	37% (14/38)	
Meningioma	1% (1/107)	0% (0/38)	
No enhancing lesion	29% (32/107)	42% (16/38)	

Table 1: Demographics and baseline characteristics for included patients in training and test samples. Statistical significance of differences between training and test samples was estimated using the Fisher exact test for gender and indication, and the Wilcoxon-Mann-Whitney test for age and weight.

MRI Protocol

Two imaging machines from the same manufacturer (General Electric, Milwaukee, USA) were used: Optima MR450w 1.5T and Discovery MR750w 3T. Table 2 details the MRI parameters for both machines. T1, T2-FLAIR, and DWI sequences were acquired first. A 0.025mmol/kg dose of gadoterate meglumine (Dotarem, Guerbet, Villepinte, France) was then injected, and a T1-weighted sequence was acquired. A second 0.075mmol/kg dose of the same GBCA was injected, and another T1 sequence with identical parameters was acquired. These two images will be called low-dose and reference contrast-enhanced T1 in the rest of the article, and abbreviated low-T1c and ref-T1c. The median delay between the acquisition of low-T1c and ref-T1c sequences was 6 minutes and 5 seconds (inter-quartile

range 2 minutes and 10 seconds). Note that although our two-injection protocol cannot be considered as strictly equivalent to a single-injection one, the contrast uptake dynamics typically feature a fast wash-in followed by a slow wash-out (39), suggesting that similar enhancement patterns would be produced in practice. Note also that two-injection MRI protocols are already recommended as a possible standard of care when a perfusion sequence is included, the first injection playing the role of preload bolus (40).

Machine	Weighting	Sequence	TR	TE	Slice thickness
Optima MR450w 1.5T Installed in 2016, 70cm tunnel, 32 channels, 50cm z-axis FOV, gradients 40mT/m SR 200T/m/s.	T1 pre-contrast	3D rapid gradient echo	9ms	4.2ms	1mm
	T2-Flair	Turbo spin echo	7002ms	138ms	1.4mm
	DWI	EPI, two b-values (0 and 1000 mm^2/s)	3349ms	79ms	4mm
	T1 post-contrast	3D rapid gradient echo	6.1ms	1.2ms	1mm
Discovery MR750w 3T Installed in 2012, 70cm tunnel, 32 channels, 50cm z-axis FOV, gradients 44mT/m SR 200T/m/s.	T1 pre-contrast	3D rapid gradient echo	5.9ms	2.1ms	1mm
	T2-Flair	Turbo spin echo	7002ms	118ms	1mm
	DWI	EPI, two b-values (0 and 1000 mm^2/s)	5375ms	62.6ms	3mm
	T1 post-contrast	3D rapid gradient echo	6.1ms	2.1ms	1mm

Table 2: MRI parameters.

Deep learning method

Preprocessing

Only one MRI exam was included per patient: if several visits were conducted, the latest ones were excluded. Incomplete exams with at least one missing sequence among DWI, T2-Flair, T1, low-T1c, or ref-T1c were also excluded. Apparent diffusion coefficient (ADC) maps were automatically created. Imaging sequences were co-registered on the ICBM 2009a nonlinear symmetric brain atlas[1] using the FSL FLIRT[2] v6.0 pipeline with 12 degrees of freedom, resampled to a 1mm isotropic resolution using spline interpolation, skull-stripped with HD-BET[3], and cropped to a common image 160x192x160 size. The ADC, T2-Flair, T1, ref-T1c and low-T1c signals were standardized and linearly mapped to the [0, 1] interval after 1-percentile clipping of extreme values. A visual check was finally performed.

Deep network architecture

Figure 3 presents our deep network architecture, a full-resolution three-dimensional UNet proposed in (37) and originally adapted from (41). A contractive path encodes the input multi-modal MRI volumes into a hierarchy of features maps by alternating convolution layers with or without striding

[1] See: https://www.bic.mni.mcgill.ca/ServicesAtlases/ICBM152NLin2009 [Accessed 14th April 2021].
[2] See: https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FLIRT [Accessed 14th April 2021].
[3] See: https://github.com/MIC-DKFZ/HD-BET [Accessed 14th April 2021].

and kernel sizes 2 or 3 respectively. A symmetrical expansive path decodes the computed representation by successively applying strided transposed convolutions with kernel size 2 and standard convolutions with kernel size 3. Both short additive and long concatenative skip connections are exploited in order to facilitate gradient flows across the network. Apart from the final sigmoid, all activation functions are 0.2-LeakyReLU.

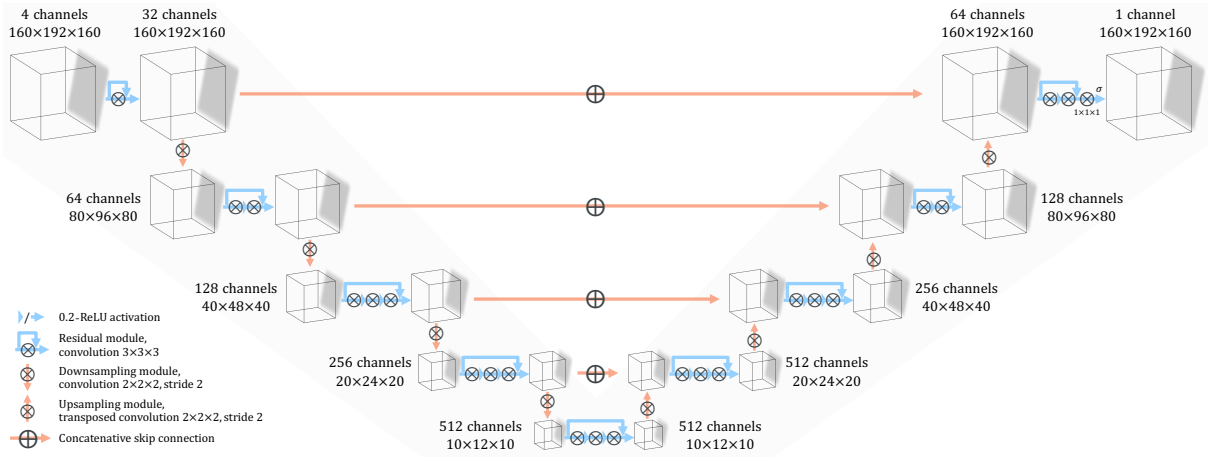

Figure 3: Architecture of the deep network, trained to predict surrogate images for contrast-enhanced T1 MRI from T1, T2-Flair, apparent diffusion coefficient (ADC) and 25%-dose contrast-enhanced T1 sequences.

