Phantom-based quality assurance for multicenter quantitative MRI in locally advanced cervical cancer

Houdt, P.J. van; Kallehauge, J.F.; Tanderup, K.; Nout, R.; Zaletelj, M.; Tadic, T.; ... ; Embrace Collaborative Grp

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original Article

Phantom-based quality assurance for multicenter quantitative MRI in locally advanced cervical cancer

Petra J. van Houdt [a,*], Jesper F. Kallehauge [b], Kari Tanderup [c], Remi Nout [d], Marko Zaletelj [e], Tony Tadic [f], Zdenko J. van Kesteren [g], Cornelius A.T. van den Berg [h], Dietmar Georg [i], Jean-Charles Côté [j], Ives R. Levesque [k], Jamema Swamidas [l], Eirik Malinen [m], Sven Telliskivi [n], Patrik Brynolfsson [o], Faisal Mahmood [p,q], Uulke A. van der Heide [a], and EMBRACE Collaborative Group

[a] *Department of Radiation Oncology, the Netherlands Cancer Institute, Amsterdam, the Netherlands;* [b] *Danish Centre for Particle Therapy, Aarhus;* [c] *Department of Clinical Medicine, Aarhus University Hospital, Denmark;* [d] *Department of Radiation Oncology, Leiden University Medical Center, the Netherlands;* [e] *Department of Radiotherapy, Institute of Oncology Ljubljana, Slovenia;* [f] *Radiation Medicine Program, Princess Margaret Cancer Center, Toronto, Canada;* [g] *Department of Radiation Oncology, Amsterdam University Medical Center;* [h] *Department of Radiotherapy, University Medical Center Utrecht, the Netherlands;* [i] *Division of Medical Radiation Physics, Department of Radiation Oncology, Medical University Of Vienna, Austria;* [j] *Department of Radiation Oncology, Centre Hospitalier de l'Universite de Montreal, Canada;* [k] *Medical Physics Unit and Gerald Bronfman Department of Oncology, McGill University, Montreal, Canada;* [l] *Department of Radiation Oncology, Tata Memorial Centre, Mumbai, India;* [m] *Department of Medical Physics, Oslo University Hospital, Norway;* [n] *Department of Radiation Oncology, North-Estonia Medical Centre, Tallinn, Estonia;* [o] *Department of Translational Sciences, Skåne University Hospital, Lund, Sweden;* [p] *Department of Oncology, Odense University Hospital; and* [q] *Department of Clinical Research, University of Southern Denmark, Odense, Denmark*

ARTICLE INFO

Article history:
Received 4 June 2020
Received in revised form 7 September 2020
Accepted 8 September 2020
Available online 12 September 2020

Keywords:
Uterine cervical neoplasms
Magnetic resonance imaging
Radiotherapy

ABSTRACT

Background and purpose: A wide variation of MRI systems is a challenge in multicenter imaging biomarker studies as it adds variation in quantitative MRI values. The aim of this study was to design and test a quality assurance (QA) framework based on phantom measurements, for the quantitative MRI protocols of a multicenter imaging biomarker trial of locally advanced cervical cancer.

Materials and methods: Fifteen institutes participated (five 1.5 T and ten 3 T scanners). Each institute optimized protocols for T2, diffusion-weighted imaging, T1, and dynamic contrast-enhanced (DCE–)MRI according to system possibilities, institutional preferences and study-specific constraints. Calibration phantoms with known values were used for validation. Benchmark protocols, similar on all systems, were used to investigate whether differences resulted from variations in institutional protocols or from system variations. Bias, repeatability (%RC), and reproducibility (%RDC) were determined. Ratios were used for T2 and T1 values.

Results: The institutional protocols showed a range in bias of 0.88–0.98 for T2 (median %RC = 1%; %RDC = 12%), −0.007 to 0.029 × 10^{-3} mm^2/s for the apparent diffusion coefficient (median %RC = 3%; %RDC = 18%), and 0.39–1.29 for T1 (median %RC = 1%; %RDC = 33%). For DCE a nonlinear vendor-specific relation was observed between measured and true concentrations with magnitude data, whereas the relation was linear when phase data was used.

Conclusion: We designed a QA framework for quantitative MRI protocols and demonstrated for a multicenter trial for cervical cancer that measurement of consistent T2 and apparent diffusion coefficient values is feasible despite protocol differences. For DCE–MRI and T1 mapping with the variable flip angle method, this was more challenging.

The interest in quantitative MRI (qMRI) for radiation oncology is growing as it has the potential to improve outcome prediction and the assessment of treatment response [1,2]. The two most investigated techniques are diffusion weighted imaging (DWI), providing images associated with cell density modulated water mobility, and dynamic contrast-enhanced (DCE–) MRI, depicting tissue perfusion and vascular permeability. Nonetheless, the use of qMRI in clinical practice is currently limited [1]. Most studies using DWI and DCE–MRI have been performed in a single-center setting or in collabora-

Abbreviations: ADC, apparent diffusion coefficient; DWI, diffusion-weighted MRI; DCE–MRI, dynamic contrast-enhanced MRI; LoA, limits of agreement; QA, quality assurance; QIBA, quantitative imaging biomarker alliance; qMRI, quantitative MRI; %RC, percentage repeatability coefficient; %RDC, percentage reproducibility coefficient; ROI, region of interest; SNR, signal-to-noise; wCV, within subject standard deviation.

[*] Corresponding author at: Department of Radiation Oncology, the Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, the Netherlands.

E-mail address: p.v.houdt@nki.nl (P.J. van Houdt).

https://doi.org/10.1016/j.radonc.2020.09.013
0167-8140/© 2020 The Authors. Published by Elsevier B.V.

tion between a few expert centers [3–7]. O'Connor et al. identified the challenges in imaging biomarker research and proposed a roadmap to go from feasibility studies to clinical implementation [8]. One of the challenges in multicenter imaging studies is that they involve a variety of MR systems, with different vendors, field strengths, scanner models, software versions, and imaging protocols [9]. This explains for example part of the variation in the apparent diffusion coefficient (ADC) from DWI between different studies [10].

One approach to reach consistency is to use standardized MRI protocols. The quantitative imaging biomarkers alliance (QIBA) is working towards consensus recommendations for standardized acquisition protocols for DWI and DCE–MRI based on existing literature [11]. For example, Malyarenko et al. showed an improved reproducibility of the ADC across different MR systems when a standardized DWI protocol was used [12]. However, standardization may be difficult to achieve in practice. As standardized protocols have to be feasible for the oldest system, this limits the use of advanced functionality of newer systems (e.g. parallel imaging techniques) affecting the scanning time for patients. Therefore, a more feasible approach for large-scale studies could be to give more freedom for protocol development such that the protocols can be optimized for each system individually based on system possibilities and local preferences. However, before initiating clinical trials quality assurance to assess the performance of the qMRI protocols between institutes will be even more crucial [8,13].

