Daily quantitative MRI for radiotherapy response monitoring

Kooreman, E.S.

DISCUSSION

Hybrid MR-linac systems create an opportunity for daily non-invasive treatment response monitoring using qMRI without an increase in patient burden. However, from an MRI perspective, these machines are adjusted to allow the treatment beam to pass through. Because these adjustments might influence the qMRI measurements compared to diagnostic systems, their capabilities in terms of qMRI need to be evaluated.

As mentioned in the introduction, the imaging biomarker roadmap presented by O'Connor *et al.* identifies three tracks that are followed in parallel by a biomarker on its way from inception to clinical adoption [12]. When a new machine is introduced, like the Unity MR-linac in our hospital in 2016, their impact on the acquisition of QIBs should be validated. Accuracy, repeatability, and reproducibility are influenced by the machine that the QIB is measured on and these measures need to be determined for each QIB on a new machine.

PHANTOM MEASUREMENTS

The main value of phantom measurements in QIB studies is that they allow the assessment of the accuracy of QIBs, which is influenced the specific machine, acquisition settings and data analysis. Also, repeated phantom measurements are useful for monitoring the longitudinal stability of the system [155]. This can be done directly in qMRI phantoms, but other phantoms are also used to measure more generic characteristics of a system, such as the B_0 and B_1 field [74,156,157]. Monitoring these over time could prevent changes in these fields to unknowingly influence longitudinal qMRI measurements. For instance, Subashi *et al.* detected a deterioration of the SNR of their system which they were able to identify by consistently performing QA measurements in phantoms [156]. They identified table electronics as the source of additional noise, and after replacing these electronic components, the SNR was improved. In the same study, they assessed the stability of T_1, T_2, and ADC mapping by repeatedly measuring qMRI phantoms over the course of four weeks. They found excellent repeatability with low %CV, consistent with our findings from Chapter 2.

Phantom measurements can also be used to compare acquisition sequences. In a recent study by McDonald *et al.* three DWI sequences were compared with the help of a diffusion phantom, and performance of the MR-linac was compared to the performance of a 1.5 T diagnostic system [158]. The sequences they tested included non-EPI sequences, which are important for tumor sites with a high variation in the magnetic susceptibility such as the head and neck, where transitions between air and human tissue cause large deformations in images acquired with EPI sequences.

In the studies described in Chapter 2, 3, and 6, phantoms measurements were essential. In Chapter 2, they are used as a tool for the initial validation of qMRI capabilities on the Unity MR-linac for a selection of sequences. This allowed us to demonstrate the feasibility of qMRI measurements on the Unity MR-linac, showing accurate results. For patient studies, similar phantom measurements should be performed for each specific sequence or modification to a sequence, to reassess the accuracy, repeatability and reproducibility [2]. In the study presented in Chapter 3, phantoms were used for further specific characterization of the Unity MR-linac with respect to DWI. By using a large phantom filled with a homogeneous liquid, we were able to demonstrate a large spatial variation in the ADC value when measuring away from the iso-center. Another use of a phantom is described in Chapter 6, where we implemented an acquisition sequence to measure $T_{1\rho}$ relaxation times. By producing a phantom with different $T_{1\rho}$ relaxation times, we could compare our measurements to previously published results, validating that we were able to accurately measure $T_{1\rho}$.

TREATMENT RESPONSE MONITORING ON THE UNITY MR-LINAC

As the MR-linac enables the acquisition of daily qMRI without an increase in patient burden, this opens opportunities to measure changes with a much higher temporal resolution than before. Previously, studies assessing QIBs for radiation treatment response purposes typically measure at three timepoints: before treatment, once during treatment, and after treatment [89]. A QIB has the most value in terms of treatment response monitoring when it can predict treatment outcome before, or early during treatment. Potentially, this would enable treatment adaptation, or in a different scenario the QIB could serve as a biological target for stopping the treatment [90]. For this, QIBs that change over time because of treatment or disease progression are valuable. For instance, a recent meta-analysis showed no predictive value of pre-treatment tumor ADC values for treatment outcome in cervical cancer, but a meta-analysis of the early change in the ADC values during treatment showed potential as a predictive QIB [159,160]. This thesis contains studies that investigated if changes were measurable during treatment for IVIM and DCE parameters in prostate cancer patients (Chapters 4 and 5), and $T_{1\rho}$ in rectal cancer patients (Chapter 6). Chapter 4 contains encouraging results, where no change was measured in the D values of the healthy prostate, but D values increased in the tumor. However, while results from all these studies indicate that changes are measurable, this is on the group level. To move further towards personalized treatment, it is important to reduce the noise in the measured QIBs, either by improvements of acquisition sequences, or in post-processing.