Training and evaluating the model

The training sample of 107 patients was randomly partitioned into five subsets containing 21 or 22 patients each, and the deep network was repeatedly trained to process the ADC, T2-Flair, T1 and low-T1c sequences in order to predict virtual contrast-enhanced T1 images (vir-T1c), taking the ref-T1c images as ground truth. At each of the five repetitions, a particular training subset was selected as tuning set and used to select the final model weights, calibrated on the remaining 80% of the data using the L2 loss function and the Adam optimizer with default hyperparameters. Across the 300 training epochs, images were flipped with 50% probability around each axis for data augmentation purposes. Mirroring was also used to process the test MRIs, where all eight possible configurations were exhaustively computed for each of the five models in order to provide test-time augmentation. Simple Euclidian averaging was finally used to combine the forty resulting predictions. This test-time augmentation scheme allowed the gain of 2 percentage points on average in terms of the structural similarity metric, evaluated on the training set in a 5-fold cross-validation approach.

Training and evaluation tasks were performed using a single Azure NC6s workstation, equipped with a Nvidia V100 GPU. Half-precision encoding was exploited in order to minimize the GPU memory footprint and accelerate computations. Training the deep learning model took approximately 14 hours and 30 minutes for each of the five folds. Predicting the 38 test vir-T1c images took 10.2 minutes in total, which corresponds to a mean prediction time of 16 seconds per exam. *A link to a public repository containing our Python code will be provided here upon acceptance of our manuscript.*

Evaluation

Automated analysis using voxel-wise metrics

For all the 38 patients in the test sample, the structural similarity index (38), peak signal-to-noise ratio, and area under the receiver operating characteristic curve were computed in order to evaluate the

discrepancies between the vir-T1c and ref-T1c images. These automated voxel-wise metrics were also computed between the low-T1c and ref-T1c sequences, in order to provide baseline performance measures: differences were compared using two-tailed t-tests. For all performance metrics and MRI sequences, only the brain voxels were evaluated. The receiver operating characteristic curves were computed using Scikit-Learn, following the methodology of (36): the precontrast T1 sequences were subtracted from the corresponding vir-T1c, low-T1c and ref-T1c images in order to define enhancement maps, which were 0-clipped. The reference maps were binarized using Otsu filters.

Reader study protocol

A read was performed in order to evaluate the suitability of vir-T1c images for lesion detection, taking ref-T1c as the reference standard. The suitability of low-T1c sequences was also evaluated, in order to provide a baseline for comparison. Two radiologists, one neuroradiologist specialized in oncology with 10 years of experience in reporting brain MRIs (X.X.), one in-training radiology resident with 4 months of experience in neuro-oncology (Y.Y.), analyzed in sub-regions both low-T1c and vir-T1c images for the 38 patients included in the test sample. These reads were organized in two phases, separated by an idle period of two weeks to eliminate recall bias. During the first phase, both readers independently annotated one image per test patient, half of them being low-T1c and the other 19 being vir-T1c. The second phase mirrored the first one, in such a way that both low-T1c and vir-T1c images were read once by both radiologists for all test patients, for a total of 152 reads. The most experienced radiologist then read the 38 ref-T1c sequences in order to identify reference lesions. Readers were instructed to grade the overall image quality on a 5-point Likert scale ranging from 1 (poor) to 5 (excellent), and to mark the long and short orthogonal axes of all visible enhancing lesions. Annotations were established for each image independently, without access to any complementary information at the exception of the precontrast T1 sequence. Exams with clearly more than 20 lesions were excluded, in order to circumscribe the annotation burden. A specific annotation tool based on 3D Slicer v4.11.0 was developed to accelerate the image review process. Readers were not blinded to the nature of the postcontrast images, because it could be easily guessed from the image appearance and would have no bias-reducing effect. Reading times were recorded for both readers.

Statistical analysis

Lesion tracking was performed automatically, by pairing together lesions whose centers were separated by a distance smaller than the halved average of their long and short axes. For each reader, the lesion detection sensitivity (SE) was computed as the ratio of the number of lesions detected in the low-T1c or vir-T1c images to the number of lesions identified in corresponding ref-T1c images. The false detection rate (FDR) was computed as the average number of false positive lesions found in low-T1c or vir-T1c images across test patients. The positive predictive value of detected lesions (PPV) was defined as the fraction of lesions found in the low-T1c or vir-T1c images which correspond to a reference lesion found in the ref-T1c sequences. The F1-score (F1) was computed as the harmonic mean between SE and PPV. These performance metrics were computed in nested evaluation configurations where only lesions larger than 0 (i.e. all lesions), 2.5, 5, 7.5 and 10mm were taken into consideration. The differences between low-T1c and vir-T1c detection performances were compared using the McNemar mid-p test for SE (42) and the Wilcoxon signed-rank test for FDR. Clopper-Pearson and Student's t-distribution 95% confidence intervals were determined for SE and FDR respectively. Average reading times and image quality grades were compared between radiologists and postcontrast modalities using t-tests. All analyses were conducted using SciPy 1.5.2. All tests were two-tailed, and the 5% level was used for statistical significance.