Quality assurance (QA) with phantoms can be used to estimate the bias, repeatability, and reproducibility of the qMRI parameters [9,11]. In this study, we designed a QA framework to reduce differences in qMRI parameters between institutes without having to use fixed MR protocols. As a first step we introduced a set of benchmark protocols to investigate whether deviations in qMRI values between institutes result from protocol differences or from system variations. Second, we measured qMRI values from protocols that were optimized by each institute. We applied this QA framework to the qMRI protocols optimized for a multicenter imaging biomarker discovery trial of locally advanced cervical cancer (IQ-EMBRACE, clinicaltrials.gov NCT03210428). In the trial patients undergo an extended pre-treatment MRI examination with T2 mapping, DWI, and DCE–MRI to assess whether they can be used to predict disease free survival. As the trial is a sub-study of the EMBRACE-II trial [14], the treatment of patients is homogeneous across centers creating a unique opportunity for biomarker discovery. The aim of this study was to design a QA framework and test it for the qMRI protocols of the IQ-EMBRACE trial to assess the variation in qMRI values in the participating institutes.

Materials and methods

Design of quality assurance framework

Fig. 1 illustrates the design of the QA framework for validation of T2, DWI, and DCE–MRI based on measurements with calibration phantoms with known values ($qMRI_{true}$). First, benchmark protocols, which are established protocols without acceleration, and too slow for clinical use, were tested. By comparing the measured values to $qMRI_{true}$ the effect of system variations were investigated, including differences in hardware, such as gradient systems and coil-set-up, as well as vendor-specific implementations of the sequences. Next, the MR protocols used for the patients in the IQ-EMBRACE trial were optimized per institute (i.e. institutional protocols). As system differences were assessed with benchmark protocols, any additional variations in qMRI values measured with the institutional protocols should originate from differences in parameter settings or sequence choice. The evaluation of DCE–MRI consisted of multiple steps, because direct evaluation of pharmacokinetic parameters (e.g. K^{trans}) with phantoms and benchmark protocols is not practically feasible. Consistent with the QIBA DCE–MRI profile, we therefore tested the performance of T1 mapping necessary for the conversion of DCE–MRI signal intensities to contrast agent concentration values, signal stability to investigate magnet stability, and signal linearity to investigate the accuracy of the conversion of signal intensity to concentration values [15]. As a result, a benchmark protocol was not included in the QA for DCE–MRI.

Data acquisition

In total 15 centers participated in the QA measurements with MR systems from three vendors (five 1.5 T systems and ten 3 T systems, Table 1). The measurements were done between 2017 and 2020.

The parameters of the benchmark protocols were specified in detail in Table 2. For T2 mapping, a non-accelerated multi-echo spin-echo sequence was used. For DWI, the echo-planar imaging (EPI) sequence for phantom scans specified by QIBA was used [16]. For T1 mapping, as one of the validation steps for DCE–MRI, an inversion recovery acquisition was used as the benchmark.

For the institutional sequences more freedom in protocol settings was allowed. However, a few minimal requirements were set, for example to ensure full tumor coverage (Table 3). Details of the institutional protocols are summarized in Supplementary Tables 1–4. All institutes used a multi-echo spin echo based

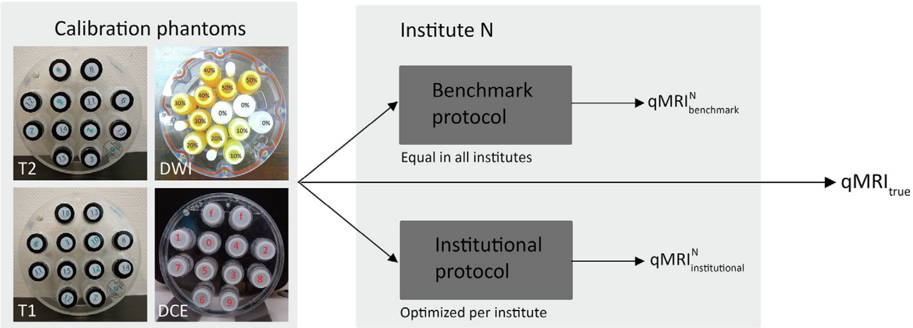

Fig. 1. Overview of the QA framework illustrated for one institute. Calibration phantoms were used with known values ($qMRI_{true}$) for T2, T1, ADC, and concentration values of DCE. First, benchmark protocols, which are established protocols without acceleration, were tested. By comparing the measured values to $qMRI_{true}$ the effect of system variations could be investigated. Next, the MR protocols used for the patients in the IQ-EMBRACE trial were optimized per institute (i.e. institutional protocols). As system differences were assessed with benchmark protocols, any additional variations in qMRI values measured with the institutional protocols should originate from differences in parameter settings or sequence choice.

Table 1
Description of all systems.

Institute Code	Field Strength	Vendor	Type
A	1.5 T	Philips	Ingenia
B	1.5 T	Philips	Ingenia
C	1.5 T	Philips	Ingenia
D	1.5 T	Siemens	Aera
E	1.5 T	GE	Optima
F	3 T	Philips	Ingenia
G	3 T	Philips	Ingenia
H	3 T	Philips	Ingenia
I	3 T	Philips	Ingenia
J	3 T	Siemens	Skyra
K	3 T	Siemens	Skyra
L	3 T	Siemens	Skyra
M	3 T	Siemens	Verio
N	3 T	GE	Discovery
O	3 T	GE	Signa

sequence for T2 mapping, except for institute E and N where separate T2-weighted images were acquired with different echo times. For DWI, all institutes used an EPI-based protocol. Maximum available gradient amplitudes and slew rates were recommended to achieve the shortest echo time to minimize geometrical distortions. Five institutes (A, C, J, M, N) included more *b*-values than *b* = 0, 200, and 1000 s/mm^2 according to institutional preferences. For DCE–MRI, every institute used a variable flip angle approach for T1 mapping. The DCE–MRI sequence itself was a spoiled-gradient echo sequence with or without a Dixon technique for fat suppression. Phase data from DCE–MRI was saved in ten institutes. B1 maps were acquired in seven of the ten institutes with a 3 T system, which was based on the actual flip angle imaging method or double angle method [17,18].