TECHNICAL CHALLENGES

Signal to noise ratio

One of the most important aspects that influence the ability to measure changes over time is the SNR of the system, which influences the random variation of qMRI parameters. In general, the visual quality of qMRI maps is lower than qualitative maps, because multiple images need to be acquired in a short time. To compensate, this usually necessitates larger voxels, producing unclear images. Additionally, motion of the patient between multiple images complicates fitting of the quantitative models especially when done in a voxel-by-voxel manner to create quantitative maps. Therefore, often some kind of image registration is required before fitting the model. The Unity MR-linac differs substantially from diagnostic MRI systems, and as shown in Chapters 2-5 the SNR is an issue that keeps resurfacing.

In Chapter 2, we were unable to meet all the requirements of the QIBA tests for the DWI phantom, specifically requirements related to SNR [70]. When adjusting the sequence to enable higher SNR using larger voxels and a restricted maximum b-value, the requirements were met. In Chapter 3, considerations for the recommendations were also based largely on SNR limitations. Here, mainly the effect of the software-limited performance of the gradient system, which causes an increased TE in our DWI sequence underlies the recommendations regarding limiting the maximum b-value.

As described in the introduction, the receive coil consists of an 8-channel phased array coil, which is rigid. For certain tumor sites, such as head and neck, it is difficult to minimize the distance between the coil and the target. Improvements of the receive coils are challenging because the treatment beam imposes additional constraints to the design. Recently, Zijlema *et al.* developed a 32-channel flexible receive coil using high-impedance coil loops [161]. They demonstrated an improved SNR performance over the commercially available receive coil, especially with accelerated imaging.

For DWI, the gradient performance of the system is an important factor. The maximum gradient strength and slew rate of the Unity MR-linac system are restricted to 15 mT/m and 65 T/m/s, respectively. This is low compared to the diagnostic system it is based on, which has a maximum gradient strength of 40 mT/m and a maximum slew rate of 200 T/m/s. For DWI this means that to achieve the same b-value, the diffusion sensitizing gradients need to be extended, increasing the TE and reducing the SNR. Therefore, we recommend limiting the maximum b-value to 500 s/mm^2 for ADC mapping on the Unity system (Chapter 3).

The SNR of the system influences the repeatability, which is important for longitudinal monitoring of patients such as is the case for treatment response monitoring. In Chapter

4, the RC of IVIM parameters was determined, however it should be noted that the patients were not scanned twice before treatment in a test-retest setting, but rather the values of the first and second fraction of treatment were used. The RCs were high for the perfusion parameters of the IVIM model, meaning that the repeatability of these parameters is poor. This is in accordance with other studies, which also showed very high RCs for f and especially for D*. In a repeatability study on a 3 T system, Sun *et al.* reported a short-term %RC of 88% for D* in the prostate [162]. In a comparison between fit algorithms, Gurney-Champion *et al.* found %RCs for D* of 67% for the best performing algorithm up to 442% of the worst performing one in pancreatic tumors [163].

One of the reasons for the low repeatability could be that the IVIM sequence used in Chapter 4 was not optimized yet for measuring the perfusion parameters on the MR-linac. Our sequence contained only one specific point to measure D* (the b-value of 30 s/mm^2), while studies looking to optimize the b-value distribution for IVIM acquisitions typically include more b-values [164,165]. By measuring only four b-values, the calculation of IVIM parameters was straight forward and fast as no optimizer was needed. Currently, efforts are undertaken by Wetscherek *et al.* to specifically optimize IVIM sequences for the Unity MR-linac for multiple target sites [166]. This abstract takes a data-driven approach, where Monte-Carlo simulations are used to find a set of b-values and averages that minimize the average relative error in the D, f, and D* parameter maps. The abstract provides optimal acquisition parameters for a measurement time of 5 minutes for head and neck, brain, prostate, and rectum tumors. By using ROI averages for the fit, a better coefficient of variation is achieved than for voxel wise fits.