Results

Qualitative inspection

Figure 4 displays the algorithm's input, output, and ground-truth images for three patients from the test sample, either with brain metastases, glioma, or no enhancing lesion.

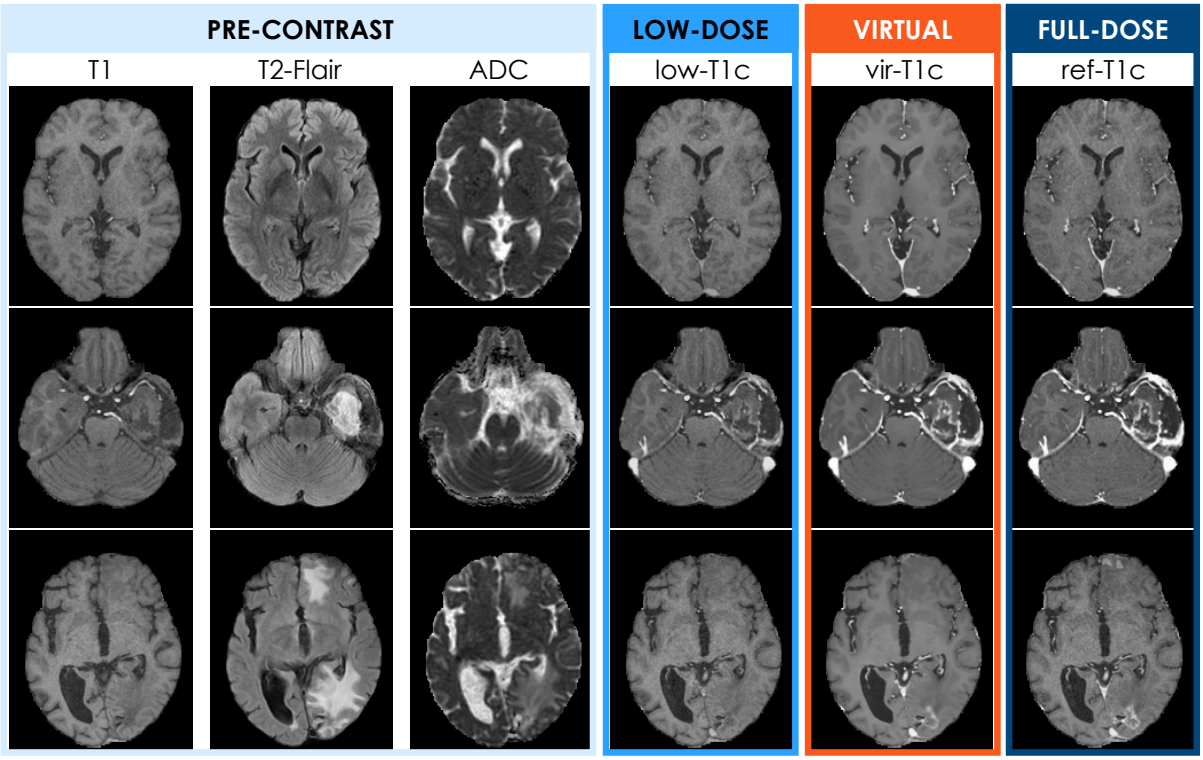

Figure 4: Axial slices of T1, T2-Flair, apparent diffusion coefficient (ADC), low-dose contrast T1 (low-T1c), virtual contrast T1 (vir-T1c) and full-dose contrast T1 (ref-T1c) images (in columns, from left to right) for three example patients. Qualitatively, the vir-T1c images seemed realistic. No obvious discrepancy between vir-T1c and ref-T1c images can be visually identified for the subject without enhancing lesion (top row: woman, 64 years old) and the patient with glioblastoma (middle row: woman, 40 years old). The last patient (bottom row: man, 74 years old) has brain metastases, clearly visible in both frontal and occipital locations on the ref-T1c image. The occipital lesion is satisfyingly restituted by the deep learning method in the sense that the lesion enhancement level and contour delineation appear similar to their reference counterparts. The frontal lesions however are largely absent from the predicted vir-T1c image. Although hyperintense frontal and occipital regions are visible in the corresponding input T2-Flair and ADC map, no frontal enhancement can be observed in the low-T1c, whereas an occipital enhancement weakly appears.

Voxel-wise analysis of the prediction performance

As detailed in Table 3, the predicted vir-T1c images reached a structural similarity index of 87.1% with respect to the ref-T1c sequences. The peak signal-to-noise ratio was 31.6dB and the area under the receiver operating characteristic curve was 99.0%. For all metrics, the vir-T1c images achieved superior performance when compared to low-T1c, highlighting the contribution of the deep leaning method. The average receiver operating curves are plotted in Figure 5. The operating point which maximizes the Youden's index for the vir-T1c sequences corresponds to a voxel-wise sensitivity of 96.4% and a

corresponding specificity of 94.8%. These values compare favorably with the related work (36), although comparisons on different data samples must be interpreted with care.

	low-T1c	**vir-T1c**	***low vs. vir***
SSIM (%)	80.4 (±6.1)	87.1 (±4.8)	*p<.001*
PSNR (dB)	27.4 (±1.8)	31.6 (±2.0)	*p<.001*
AUC (%)	96.4 (±3.1)	99.0 (±1.3)	*p<.001*

Table 3: Averaged structural similarity index (SSIM), peak signal-to-noise ratio (PSNR), and area under the receiver operating characteristic curve for the low-dose T1 sequences (low-T1c) and virtual images (vir-T1c) predicted by the deep network. Standard deviations across all test patients are given in parentheses. The performance differences between low-T1c and vir-T1c images are tested using t-tests, and p-values are reported. Significant differences (for the 5% level) are emphasized using a grey background.