We used three copies of the Eurospin II TO5 phantom (Diagnostic Sonar LTD, Livingston, Scotland) for T2 assessment. A set of twelve gel samples was chosen with T2 values ranging between 49 to 212 ms at 296 K and 3 T according to the manufacturer. For DWI, we used three copies of the Diffusion Phantom Model 128 (High Precision Devices, Inc, Boulder, Colorado, USA) filled with ice water to obtain a temperature of 273.15 K at the beginning of the measurements per phantom instructions. For DCE–MRI we used two phantoms. For the evaluation of T1 mapping, the Eurospin II TO5 phantom was used with eleven samples of T1 values ranging from 331 to 1615 ms (296 K, 3 T) according to the vendor. To assess signal stability and linearity, a series of ten tubes with varying concentration of gadolinium-based (Gd) contrast agent (Dotarem, Guerbet, France) was created in a range of 0 to 10 mM in a solution of 0.045 mM manganese chloride (native T1 ≈ 1.5 s, T2 ≈ 0.2 s). The samples fitted in the holder of the Eurospin II TO5 phantom. If the phantoms were not available locally, arrangements were made to ship them around.

The phantoms were positioned such that the samples were aligned with the main magnetic field. The DWI phantom was positioned with the center tube at the isocenter of the scanner to minimize the effect of gradient nonlinearities. The same receive coils were used as will be used in the patient MRI examinations. The institutional protocols were repeated three times within the same examination for assessment of short-term repeatability. For validation of the DCE–MRI protocol, the institutional protocol was scanned for the full five minutes to collect data on signal stability as well as signal linearity. If phase data could be saved in the protocol, an additional experiment was performed in which the tubes were scanned one by one for at least five dynamic scans at the same location in the phantom to avoid field inhomogeneity effects [19]. Temperature was measured before and after each experiment either from a tube filled with water placed next to the phantom for T2, T1, and DCE–MRI or by measuring the temperature of the ice water for DWI measurements.

Data analysis

Centralized data analysis was performed in institute I. All acquired data were uploaded in DICOM format. The analysis was done for regions of interest (ROIs), which were selected manually in the center of each tube and in the center slice of the phantom. T2, ADC, T1, and Gd concentration values were estimated from the mean signal intensities of the ROIs. Fitting details are described in Supplementary information. T2 and T1 values were corrected to a temperature of 296 K using the tables provided in the phantom manual. If a B1 map was available, T1 values and Gd concentration values were derived both with and without correction of the flip angles.

Statistical analysis

Accuracy was assessed with Bland-Altman analysis calculating the bias and the limits of agreement for the average of the repeated measurements per institute [20]. The true values were plotted on the *x*-axis with the differences on the *y*-axis (ΔqMRI). Kendall's tau test was used to test whether the differences were proportional to the magnitude of the true value [21]. If this was the case ($p < 0.05$), relative values were used to calculate bias by taking the ratio of the measurement and the true value [20]. As the bias is calculated for all ROIs of one institute together, we also calculated the percentage error (%error) for each ROI of each institute to be comparable with previous literature [12,22,23]:

$$\%\text{error} = 100 * \frac{|\Delta\text{qMRI}|}{\text{reference}} \qquad (1)$$

To estimate the short-term repeatability, the percentage repeatability coefficient (%RC) was estimated as 2.77 times the within-subject coefficient of variation (wCV) calculated from the

Table 2
Details of benchmark protocols.

Parameter	T2	DWI	T1 for DCE–MRI
Sequence type	Multi-echo spin echo	Single-shot EPI	Inversion recovery series
FOV (mm^3)	250 × 250 × 4	220 × 220 × 100	250 × 250 × 4
Voxel size (mm^3)	2 × 2 × 4	1.72 × 1.72 × 4	2 × 2 × 4
TR (ms)	2000	10,000	8000
TE (ms)	Max TE = 200*	Shortest	Not specified
TI (ms)	n.a.	n.a.	30, 50, 75, 100, 150, 200, 250, 300, 400, 500, 750, 1000, 1250, 1500, 2000, 4000
No. of averages	1	1	1
b-Values (s/mm^2)	n.a.	0, 500, 900, 2000	n.a.
Acceleration factor	no	2	No

* For Philips systems use minimal dTE that resulted in perfect (i.e. nontruncated) pulses.

Table 3
Overview of specified protocol parameters for institutional protocols. Not specified indicates that institutes were free to adjust this parameter to their needs; n.a. indicates that this parameter was not relevant for this protocol.

Parameter	T2	DWI	T1 for DCE–MRI	DCE–MRI
Minimal FOV (mm^3)	260 × 260 × 100	260 × 260 × 100	260 × 260 × 100	260 × 260 × 100
Imaging plane	Transversal	Transversal	Transversal	Transversal
Maximal slice thickness (mm)	5	5	5	5
TR (ms)	2000–5000	>2000	Not specified	≤7
TE (ms)	Max TE ≈ 200*	90	Not specified	Not specified
Number of echoes	≥5	n.a.	n.a.	n.a.
Flip angle	Not specified	Not specified	Not specified	≥20
Fat suppression	no	yes	Not specified	Not specified
b-Values (s/mm^2)	n.a.	At least b = 0, 200, 1000	n.a.	n.a.
Dynamic interval (s)	n.a.	n.a.	n.a.	≤5
Total acquisition time (min)	Not specified	Not specified	Not specified	≥5

* For Philips systems: dTE that resulted in perfect (i.e. nontruncated) pulses.

repeated measurements [11]. In a similar way, the inter-institutional reproducibility was estimated by the percentage reproducibility coefficient (%RDC) as 2.77 times wCV calculated from the first measurement of all institutes. In addition, the DWI data were analysed with the QIBAPhan software package (R1.4) to be sure that parameters were calculated in the same way and could directly be compared to the QIBA requirements [16]. The software package calculated voxel-wise ADC maps and extracted amongst others ADC bias, ADC error, and RC.

To assess signal stability of the DCE–MRI data, we calculated the coefficient of variation (CV) per phantom tube from the standard deviation of the signal intensities for all dynamic scans. From that, the mean CV per institute was calculated. To investigate the effect of B1 correction on the T1 values and Gd concentration values of DCE–MRI, the difference in %error with and without B1 correction was calculated.

Assessment of protocol updates

If modification of an institutional protocol was needed after the phantom measurements, additional experiments were performed to test the updated institutional protocol with respect to the already validated benchmark protocol of the initial phantom measurements. In case the calibration phantoms were no longer available at an institute, simpler phantoms were created locally. These phantoms were scanned with the benchmark protocol, the old institutional protocol and the updated institutional protocol.

Results

Short-term repeatability was assessed everywhere except for T2 mapping at institute B and for ADC mapping at institute J. The DCE–MRI stability measurements were not performed at institutes C and J. At this stage institute O has not yet been accredited for participation in the IQ-EMBRACE trial.