Novel deep learning based algorithms are promising techniques that can reduce the RC in quantitative parameter maps. In a study by Kaandorp *et al.*, a neural network outperformed conventional least squares based and Bayesian algorithms in terms of accuracy and repeatability of the IVIM parameters in a cohort of pancreatic cancer patients, although accuracy was only assessed on simulated data [167].

Noisy parameter maps also played a role in Chapter 5. While we did find some statistically significant longitudinal correlations between IVIM and DCE parameters, these correlations were low and mainly present in the transition zone, the largest ROI. The correlations could be reduced by the noisiness of the parameter maps on the patient level, which has the least influence on the largest ROI as median values were used. Future efforts should include improving the repeatability of both IVIM and DCE on the Unity MR-linac. The improvements mentioned above, such as a better receive coil, improved fitting methods, and optimized acquisition parameters would also be beneficial for both techniques.

Spatial inhomogeneity

An aspect of ADC mapping on the Unity system that deserves further attention is the spatial dependence of the ADC value. The current recommendation from Chapter 3 includes the advice to be careful when scanning away from the iso-center in the xy-plane. Within a radius of 7 cm the variation of the ADC that we found was < 5% of the mean value in the iso-center, while outside of that the ADC value can steeply increase up to 600% of the value in the iso-center. This might be difficult to adhere to in practice for certain tumor positions, for instance in liver cancer patients. As the pattern seems consistent over multiple measurements on multiple machines, longitudinal measurements might still be useful with careful patient positioning. We hypothesize that this spatial variation is caused by eddy currents which are increased in the Unity MR-linac because of the split gradient coil. The influence of eddy currents on DWI using EPI sequences and correction methods have been proposed before [168–170]. However, these techniques require the acquisition of additional scans, advanced sequence design, or the use of field probes. As eddy current properties are system specific, such measurements would have to be performed on each system separately. Additionally, extensive knowledge about reconstruction algorithms is needed to implement these corrections.

TUMOR SITES

Clinically, our results suggest that there is a limited value for prostate cancer patients in terms of treatment response monitoring using qMRI, because the measured changes during treatment are low compared to the RCs. The largest change in D from our mixed effects model over 20 fractions on the group level was found for patients with a high ISUP score and constituted an increase of about 14 % compared to the mean value on the first fraction. This corresponds to the 14 % increase Foltz *et al.* found in the tumor after six weeks of treatment [99], however it is well below our estimate of the RC for the D, which was about 36 % of the mean value from the tumors on the first fraction. Another factor limiting the value in the case of prostate cancer, is the time frame in which the QIBs are measured. Currently, most low- and intermediate-risk prostate cancer patients in our clinic are treated with a schedule of 5 × 7.25 Gy and trials investigating ultra-hypofractionation with only a single fraction are currently recruiting [171]. This is too short for treatment adaption based on QIBs although monitoring can also take place after treatment.

Prostate cancer patients were selected as first subjects for qMRI on the MR-linac for multiple reasons. Multiparametric protocols, including DWI and DCE are well established [172,173], the organ is positioned in the center of the body and close to the iso-center of the system, the movement is relatively small, and at our center a large

number of these patients were eligible for treatment on the MR-linac. This allows for the assessment of important aspects of QIBs that can reveal challenges also applicable to other tumor sites. For instance, the repeatability we found can be expected to be worse in other tumor sites were e.g. movement due to breathing or bowel motion, or tumor regression can complicate things.

Other groups have investigated changes in QIBs during radiotherapy in different organ sites. Yang *et al.* measured three sarcoma and three head and neck cancer patients on a 0.35 T MRI integrated with three cobalt-60 heads [17,18]. They analyzed their patients individually, and report varying results. Most notably, they found a drop of 33 % in the ADC value of a head and neck cancer patient, who was later shown by biopsy to have a necrotic core. On the Unity MR-linac, Thorwarth *et al.* showed preliminary results in a single head and neck cancer patient, where they found a doubling of the ADC value during treatment (from about 0.75 to 1.5 × 10^{-3} mm^2/s) [174]. Such large changes on an individual basis are encouraging results for treatment response potential in these tumor sites.