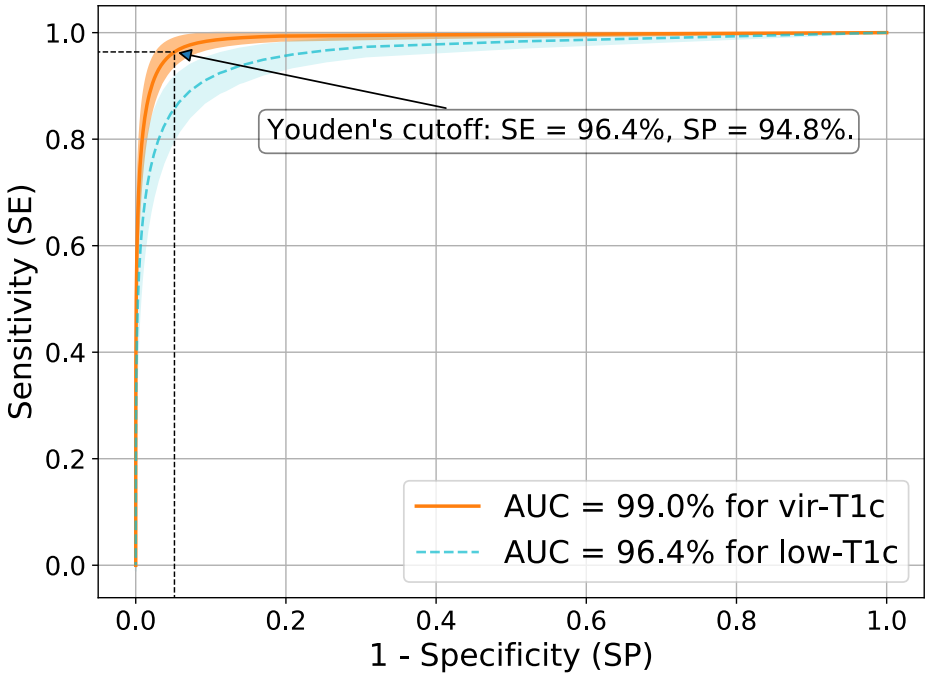

Figure 5: Average receiver operating curves for the virtual (vir-T1c) and low-dose (low-T1c) sequences, with 95% confidence intervals. The corresponding AUC values are given in the legend. Voxel-wise sensitivity and specificity are given for the Youden's operating point of the vir-T1c curve.

Image quality grades

The vir-T1c images obtained superior image quality grades than both low-T1c (p<.001 for both readers) and ref-T1c (p=.008 for R1, not applicable for R2). Table 5 provides the detailed results. On average across readers, image quality was graded 3.3/5 for vir-T1c, 2.9/5 for ref-T1c, and 2.1/5 for low-T1c. No significant difference was found between readers.

	IQ ±std (grades on a 5-point scale)			***low vs. ref***	***vir vs. ref***	***low vs. vir***
	low-T1c	**vir-T1c**	**ref-T1c**			
Reader 1	2.1 (±0.7)	3.3 (±0.8)	2.9 (±0.9)	*p<.001*	*p=.008*	*p<.001*

Reader 2	2.1 (±0.8)	3.3 (±0.7)	*n/a*	*n/a*	*p<.001*
R1 vs. R2	*p=.70*	*p=.68*			

Table 4: Overall image quality (IQ) subjective scores, expressed on 5-point Likert scales ranging from 1 (poor) to 5 (excellent). Performance across postcontrast modalities and across readers are compared using two-tailed t-tests, and p-values are reported. They are emphasized using a grey background or bold fonts if the 5% significance level is met.

Lesion detection performance

A total of 91 enhancing lesions were identified on the ref-T1c images. The smallest lesion had a long axis length of 2.6mm, for an average size across lesions of 14mm (±12 of standard deviation). Table 5 details the lesion detection performance achieved by vir-T1c images, taking the ref-T1c lesions as reference. On average across readers, SE of vir-T1c reached 83% for lesions larger than 10mm. The corresponding FDR was 0.08 lesion/patient, giving in turn a PPV of 92% for detected lesions. The F1-score was 88%. Lesion detection performance dropped when smaller lesions where considered. For lesions larger than 5mm, SE was 67% and FDR was 0.42 lesion/patient. Accordingly, PPV decreased to 76% and F1 to 71%. With all lesions included as detection targets regardless of their size, SE was 56%, FDR 0.50 lesion/patient, PPV 73% and F1 63%.

Table 5 also reports the lesion detection performance of low-T1c sequences, in order to provide a baseline for comparison. Significantly more reference lesions were identified on vir-T1c images than on low-T1c: respective SE were 53% versus 42% (p=.01) for reader 1 (R1: X.X.), and 59% versus 37% (p<.001) for reader 2 (R2: Y.Y.). However, more false positive lesions were identified on vir-T1c images than on low-T1c: respective FDR were 0.47 lesion/patient versus 0.11 lesion/patient for R1 (p=.04), and 0.53 lesion/patient versus 0.11 lesion/patient for R2 (p=.02). In consequence, for both readers PPV was lower for vir-T1c than for low-T1c images (73% versus 90% for R1, 73% versus 89% for R2). The composite F1 metric, which summarizes SE and PPV, was however higher for vir-T1c than for low-T1c: 61% versus 57% for R1 and 65% versus 53% for R2.

SE also increased for low-T1c when only larger lesions were considered, but SE remained always more elevated for vir-T1c by a margin of 14 percentage points at least on average across readers. Similarly, F1 remained consistently higher for vir-T1c images. On the other hand, although the FDR and PPV metrics were initially at the advantage of low-T1c images with all lesion sizes included, they reached similar values at the 10mm threshold: 0.08 lesion/patient for FDR and 91% or 92% for PPV.