For T2 mapping, we observed an underestimation of the T2 value for both benchmark and institutional protocols for all institutes (Fig. 2A). As the differences increased for larger T2 values, ratios were used for bias estimation (Supplementary Fig. 1A). The bias in relation to the true values ranged between 0.90 and 0.98 for the benchmark and between 0.88 and 0.98 for the institutional protocols. The %error ranged between 4 and 10% for the benchmark protocol and between 4 and 12% for the institutional protocol. The median %RC for short-term repeatability was 1% (range 0–2%). The %RDC across institutes was 12%.

For ADC, the bias ranged between 0.002 and 0.040 × 10^{-3} mm^2/s for the benchmark protocol and between −0.007 and 0.029 × 10^{-3} mm^2/s for the institutional protocol (Fig. 2B). For the institutional protocol, the data from institute N were not used in the analysis as the temperature increased to 275 K at the time of these measurements. The %error of the center tube ranged between 0 and 3% for the benchmark and between 1–4% for the institutional protocol. Median %RC of the institutional protocol was 3% (range 1–13%), whereas the %RDC was 18%. Comparing the ADC estimates to the criteria of the QIBA DWI profile indicates that the bias criterion was met in all institutes for the benchmark protocol and in all but one for the institutional protocol (Supplementary information). The ADC error was higher than the QIBA requirements in twelve institutes for the benchmark protocol and in three for the institutional protocol, whereas the RC was higher in seven and three institutes, respectively.

For T1 mapping, as part of the validation of DCE–MRI, we observed that the T1 values estimated from the benchmark protocol were closer to the true values than those estimated with the institutional protocols (Fig. 2C). Similar to T2 values, a dependency was observed between the absolute differences and the magnitude of the T1 values. The range in relative bias in T1 values was smaller in the benchmark sequences (0.97–1.05) compared to the institutional sequence (0.39–1.29) (Supplementary Fig. 1B). The relative bias in institute E was about three times larger than the bias of the other institutes as a gradient echo sequence was used instead of a spoiled gradient echo sequence. The results of this sequence were not taken into account in further analysis. The %error ranged between 1 and 6% for the benchmark protocol and between 5 and 29% for the institutional protocol. The median %RC for short-term repeatability was 1% (range 1–10%). The %RDC across institutes was 33%. For six of seven institutes where B1 map correction was applied, the %error decreased with a median of 3% (range 3–8%), whereas in one institute the %error increased with 7%.

For the evaluation of the DCE–MRI protocol itself, we observed that the median CV as a measure for signal stability ranged between 0.2 and 1.5% across institutes. Furthermore, the assessment of signal linearity showed a vendor-specific non-linear relation for the magnitude data between measured and true Gd concentrations for all institutes (Fig. 3a). Up to concentrations of 2 mM, the relation was linear for Siemens and GE systems, whereas Philips systems showed an underestimation. For larger Gd concentrations the differences between institutes increased. B1 correction improved the %error for Philips (−0.5% to 1.6% at 2 mM) and Siemens systems (3.6% and 7.4% at 2 mM; 29% and 152% at 10 mM). On one GE system B1 correction led to negative Gd concentration values. For phase data, a linear relation for the whole range of concentration values was present in eight of ten institutes (Fig. 3b). The phase data of institutes F and H resulted in erroneous concentration values, most likely due to phase shift corrections performed as the data of each tube was collected in a separate acquisition.

The QA procedure allowed for changes in the protocol after the phantom measurements were performed. At institute E further