OTHER QUANTITATIVE TECHNIQUES

Chan *et al.* showed the feasibility of chemical exchange saturation transfer (CEST) MRI on the Unity MR-linac [175], validating results in a phantom, and measuring differences between high- and low-grade tumors of central nervous system (CNS) cancer patients. They demonstrated changes in CEST parameters on the cohort level in 54 patients, but also in individual patients. The same group assessed DWI in CNS in 24 patients on the MR-linac, and compared results to measurements of 20 patients (six patients were scanned on both systems) on a 1.5 T diagnostic system [123]. They found a negative bias in ADC values from the MR-linac, which increased with an increasing ADC value. In the 20 patients scanned on the MR-linac, they were able to identify individual patients based on an increase, decrease, both, or no change of ADC values in the GTV. The repeatability of the MR-linac with the 8-channel receive array they found was similar to a diagnostic 1.5 T system using a 16-channel birdcage coil, and better than we found in the prostate (Chapter 4). For white matter, the wSD was 0.0097 × 10^{-3} mm^2/s compared to 0.033 × 10^{-3} mm^2/s in the prostate [123]. Having a better repeatability in CNS cancer patients is encouraging for QIBs for treatment response monitoring in this group of patients.

Another interesting technique called MR fingerprinting [176], which can produce T_1 and T_2 maps in an extremely short amount of time (about 5 seconds for a single slice) has been shown to be feasible on the MRIdian system and the Unity system [177–179].

BIOLOGICAL LINK

Biological links for the ADC, DCE, and IVIM parameters are relatively well established [23]. The ADC corresponds to cell density and DCE and IVIM parameters correspond to perfusion characteristics. Currently, a link between the $T_{1\rho}$ relaxation time and radiation induced changes is not clear. $T_{1\rho}$ has been investigated in relation to measuring damage in OARs in head and neck cancer, specifically in the parotid glands which are sensitive to radiation [146]. In this study they found an increase in $T_{1\rho}$ values of the parotids during radiation treatment. Citing animal studies, the authors hypothesized that an early build-up of collagen, eventually leading to fibrosis could be the underlying biological mechanism. Fibrosis is also common in rectal cancer, and differentiating between fibrosis and residual tumor using conventional MRI techniques is difficult [153]. It is possible that the early onset of the formation of fibrosis causes an increase in $T_{1\rho}$. Establishing such a biological link would be an important next step in the development of $T_{1\rho}$ as a QIB.

DATA ANALYSIS

A challenging aspect of the daily acquisition of qMRI scans is the large amount of data that is generated for each patient. For instance, 43 patients were included in the study described in Chapter 3, who received 20 fractions of radiotherapy. These 43 patients would get 860 scanning sessions in total, with multiple scans per session. In our workflow, these scans were all registered to the first fraction, and delineations for each patient on the first fraction were needed. The IVIM scans consisted of 4 b-values, and scans were checked for motion between b-values. This is feasible to do in a research setting but would require substantial additional work on the daily in the clinic. A proper infrastructure is needed to handle this amount of data, and automation tools should be developed to lower the workload and improve the reproducibility. Auto-registration, auto-segmentation, and automated fitting procedures should be developed before the use of daily QIB informed therapy can become reality.

The statistical analysis of longitudinal imaging data should be considered carefully, as multiple data points from the same patient are correlated. Depending on the research question, proper statistical tests should be selected to account for this, such as the mixed effects models used in Chapter 4 or the repeated measures correlation in Chapter 5 of this thesis. As stated in Simpson's paradox, ignoring clustering when determining correlations can potentially lead to correlation coefficients with an opposite sign as when you take the clustering into account [136].

CLINICAL IMPLEMENTATION OF QIBs

The adaptation of qMRI and QIBs into the clinic is a long and slow process with many challenges, and groups such as QIBA, QIN, AAPM, and EIBALL have been formed with the goal to move QIB research further towards clinical adoption by specifically focusing on education, standardization of acquisition and software, and improvement of accuracy, repeatability, and reproducibility [12,19,62,63,89,180–183]. The MR-linac provides an opportunity to simplify some of these difficulties by providing an identical platform to multiple centers, thereby allowing sharing of MRI protocols and therefore easier pooling of patients.