		Reader 1			**Reader 2**			***R1 vs. R2***		**Reader average**	
		low-T1c	**vir-T1c**	***low vs. vir***	**low-T1c**	**vir-T1c**	***low vs. vir***	***low-T1c***	***vir-T1c***	**low-T1c**	**vir-T1c**
All lesions (≥2.5 mm)	**SE**	42% (38/91) [32; 53]	53% (48/91) [42; 63]	*p=.01*	37% (34/91) [27; 48]	59% (54/91) [49; 70]	*p<.001*	*p=.13*	*p=.07*	40%	**56%**
	FDR	0.11 (4/38) [±0.10]	0.47 (18/38) [±0.38]	*p=.04*	0.11 (4/38) [±0.10]	0.53 (20/38) [±0.39]	*p=.02*	*p>.99*	*p=.68*	**0.11**	0.50
	PPV	90% (38/42)	73% (48/66)	*n/a*	89% (34/38)	73% (54/74)	*n/a*	*n/a*		**90%**	73%

	F1	57%	61%		53%	65%				55%	**63%**
≥5 mm	**SE**	51% (38/74) [39; 63]	64% (47/74) [52; 74]	*p=.02*	46% (34/74) [34; 58]	70% (52/74) [59; 80]	*p<.001*	*p=.13*	*p=.11*	49%	**67%**
	FDR	0.11 (4/38) [±0.10]	0.39 (15/38) [±0.33]	*p=.07*	0.11 (4/38) [±0.10]	0.45 (17/38) [±0.36]	*p=.04*	*p>.99*	*p=.78*	**0.11**	0.42
	PPV	90% (38/42)	76% (47/62)	*n/a*	89% (34/38)	75% (52/69)	*n/a*	*n/a*		**90%**	76%
	F1	66%	69%		61%	73%				63%	**71%**
≥7.5 mm	**SE**	65% (34/52) [51; 78]	81% (42/52) [67; 90]	*p=.02*	60% (31/52) [45; 73]	87% (45/52) [74; 94]	*p<.001*	*p=.22*	*p=.22*	62%	**84%**
	FDR	0.11 (4/38) [±0.10]	0.21 (8/38) [±0.19]	*p=.23*	0.08 (3/38) [±0.09]	0.26 (10/38) [±0.28]	*p=.13*	*p=.56*	*p=.76*	**0.09**	0.24
	PPV	89% (34/38)	84% (42/50)	*n/a*	91% (31/34)	82% (45/55)	*n/a*	*n/a*		**90%**	83%
	F1	76%	82%		72%	84%				74%	**83%**
≥10 mm	**SE**	71% (30/42) [55; 84]	83% (35/42) [69; 93]	*p=.07*	67% (28/42) [50; 80]	83% (35/42) [69; 93]	*p=.02*	*p=.38*	*p>.99*	69%	**83%**
	FDR	0.11 (4/38) [±0.10]	0.03 (1/38) [±0.05]	*p=.08*	0.05 (2/38) [±0.07]	0.13 (5/38) [±0.14]	*p=.18*	*p=.16*	*p=.046*	0.08	0.08
	PPV	88% (30/34)	97% (35/36)	*n/a*	93% (28/30)	88% (35/40)	*n/a*	*n/a*		91%	**92%**
	F1	79%	90%		78%	85%				78%	**88%**

Table 5: Sensitivity (SE, expressed in percentages of the total number of reference lesions), false detection rate (FDR, expressed in average number per patient), positive predictive value (PPV, expressed in percentages of the total number of detected lesions), and F1-score (F1, harmonic mean between SE and PPV), in nested evaluation configurations where only lesions larger than some threshold (for their long axis) are considered. Two-tailed McNemar mid-p and Wilcoxon signed-rank tests are used for SE and FDR respectively, and p-values are reported. They are emphasized using a color background if the 5% significance level is met. Clopper-Pearson and Student's t-distribution 95% confidence intervals are also reported for SE and FDR respectively, between brackets. Average performance metrics across readers for low-T1c and vir-T1c postcontrast modalities are indicated in the two rightmost columns.

Influence of the reader's experience

Table 5 showed that the in-training reader (R2) benefited from the virtual images in larger proportions than the experienced reader (R1): the all-lesion sensitivity of R2 was higher with vir-T1c in comparison to low-T1c by a margin of 22%, versus 11% for R1. This difference came from both the higher SE of R1 with low-T1c (42% versus 37% for R2), and the higher SE of R2 with vir-T1c images (59% versus 53% for R1). This performance pattern, more lesions identified by R1 than R2 on low-T1c and vice-versa for vir-T1c, held for both true and false positives, and across all lesion size thresholds. Table 6 reports the

average reading times across postcontrast modalities and readers. For both low-T1c and vir-T1c, the experienced reader R1 was faster than the in-training reader R2 (p<.001 and p=.01 respectively). As the ref-T1c images were only read by R1, no comparison could be performed in this case.

	RT ±std (minutes' and seconds'')			***low vs. ref***	***vir vs. ref***	***low vs. vir***
	low-T1c	**vir-T1c**	**ref-T1c**			
Reader 1	**2'25''** (±2'30'')	**2'48''** (±2'20'')	3'59'' (±3'42'')	*p=.04*	*p=.12*	*p=.51*
Reader 2	5'21'' (±4'48'')	4'42'' (±3'42'')	*n/a*	*n/a*		*p=.52*
R1 vs. R2	***p<.001***	***p=.01***				

Table 6: Average exam reading times (RT), expressed in seconds. Results across postcontrast modalities and across readers are compared using two-tailed t-tests, and p-values are reported. They are emphasized using a grey background or bold fonts if the 5% significance level is met.