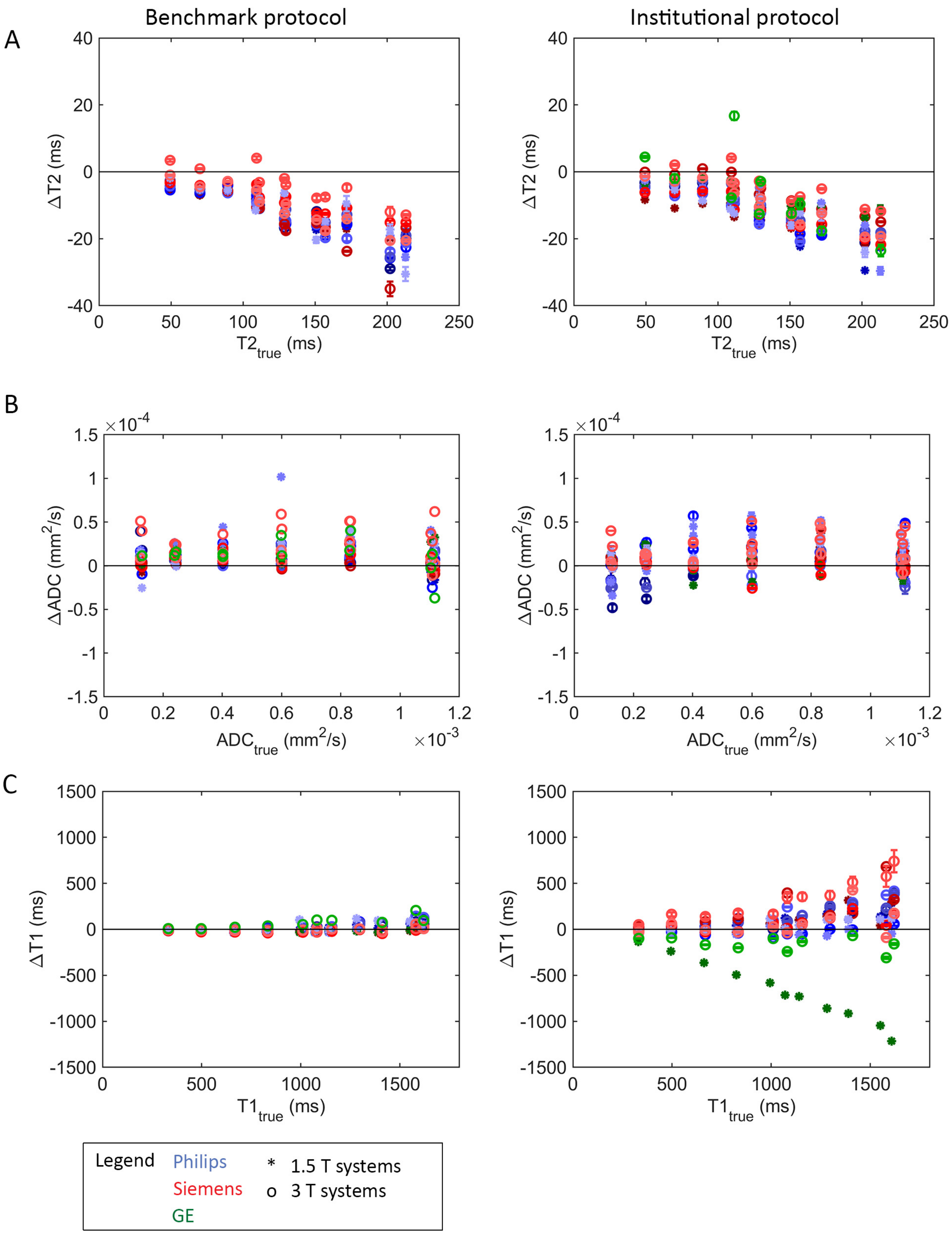

Fig. 2. Summary of the results of the benchmark (left column) and institutional protocol (right column). Absolute differences are shown between the measured values and the true values for T2 (A), ADC (B), and T1 (C). For the institutional T1 protocol, the data are shown without B1 correction. Error bars represent the standard deviation of the repeated measurements. The colors of the markers refer to different vendors, whereas the symbols refer to the field strength. The reason for the outlier results in the institutional T1 protocol (C) is that a gradient echo sequence was used and not a spoiled-gradient echo sequence. The protocol was modified later, which resulted in a lower bias.

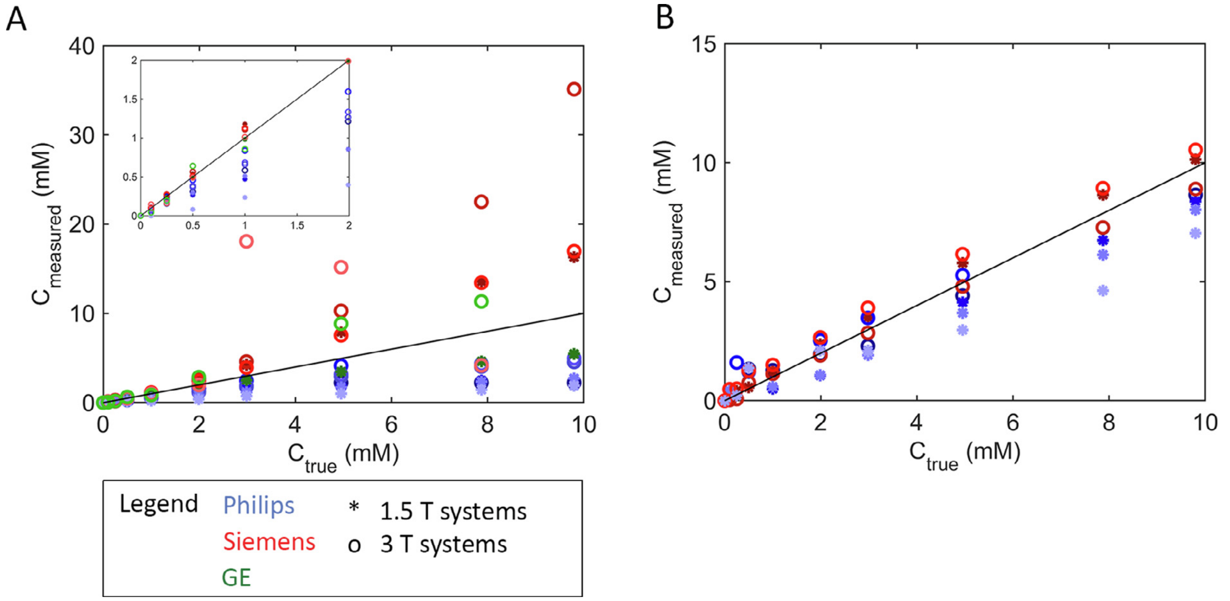

Fig. 3. The measured concentration values with the DCE protocols versus the true concentration values of the phantom samples. (A) Shows the results for the magnitude part of the data with the insert showing a zoomed image between 0 and 2 mM. The results without B1 correction are shown. (B) Shows the results for the phase part of the data. The colors of the markers refer to different vendors, whereas the symbols refer to the field strength. The black solid line represents the identity line.

optimization of the institutional T1 mapping protocol after the QA measurements reduced the bias to 0.73 with respect to the benchmark protocol. This was tested with a phantom consisting of four 30 mL tubes filled with water and contrast agent (Gadovist, Bayer AG, Germany) with estimated T1 values of 600, 1000, 1200, and 1600 ms. At institute N where the temperature increased during the DWI measurements, the benchmark DWI protocol was repeated at 296 K as well. As the bias of the benchmark protocol at 273 K was acceptable, the ADC values of the institutional protocol could be compared to the benchmark data at 275 K instead of the true values at 273 K showing a bias similar as the other institutes.

Discussion

In this study, we designed a QA framework for qMRI studies in the presence of a large variety of MR systems and institutional protocols. We applied this framework to assess the performance of T2 mapping, DWI, and DCE–MRI across fifteen institutes participating in the IQ-EMBRACE trial. Despite the freedom to choose the imaging method (e.g. inversion recovery, look-locker, or variable flip angle method for T1 mapping), many institutes opted for the same basic method. By including benchmark protocols in the QA framework, we enabled modification of the institutional protocols during the trial by comparing the updated protocol with the benchmark protocol.

For T2 mapping, we observed a consistent underestimation of the T2 values with the benchmark and institutional protocols compared to the values provided by the manufacturer, which could be explained by the presence of stimulated echoes due to imperfect refocusing pulses for example [24–26]. Interestingly, the bias and %error were similar between benchmark and institutional protocols suggesting that the larger variation in protocol settings that was present in the institutional protocols did not affect the accuracy of the T2 values. Short-term repeatability values were similar as reported in previous studies [22,27]. The %RDC of 12% was higher than reported in a previous study [27] employing MR systems from a single vendor using the same acquisition scheme. To our knowledge, T2 mapping data in cervical cancer has not been reported yet to get an indication whether the reproducibility will be sufficient to distinguish non-responders and responders.

For ADC, the bias with both the benchmark and the institutional protocols satisfied the QIBA DWI criteria [16]. However, for the benchmark protocol ADC error and RC exceeded the limits in most institutes. An explanation for these findings could be that we used the coil set-up for scanning patients with cervical cancer instead of the head coil as recommended in the phantom instructions, which could result in a lower SNR. This is illustrated by the results of the institutional protocols where larger voxel sizes and lower maximum *b*-values were used, resulting in more institutes passing the ADC error and RC requirements. The %RDC reported in this study was higher than reported by Malyarenko et al. for standardized DWI protocols [12]. However, in their calculation only the center tube was taken into account, whereas we included all samples in the %RDC calculation. Phantom positioning is most likely the main explanation for these differences. The center tube was always in the iso-center of the MRI, but the phantom was rotated differently meaning that the position of the other tubes varied between institutes. This results in extra variation explained by gradient nonlinearities. When repeating our analysis for the center tube only, the % RDC was 3.1% which is close to the reproducibility they reported for using torso coils. Some previously reported differences in ADC values between responders and non-responders [28,29] were smaller than the %RDC of 18% we found in this study, but larger differences have also been reported [30].