Discussion

A pressing challenge for contemporary radiology is to reduce the routine usage of gadolinium-based contrast agents (GBCAs) without degrading diagnostic accuracy. Among other leads, deep learning algorithms able to enhance low-dose contrast-enhanced MRI sequences have been recently proposed and qualitatively validated. In this study, we trained a deep network to predict, from precontrast and 25%-dose MRI sequences, virtual images that approximate reference full-dose contrast-enhanced T1 images (see Figure 3).

The images predicted by the deep learning method reached a structural similarity index of 87.1% with respect to the reference T1 MRIs, a peak signal-to-noise ratio of 31.6dB, and an area under the receiver operating curve of 99.0%. At the operating point which maximizes the Youden's index, voxel-wise sensitivity and specificity reached 96.4% and 94.8% respectively. In addition to this initial and automated analysis, a reader study showed that the virtual images achieved a lesion detection SE of 88% for lesions larger than 10mm, along with an FDR of 0.08 lesion/patient, a PPV of 92% for detected lesion, and a F1-score of 88%, unveiling a promising performance regime. In addition, the virtual images were better graded than their reference counterparts in terms of overall quality (p=.008). However, the reader study also showed that many small-sized lesions were missed by the method: the average SE dropped to 56% when all lesions were included regardless of their size. The FDR increased to 0.50 lesion/patient, showing that small-size false positive lesions are sometimes created by the method. As a result, the PPV dropped to 73%, and F1 to 63%.

From a technical point of view, this deep learning method can be understood as an incremental improvement over previously published approaches (35,36), notably in terms of network capacity and data set size. To the best of our knowledge however, this study is the first to quantitatively evaluate the suitability of virtual images for lesion detection. Although the grade-based reading protocols followed in (35,36) are interesting first-line validation approaches, their results cannot anticipate the true clinical impact of such image prediction methods. In particular, our study showed, on the one hand, that the virtual images were largely preferred by radiologists for their overall appearance qualities (in line with (35,36)), but, on the other hand, that many lesions were actually absent. Small-sized lesions proved to be particularly challenging.

This modest absolute SE is the major limitation of the deep learning method, not suitable at this stage for straightforward clinical replacement of standard 0.1mmol/kg contrast sequences. The prediction

of false positive lesions is a second limitation, with similar consequences. These results show that full-dose injections of GBCAs remain essential for accurate diagnosis in neuro-oncology, at least in the setting of this study. However, it should be noted that the reading radiologists were not particularly trained to read the virtual MRIs, whose general smooth aspect was found qualitatively quite different from standard sequences. We may hypothesize that this lack of training acted as an adverse evaluation bias against virtual images, which might have led to an underestimation of their intrinsic potential. Furthermore, targeting a sub-millimetric resolution in the resampling step of the preprocessing pipeline may offer a natural yet powerful perspective of technical improvement for our method. Similarly to (43) in the case of biomedical image segmentation, cascaded network architectures could be advantageoulsy exploited in order to handle the resulting increase in GPU memory footprint.

To highlight more minor limits, we incidentally observed that automatic skull-stripping led, in some cases, to the removal of benign peripheral lesions such as meningioma. Future work should weigh the opportunity to either improve or avoid altogether this preprocessing step. In the same vein, future work may also question wether all four input modalities (precontrast and low-dose T1, T2-Flair, ADC map) are significantly contributing to the performance of the method, similarly to what was done in (36,37) using simple voxel-wise metrics. Indeed, requiring multiple sequences can be expected to limit in practice the usability and the robustness of the method.

More fundamentally, future work should take advantage of the promising performance of our deep learning method for medium-sized to large brain lesions (larger than 10mm along their long axis) to demonstrate its suitability as a replacement of MRI sequences injected at 0.1mmol/kg. We believe that a population such as, for instance, returning patients with benign tumor, low-grade glioma or multiple sclerosis could benefit from the proposed algorithm to minimize their exposure to GBCAs without sacrificing the quality of their follow-up visits. Beyond neuroimaging, abdominal imaging could also benefit from contrast dose reduction. The proposed methodology could finally be opportunistically repurposed from a dose-minimization to a contrast-maximization objective by using standard and double-dose sequences as respective input and ground truth images to train the deep network, with the aim to improve the sensititvity of routine contrast-enhanced MRI. In any case, such image prediction methods would need to be further evaluated in a prospective multicentric multireader fashion in order to obtain all necessary regulatory approvals and become available to clinical pratice. As an intermediate step on this validation roadmap, our study demonstrated the importance to include detection endpoints and to stratify the performance with respect to the lesion size.

References

1. Runge VM. Safety of the gadolinium-based contrast agents for magnetic resonance imaging, focusing in part on their accumulation in the brain and especially the dentate nucleus. Invest Radiol. LWW; 2016;51(5):273–279. doi: 10.1097/RLI.0000000000000273.

2. Caravan P, Ellison JJ, McMurry TJ, Lauffer RB. Gadolinium(III) chelates as MRI contrast agents: Structure, dynamics, and applications. Chem Rev. ACS; 1999;99(9):2293–2352. doi: 10.1021/cr980440x.

3. Grobner T. Gadolinium - A specific trigger for the development of nephrogenic fibrosing dermopathy and nephrogenic systemic fibrosis? Nephrol Dial Transplant. Oxford Academic; 2006;21(4):1104–1108. doi: 10.1093/ndt/gfk062.

4. Marckmann P, Skov L, Rossen K, et al. Nephrogenic systemic fibrosis: Suspected causative role of gadodiamide used for contrast-enhanced magnetic resonance imaging. J. Am. Soc. Nephrol. American Society of Nephrology; 2006. p. 2359–2362. doi: 10.1681/ASN.2006060601.

5. Kanda T, Ishii K, Kawaguchi H, Kitajima K, Takenaka D. High signal intensity in the dentate nucleus and globus pallidus on unenhanced T1-weighted MR images: Relationship with increasing cumulative dose of a gadoliniumbased contrast material. Radiology. Radiological Society of North America Inc.; 2014;270(3):834–841. doi: 10.1148/radiol.13131669.

6. McDonald RJ, McDonald JS, Kallmes DF, et al. Intracranial gadolinium deposition after contrast-enhanced MR imaging. Radiology. Radiological Society of North America Inc.; 2015;275(3):772–782. doi: 10.1148/radiol.15150025.

7. Radbruch A, Weberling LD, Kieslich PJ, et al. Gadolinium retention in the dentate nucleus and globus pallidus is dependent on the class of contrast agent. Radiology. Radiological Society of North America Inc.; 2015;275(3):783–791. doi: 10.1148/radiol.2015150337.

8. Zivadinov R, Bergsland N, Hagemeier J, et al. Cumulative gadodiamide administration leads to brain gadolinium deposition in early MS. Neurology. Lippincott Williams and Wilkins; 2019;93(6):E611–E623. doi: 10.1212/WNL.0000000000007892.

9. Debevits JJ, Munbodh R, Bageac D, et al. Gray Matter Nucleus Hyperintensity after Monthly Triple-Dose Gadopentetate Dimeglumine with Long-term Magnetic Resonance Imaging. Invest Radiol. Lippincott Williams and Wilkins; 2020;55(10):629–635. doi: 10.1097/RLI.0000000000000663.

10. Strzeminska I, Factor C, Robert P, et al. Long-Term Evaluation of Gadolinium Retention in Rat Brain after Single Injection of a Clinically Relevant Dose of Gadolinium-Based Contrast Agents. Invest Radiol. Lippincott Williams and Wilkins; 2020;55(3):138–143. doi: 10.1097/RLI.0000000000000623.

11. Smith TE, Steven A, Bagert BA. Gadolinium deposition in neurology clinical practice. Ochsner J. Ochsner Clinic; 2019;19(1):17–25. doi: 10.31486/toj.18.0111.

12. Radbruch A, Richter H, Fingerhut S, et al. Gadolinium Deposition in the Brain in a Large Animal Model: Comparison of Linear and Macrocyclic Gadolinium-Based Contrast Agents. Invest Radiol. Lippincott Williams and Wilkins; 2019;54(9):531–536. doi: 10.1097/RLI.0000000000000575.

13. Bower D V., Richter JK, Von Tengg-Kobligk H, Heverhagen JT, Runge VM. Gadolinium-Based MRI Contrast Agents Induce Mitochondrial Toxicity and Cell Death in Human Neurons, and Toxicity Increases With Reduced Kinetic Stability of the Agent. Invest Radiol. Lippincott Williams and Wilkins; 2019;54(8):453–463. doi: 10.1097/RLI.0000000000000567.

14. Radbruch A, Richter H, Bücker P, et al. Is Small Fiber Neuropathy Induced by Gadolinium-Based Contrast Agents? Invest Radiol. Lippincott Williams and Wilkins; 2020;55(8):473–480. doi: 10.1097/RLI.0000000000000677.

15. Wang S, Hesse B, Roman M, et al. Increased Retention of Gadolinium in the Inflamed Brain after Repeated Administration of Gadopentetate Dimeglumine: A Proof-of-Concept Study in

Mice Combining ICP-MS and Micro- And Nano-SR-XRF. Invest Radiol. Lippincott Williams and Wilkins; 2019;54(10):617–626. doi: 10.1097/RLI.0000000000000571.

16. European Medicines Agency. EMA's final opinion confirms restrictions on use of linear gadolinium agents in body scans. 2017. Document Reference: EMA/625317/2017.

17. FDA warns that gadolinium-based contrast agents (GBCAs) are retained in the body. U.S. Food Drug Adm. Website. 2017. www.fda.gov/Drugs/DrugSafety/ucm589213.htm. Accessed February 25, 2021.

18. Falk Delgado A, Van Westen D, Nilsson M, et al. Diagnostic value of alternative techniques to gadolinium-based contrast agents in MR neuroimaging—a comprehensive overview. Insights Imaging. Springer Verlag; 2019. doi: 10.1186/s13244-019-0771-1.

19. Zhou J, Payen JF, Wilson DA, Traystman RJ, Van Zijl PCM. Using the amide proton signals of intracellular proteins and peptides to detect pH effects in MRI. Nat Med. Nat Med; 2003;9(8):1085–1090. doi: 10.1038/nm907.

20. Ward KM, Aletras AH, Balaban RS. A New Class of Contrast Agents for MRI Based on Proton Chemical Exchange Dependent Saturation Transfer (CEST). J Magn Reson. Academic Press Inc.; 2000;143(1):79–87. doi: 10.1006/jmre.1999.1956.

21. Kamimura K, Nakajo M, Yoneyama T, et al. Amide proton transfer imaging of tumors: theory, clinical applications, pitfalls, and future directions. Jpn. J. Radiol. Springer Tokyo; 2019. p. 109–116. doi: 10.1007/s11604-018-0787-3.

22. Burtscher IM, Holtas S. Proton MR spectroscopy in clinical routine. J Magn Reson Imaging. J Magn Reson Imaging; 2001;13(4):560–567. doi: 10.1002/jmri.1079.

23. Haller S, Zaharchuk G, Thomas DL, Lovblad KO, Barkhof F, Golay X. Arterial spin labeling perfusion of the brain: Emerging clinical applications. Radiology. Radiological Society of North America Inc.; 2016. p. 337–356. doi: 10.1148/radiol.2016150789.