The results we found for both the benchmark and institutional baseline T1 mapping protocol are similar to those obtained in a recent multicenter phantom T1 study [23]. As the large uncertainty in the institutional T1 mapping data will introduce extra uncertainty in the DCE–MRI analysis, the use of a fixed T1 value could be considered instead, like in prostate data for example [31]. The stability of the DCE–MRI signal was considered good with a median CV below 1.5% which will have negligible impact on the pharmacokinetic parameters [32]. The results found for the estimated concentration values with magnitude and phase are in line with previously reported results [33,34]. The vendor-specific relation between measured and true concentration values for magnitude values suggests that this is related to the vendor implementation of the spoiled gradient echo sequence such as RF spoiling [35], for example. The large variations for higher concentration values indicate that arterial input functions derived from magnitude data only will be inaccurate, especially for the peak height. Instead using the combination of the magnitude and phase data, i.e. complex data, will result in more accurate arterial input functions [36,37]. However, not all institutes were able to save the phase data due to scanner software limitations, consequently a population or reference AIF will be used in the analysis of the patient data. In tumor tissue, usually lower concentrations of contrast agent are present (between 0 and 1 mM [38,39]). However, even for these concentrations values we observed substantial variation across institutes, which will impact the accuracy of the pharmacokinetic parameters as the conversion from signal intensity to concentration value is an essential step in the analysis. Errors in concentration values will directly propagate to the K^{trans} values estimated with the Tofts model. For example, a 50% underestimation of concentration value will result in a 50% underestimation of K^{trans}. k_{ep} is less sensitive for the variations in concentration values and might therefore, be a more reproducible parameter to be compared between institutes [40]. Several studies suggest that DCE–MRI is prognostic for cervical cancer [3,41–43]. However, difference in methods of analyzing the data make a direct comparison difficult and except for [44], no absolute K^{trans} values of responders and non-responders were reported.

While the results of the phantom measurements in this study are specific for the qMRI protocols optimized for the IQ-EMBRACE trial, the QA framework can be applied generally. For other tumor sites, qMRI protocols may need to be modified. Then, new phantom measurements are required to ensure consistent qMRI values. The results of the phantom measurements can be used for sample size calculations in clinical trial design [13]. In addition, as the reproducibility of T2 and ADC was good, an institutional bias is not to be expected in the patient data. As a result, the patient data that will be acquired in the IQ-EMBRACE trial can be used to determine thresholds or build prediction models to separate between responders and non-responders. Furthermore, the results of the phantom data for DCE–MRI and T1 mapping might explain an institutional bias in the patient data that will be collected and could potentially be used to apply institute-specific corrections to reduce the variations between institutes. In addition, the current data can be a starting point for the development of recommendations for quantitative MR protocols for cervical cancer, in line with the recommendations of QIBA for DWI and DCE–MRI protocols for brain, breast, prostate, liver, and head and neck [11]. For a large part the protocols for cervix can be based on prostate protocols. However, cervical tumors are larger than prostate tumors and therefore larger field-of-views are required. This puts constraints on other parameters to limit the acquisition time.

While the benchmark sequences would ideally be identical, in practice this is challenging. Even if protocol parameters are specified, this does not mean that the sequences are implemented identically by different vendors. Differences in the qMRI values of the benchmark protocols can be related to differences in sequence implementation as well as hardware differences, such as gradient systems and coil set-up. However, despite these challenges, the qMRI values of the benchmark sequences were consistent and reproducible across institutes. As a certain degree of freedom was allowed for the optimization of the institutional protocols, we could assess whether differences in qMRI values obtained with these institutional protocols were related to differences in protocol parameters or mainly due to more general system differences. Additional sources of variation, like gradient nonlinearities for ADC [12], were not considered in this study. For logistical reasons we were not able to use the same phantom copy for all measurements, which may have introduced unnecessary variation. Phantom measurements are only a part of the validation of qMRI protocols. Aspects like SNR, image artifacts and day-to-day variations in patients were not taken into account [10,45]. This needs to be investigated with healthy volunteers or in patient test–retest studies.

In conclusion, to meet the challenge of multicenter MRI biomarker studies, we designed and tested a QA framework with calibration phantoms to assess the variation in qMRI values. The framework allows updates of institutional protocols during a running trial by comparing them with the benchmark protocols. While allowing some variability in scan protocols, consistent ADC and T2 values were obtained. For DCE–MRI and T1 mapping with variable flip angle mapping this was more challenging. The results of the phantom measurements can be used for sample size calculations and to apply corrections to the acquired patient data to further reduce the variation.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: This work was funded by a grant from Varian Medical Systems received by Kari Tanderup and Uulke van der Heide. Relevant financial activities outside the submitted work are reported for the following authors: Uulke van der Heide receives research funding from Philips Healthcare and Elekta AB. Remi Nout receives research funding from Elekta AB, Varian Medical Systems, and Accuray. Cornelius van den Berg receives research funding from Philips Healthcare.

Acknowledgements

We thank the following people for their assistance with phantom measurements and/or protocol optimization: Robert Hudej, PhD (Dept. of Radiotherapy, Institute of Oncology Ljubljana), Michaela Daniel (Dept. of Radiation Oncology, Medical University of Vienna), Pascal A. Baltzer, MD (Dept. of Biomedical Imaging and Image-guided Therapy, Medical University of Vienna), Supriya Chopra, MD PhD (ACTREC, Tata Memorial Centre), Kelly McPhee, PhD and Warren Foltz, PhD (Radiation Medicine Program, Princess Margaret Cancer Center) and Djordje Kovacevic (Oslo University Hospital). We thank Iris Walraven, PhD (Dept. of Radiation Oncology, the Netherlands Cancer Institute) for consultation on statistical analysis. This work was funded by a research grant from Varian Medical Systems received by Kari Tanderup, PhD and Uulke van der Heide, PhD.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2020.09.013.

References

[1] Press RH, Shu HKG, Shim H, Mountz JM, Kurland BF, Wahl RL, et al. The use of quantitative imaging in radiation oncology: A quantitative imaging network (QIN) perspective. Int J Radiat Oncol Biol Phys 2018;102:1219–35. https://doi.org/10.1016/j.ijrobp.2018.06.023.

[2] Hall WA, Paulson ES, van der Heide UA, Fuller CD, Raaymakers BW, Lagendijk JJW, et al. The transformation of radiation oncology using real-time magnetic resonance guidance: A review. Eur J Cancer 2019;122:42–52. https://doi.org/10.1016/j.ejca.2019.07.021.

[3] Halle C, Andersen E, Lando M, Aarnes E-K, Hasvold G, Holden M, et al. Hypoxia-induced gene expression in chemoradioresistant cervical cancer revealed by dynamic contrast-enhanced MRI. Cancer Res 2012;72:5285–95. https://doi.org/10.1158/0008-5472.CAN-12-1085.

[4] Daniel M, Andrzejewski P, Sturdza A, Majercakova K, Baltzer P, Pinker K, et al. Impact of hybrid PET/MR technology on multiparametric imaging and

treatment response assessment of cervix cancer. Radiother Oncol 2017;125:420–5. https://doi.org/10.1016/j.radonc.2017.10.036.

[5] Akkus Yildirim B, Onal C, Erbay G, Cem Guler O, Karadeli E, Reyhan M, et al. Prognostic values of ADCmean and SUVmax of the primary tumour in cervical cancer patients treated with definitive chemoradiotherapy. J Obstet Gynaecol (Lahore) 2018;39:224–30. https://doi.org/10.1080/01443615.2018.1492528.

[6] Schurink NW, Lambregts DMJ, Beets-Tan RGH. Diffusion-weighted imaging in rectal cancer: current applications and future perspectives. Br J Radiol 2019;92:20180655. https://doi.org/10.1259/bjr.20180655.

[7] Noij DP, de Jong MC, Mulders LGM, Marcus JT, de Bree R, Lavini C, et al. Contrast-enhanced perfusion magnetic resonance imaging for head and neck squamous cell carcinoma: A systematic review. Oral Oncol 2015;51:124–38. https://doi.org/10.1016/j.oraloncology.2014.10.016.

[8] O'Connor JPB, Aboagy EO, Adams JE, Aerts HJWL, Barrington SF, Beer AJ, et al. Imaging biomarker roadmap for cancer studies. Nat Rev Clin Oncol 2016;14:169–86. https://doi.org/10.1038/nrclinonc.2016.162.

[9] DeSouza NM, Winfield JM, Waterton JC, Weller A, Papoutsaki MV, Doran SJ, et al. Implementing diffusion-weighted MRI for body imaging in prospective multicentre trials: current considerations and future perspectives. Eur Radiol 2018;28:1118–31. https://doi.org/10.1007/s00330-017-4972-z.

[10] Jafar MM. Diffusion-weighted magnetic resonance imaging in cancer: Reported apparent diffusion coefficients, in-vitro and in-vivo reproducibility. World J Radiol 2016;8:21–49. https://doi.org/10.4329/wjr.v8.i1.21.

[11] Shukla-Dave A, Obuchowski NA, Chenevert TL, Jambawalikar S, Schwartz LH, Malyarenko D, et al. Quantitative imaging biomarkers alliance (QIBA) recommendations for improved precision of DWI and DCE-MRI derived biomarkers in multicenter oncology trials. J Magn Reson Imaging 2019;49:e101–21. https://doi.org/10.1002/jmri.26518.

[12] Malyarenko D, Galbán CJ, Londy FJ, Meyer CR, Johnson TD, Rehemtulla A, et al. Multi-system repeatability and reproducibility of apparent diffusion coefficient measurement using an ice-water phantom. J Magn Reson Imaging 2013;37:1238–46. https://doi.org/10.1002/jmri.23825.

[13] Obuchowski NA, Mozley PD, Matthews D, Buckler A, Bullen J, Jackson E. Statistical considerations for planning clinical trials with quantitative imaging biomarkers. J Natl Cancer Inst 2019;111:19–26. https://doi.org/10.1093/jnci/djy194.

[14] Pötter R, Tanderup K, Kirisits C, de Leeuw A, Kirchheiner K, Nout R, et al. The EMBRACE II study: The outcome and prospect of two decades of evolution within the GEC-ESTRO GYN working group and the EMBRACE studies. Clin Transl Radiat Oncol 2018;9:48–60. https://doi.org/10.1016/j.ctro.2018.01.001.

[15] DCE MRI Biomarker Committee. Profile: DCE MRI Quantification, Quantitative Imaging Biomarkers Alliance. Version 1.6. QIBA, Dec 13, 2011. Available from: http://qibawiki.rsna.org/index.php/Profiles.

[16] DWI MR Biomarker Committee. Diffusion-Weighted Magnetic Resonance Imaging Profile, Quantitative Imaging Biomarkers Alliance. Version 2019-Feb-05. QIBA, Feb 05, 2019. Available from: http://qibawiki.rsna.org/index.php/Profiles.

[17] Stollberger R, Wach P. Imaging of the active B1 field in vivo. Magn Reson Med 1996;35:246–51. https://doi.org/10.1002/mrm.1910350217.

[18] Yarnykh VL. Actual flip-angle imaging in the pulsed steady state: A method for rapid three-dimensional mapping of the transmitted radiofrequency field. Magn Reson Med 2007;57:192–200. https://doi.org/10.1002/mrm.21120.

[19] Akbudak E, Norberg RE, Conturo TE. Contrast-agent phase effects: An experimental system for analysis of susceptibility, concentration, and bolus input function kinetics. Magn Reson Med 1997;38:990–1002. https://doi.org/10.1002/mrm.1910380619.

[20] Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999;8:135–60. https://doi.org/10.1191/096228099673819272.

[21] Bland JM, Altman DG. Statistics notes: Measurement error proportional to the mean. BMJ 1996;313:106. https://doi.org/10.1136/bmj.313.7049.106.

[22] Kooreman ES, van Houdt PJ, Nowee ME, van Pelt VWJ, Tijssen RHN, Paulson ES, et al. Feasibility and accuracy of quantitative imaging on a 1.5 T MR-linear accelerator. Radiother Oncol 2019;133:156–62. https://doi.org/10.1016/j.radonc.2019.01.011.

[23] Bane O, Hectors SJ, Wagner M, Arlinghaus LL, Aryal MP, Cao Y, et al. Accuracy, repeatability, and interplatform reproducibility of T1 quantification methods used for DCE-MRI: Results from a multicenter phantom study. Magn Reson Med 2018;79:2564–75. https://doi.org/10.1002/mrm.26903.

[24] Liney GP, Knowles AJ, Manton DJ, Turnbull LW, Blackband SJ, Horsman A. Comparison of conventional single echo and multi-echo sequences with a fast spin-echo sequence for quantitative T2 mapping: Application to the prostate. J Magn Reson Imaging 1996;6:603–7. https://doi.org/10.1002/jmri.1880060408.