24. Le Bihan D, Breton E, Lallemand D, Grenier P, Cabanis E, Laval-Jeantet M. MR imaging of intravoxel incoherent motions: Application to diffusion and perfusion in neurologic disorders. Radiology. Radiology; 1986;161(2):401–407. doi: 10.1148/radiology.161.2.3763909.

25. Kim DY, Kim HS, Goh MJ, Choi CG, Kim SJ. Utility of intravoxel incoherent motion MR imaging for distinguishing recurrent metastatic tumor from treatment effect following gamma knife radiosurgery: Initial experience. Am J Neuroradiol. American Society of Neuroradiology; 2014;35(11):2082–2090. doi: 10.3174/ajnr.A3995.

26. Suh CH, Kim HS, Lee SS, et al. Atypical imaging features of primary central nervous system lymphoma that mimics glioblastoma: Utility of intravoxel incoherent motion MR imaging. Radiology. Radiological Society of North America Inc.; 2014;272(2):504–513. doi: 10.1148/radiol.14131895.

27. Federau C, Meuli R, O'Brien K, Maeder P, Hagmann P. Perfusion measurement in brain gliomas with intravoxel incoherent motion MRI. Am J Neuroradiol. AJNR Am J Neuroradiol; 2014;35(2):256–262. doi: 10.3174/ajnr.A3686.

28. Bendszus M, Roberts D, Kolumban B, et al. Dose Finding Study of Gadopiclenol, a New Macrocyclic Contrast Agent, in MRI of Central Nervous System. Invest Radiol. Lippincott Williams and Wilkins; 2020;55(3):129–137. doi: 10.1097/RLI.0000000000000624.

29. Rodríguez-Galván A, Rivera M, García-López P, Medina LA, Basiuk VA. Gadolinium-containing carbon nanomaterials for magnetic resonance imaging: Trends and challenges. J Cell Mol Med. Blackwell Publishing Inc.; 2020;24(7):3779–3794. doi: 10.1111/jcmm.15065.

30. Lancelot E, Raynaud JS, Desché P. Current and Future MR Contrast Agents: Seeking a Better Chemical Stability and Relaxivity for Optimal Safety and Efficacy. Invest. Radiol. Lippincott Williams and Wilkins; 2020;55(9):578–588. doi: 10.1097/RLI.0000000000000684.

31. Erstad DJ, Ramsay IA, Jordan VC, et al. Tumor Contrast Enhancement and Whole-Body Elimination of the Manganese-Based Magnetic Resonance Imaging Contrast Agent Mn-PyC3A. Invest Radiol. Lippincott Williams and Wilkins; 2019;54(11):697–703. doi: 10.1097/RLI.0000000000000593.

32. Sudarshana DM, Nair G, Dwyer JT, et al. Manganese-enhanced MRI of the brain in healthy volunteers. Am J Neuroradiol. American Society of Neuroradiology; 2019;40(8):1309–1316. doi: 10.3174/ajnr.A6152.

33. Bales BC, Grimmond B, Johnson BF, et al. Fe-HBED Analogs: A Promising Class of Iron-Chelate Contrast Agents for Magnetic Resonance Imaging. Contrast Media Mol Imaging. Hindawi Limited; 2019;2019. doi: 10.1155/2019/8356931.

34. Dadfar SM, Roemhild K, Drude NI, et al. Iron oxide nanoparticles: Diagnostic, therapeutic and theranostic applications. Adv. Drug Deliv. Rev. Elsevier B.V.; 2019;138:302–325. doi: 10.1016/j.addr.2019.01.005.

35. Gong E, Pauly JM, Wintermark M, Zaharchuk G. Deep learning enables reduced gadolinium dose for contrast-enhanced brain MRI. J Magn Reson Imaging. 2018;48(2):330–340. doi: 10.1002/jmri.25970.

36. Kleesiek J, Morshuis JN, Isensee F, et al. Can Virtual Contrast Enhancement in Brain MRI Replace Gadolinium?: A Feasibility Study. Invest Radiol. LWW; 2019;54(10):653–660. doi: 10.1097/RLI.0000000000000583.

37. Bône A, Ammari S, Lamarque J-P, et al. Contrast-enhanced brain MRI synthesis with deep learning: key input modalities and asymptotic performance. IEEE Int Symp Biomed Imaging. 2021;18:1159–1163. doi: 10.1109/ISBI48211.2021.9434029.

38. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: From error visibility to structural similarity. IEEE Trans Image Process. IEEE; 2004;13(4):600–612. doi: 10.1109/TIP.2003.819861.

39. Gordon Y, Partovi S, Müller-Eschner M, et al. Dynamic contrast-enhanced magnetic resonance imaging: fundamentals and application to the evaluation of the peripheral perfusion. Cardiovasc Diagn Ther. AME Publications; 2014;4(2):147–14764. doi: 10.3978/j.issn.2223-3652.2014.03.01.

40. Thust SC, Heiland S, Falini A, et al. Glioma imaging in Europe: A survey of 220 centres and

recommendations for best clinical practice. Eur Radiol. Springer; 2018;28(8):3306–3317. doi: 10.1007/s00330-018-5314-5.

41. Milletari F, Navab N, Ahmadi SA. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. Proc. Int. Conf. 3D Vision, 2016;4: 565–571. doi: 10.1109/3DV.2016.79.

42. Fagerland MW, Lydersen S, Laake P. The McNemar test for binary matched-pairs data: Mid-p and asymptotic are better than exact conditional. BMC Med Res Methodol. BioMed Central; 2013;13(1):1–8. doi: 10.1186/1471-2288-13-91.

43. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. Nature Research; 2021;18(2):203–211. doi: 10.1038/s41592-020-01008-z.