[25] Poon CS, Henkelman RM. Practical T2 quantitation for clinical applications. J Magn Reson Imaging 1992;2:541–53. https://doi.org/10.1002/jmri.1880020512.

[26] McPhee KC, Wilman AH. Limitations of skipping echoes for exponential T 2 fitting. J Magn Reson Imaging 2018;48:1432–40. https://doi.org/10.1002/jmri.26052.

[27] van Houdt PJ, Agarwal HK, van Buuren LD, Heijmink SWTPJ, Haack S, van der Poel HG, et al. Performance of a fast and high-resolution multi-echo spin-echo sequence for prostate T2 mapping across multiple systems. Magn Reson Med 2018;79:1586–94. https://doi.org/10.1002/mrm.26816.

[28] Erbay G, Onal C, Karadeli E, Guler OC, Arica S, Koc Z. Predicting tumor recurrence in patients with cervical carcinoma treated with definitive chemoradiotherapy: Value of quantitative histogram analysis on diffusion-weighted MR images. Acta Radiol 2017;58:481–8. https://doi.org/10.1177/0284185116656492.

[29] Meng J, Zhu L, Zhu L, Xie L, Wang H, Song L, et al. Whole-lesion ADC histogram and texture analysis in predicting recurrence of cervical cancer treated with CCRT. Oncotarget 2017;8:92442–53. https://doi.org/10.18632/oncotarget.21374.

[30] Zhao B, Cao K, Li XT, Zhu HT, Sun YS. Whole lesion histogram analysis of apparent diffusion coefficients on MRI predicts disease-free survival in locally advanced squamous cell cervical cancer after radical chemo-radiotherapy. BMC Cancer 2019;19:1–7. https://doi.org/10.1186/s12885-019-6344-3.

[31] Daniel M, Polanec SH, Wengert G, Clauser P, Pinker K, Helbich TH, et al. Intra- and inter-observer variability in dependence of T1-time correction for common dynamic contrast enhanced MRI parameters in prostate cancer patients. Eur J Radiol 2019;116:27–33. https://doi.org/10.1016/j.ejrad.2019.04.015.

[32] Garpebring A, Brynolfsson P, Yu J, Wirestam R, Johansson A, Asklund T, et al. Uncertainty estimation in dynamic contrast-enhanced MRI. Magn Reson Med 2013;69:992–1002. https://doi.org/10.1002/mrm.24328.

[33] Foltz W, Driscoll B, Lee SL, Nayak K, Nallapareddy N, Fatemi A, et al. Phantom validation of DCE-MRI magnitude and phase-based vascular input function measurements. Tomography 2019;5:77–89. https://doi.org/10.18383/j.tom.2019.00001.

[34] Korporaal JG, van den Berg CAT, van Osch MJP, Groenendaal G, van Vulpen M, van der Heide UA. Phase-based arterial input function measurements in the femoral arteries for quantification of dynamic contrast-enhanced (DCE) MRI and comparison with DCE-CT. Magn Reson Med 2011;66:1267–74. https://doi.org/10.1002/mrm.22905.

[35] Yarnykh VL. Optimal radiofrequency and gradient spoiling for improved accuracy of T1 and B1 measurements using fast steady-state techniques. Magn Reson Med 2010;63:1610–26. https://doi.org/10.1002/mrm.22394.

[36] Simonis FFJ, Sbrizzi A, Beld E, Lagendijk JJW, van den Berg CAT. Improving the arterial input function in dynamic contrast enhanced MRI by fitting the signal in the complex plane. Magn Reson Med 2016;76:1236–45. https://doi.org/10.1002/mrm.26023.

[37] Klawer EME, van Houdt PJ, Simonis FFJ, van den Berg CAT, Pos FJ, Heijmink SWTPJ, et al. Improved repeatability of dynamic contrast-enhanced MRI using the complex MRI signal to derive arterial input functions: a test-retest study in prostate cancer patients. Magn Reson Med 2019;81:3358–69. https://doi.org/10.1002/mrm.27646.

[38] Donaldson SB, West CML, Davidson SE, Carrington BM, Hutchison G, Jones AP, et al. A comparison of tracer kinetic models for T1-weighted dynamic contrast-enhanced MRI: application in carcinoma of the cervix. Magn Reson Med 2010;63:691–700. https://doi.org/10.1002/mrm.22217.

[39] Kallehauge JF, Sourbron S, Irving B, Tanderup K, Schnabel JA, Chappell MA. Comparison of linear and nonlinear implementation of the compartmental tissue uptake model for dynamic contrast-enhanced MRI. Magn Reson Med 2017;77:2414–23. https://doi.org/10.1002/mrm.26324.

[40] Li X, Cai Y, Moloney B, Chen Y, Huang W, Woods M, et al. Relative sensitivities of DCE-MRI pharmacokinetic parameters to arterial input function (AIF) scaling. J Magn Reson 2016;269:104–12. https://doi.org/10.1016/j.jmr.2016.05.018.

[41] Dickie BR, Rose CJ, Kershaw LE, Withey SB, Carrington BM, Davidson SE, et al. The prognostic value of dynamic contrast-enhanced MRI contrast agent transfer constant Ktrans in cervical cancer is explained by plasma flow rather than vessel permeability. Br J Cancer 2017;116:1436–43. https://doi.org/10.1038/bjc.2017.121.

[42] Andersen EKF, Hole KH, Lund KV, Sundfør K, Kristensen GB, Lyng H, et al. Pharmacokinetic parameters derived from dynamic contrast enhanced MRI of cervical cancers predict chemoradiotherapy outcome. Radiother Oncol 2013;107:117–22. https://doi.org/10.1016/j.radonc.2012.11.007.

[43] Lund KV, Simonsen TG, Kristensen GB, Rofstad EK. DCE-MRI of locally-advanced carcinoma of the uterine cervix: Tofts analysis versus non-model-based analyses. Radiat Oncol 2020;15:79. https://doi.org/10.1186/s13014-020-01526-2.

[44] Zhang Z, Wang Z, Zhao R. Dynamic contrast-enhanced magnetic resonance imaging of advanced cervical carcinoma: the advantage of perfusion parameters from the peripheral region in predicting the early response to radiotherapy. Int J Gynecol Cancer 2018;28:1342–9. https://doi.org/10.1097/IGC.0000000000001308.

[45] Winfield JM, Collins DJ, Priest AN, Quest RA, Glover A, Hunter S, et al. A framework for optimization of diffusion-weighted MRI protocols for large field-of-view abdominal-pelvic imaging in multicenter studies. Med Phys 2016;43:95–110. https://doi.org/10.1118/1.4937789.