Published in final edited form as:

Eur J Radiol. 2014 March ; 83(3): 487–496. doi:10.1016/j.ejrad.2013.11.016.

Interobserver agreement of semi-automated and manual measurements of functional MRI metrics of treatment response in hepatocellular carcinoma

David Bonekamp^a, Susanne Bonekamp^a, Vivek Gowdra Halappa^a, Jean-Francois H. Geschwind^a, John Eng^a, Celia Pamela Corona-Villalobos^a, Timothy M. Pawlik^{b,c}, and Ihab R. Kamel^{a,*}

^aThe Johns Hopkins School of Medicine, Department of Radiology, Baltimore, MD, United States

^bThe Sidney Kimmel Comprehensive Cancer Center, The Johns Hopkins University School of Medicine, Baltimore, MD, United States

^cThe Johns Hopkins School of Medicine, Department of Surgery, Oncology, Baltimore, MD, United States

Abstract

Purpose—To assess the interobserver agreement in 50 patients with hepatocellular carcinoma (HCC) before and 1 month after intra-arterial therapy (IAT) using two semi-automated methods and a manual approach for the following functional, volumetric and morphologic parameters: (1) apparent diffusion coefficient (ADC), (2) arterial phase enhancement (AE), (3) portal venous phase enhancement (VE), (4) tumor volume, and assessment according to (5) the Response Evaluation Criteria in Solid Tumors (RECIST), and (6) the European Association for the Study of the Liver (EASL).

Materials and methods—This HIPAA-compliant retrospective study had institutional review board approval. The requirement for patient informed consent was waived. Tumor ADC, AE, VE, volume, RECIST, and EASL in 50 index lesions was measured by three observers. Interobserver reproducibility was evaluated using intraclass correlation coefficients (ICC). $P < 0.05$ was considered to indicate a significant difference.

Results—Semi-automated volumetric measurements of functional parameters (ADC, AE, and VE) before and after IAT as well as change in tumor ADC, AE, or VE had better interobserver agreement (ICC = 0.830–0.974) compared with manual ROI-based axial measurements (ICC = 0.157–0.799). Semi-automated measurements of tumor volume and size in the axial plane before and after IAT had better interobserver agreement (ICC = 0.854–0.996) compared with manual size measurements (ICC = 0.543–0.596), and interobserver agreement for change in tumor RECIST

© 2013 Elsevier Ireland Ltd. All rights reserved.

*Corresponding author at: The Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins Medical Institutions, 600 North Wolfe Street, MRI 143, Baltimore, MD 21287, United States. Tel.: +1 410 955 4567; fax: +1 410 955 9799. ikamel@jhmi.edu (I.R. Kamel).

Conflicts of interest

Dr. Ihab Kamel is receiving grant support for his research (unrelated to the current article) from Siemens Medical and Bayer Healthcare.

size was also higher using semi-automated measurements (ICC = 0.655) compared with manual measurements (ICC = 0.169). EASL measurements of tumor enhancement in the axial plane before and after IAT (ICC = 0.758–0.809), and changes in EASL after IAT (ICC = 0.653) had good interobserver agreement.

Conclusion—Semi-automated measurements of functional changes assessed by ADC and VE based on whole-lesion segmentation demonstrated better reproducibility than ROI-based axial measurements, or RECIST or EASL measurements.

Keywords

Response assessment; Reproducibility; Observer agreement; Functional magnetic resonance imaging; Cancer treatment response

1. Introduction

Evaluation of tumor response to locoregional treatments such as intra-arterial therapy (IAT) has traditionally been made according to response criteria based on morphologic imaging assessments such as those proposed by the Response Evaluation Criteria in Solid Tumors (RECIST) [1]. However, size assessments have significant limitations including poor measurement reproducibility and the fact that tumor size alone offers only a limited assessment of response, as cell death is not in all cases accompanied by a decrease in volume [2]. Furthermore, with the increasing clinical use of antiangiogenic and biologic targeted agents, including tyrosine kinase inhibitors such as sorafenib, response assessment needs to take changes at a cellular level into account. Ideally, response assessment should help predict the likelihood of therapeutic success and provide guidance for follow-up treatment as early as possible. The European Association for the Study of the Liver (EASL) suggested a treatment response assessment based on tumor perfusion, a method recently adopted by an amendment to RECIST [1,3]. Unfortunately, there are also drawbacks that may limit EASL criteria, namely tumor rim-enhancement, enhancing granulation tissue after treatment, and central tumor necrosis in larger lesions [2]. In addition, EASL measurements are performed in the axial plane, not taking into account changes in enhancement that occur throughout the entire tumor. In order to overcome the limitations of the current response criteria, volumetric, whole-tumor, functional imaging techniques, including diffusion-weighted MRI and contrast enhanced MRI, that depict vascular and cellular processes within the entire tumor volume are being investigated for the monitoring of response to IAT.

Prior studies have evaluated the correlation between functional, volumetric MRI response to treatment and patient survival in patients with primary and metastatic liver cancer [4–7]. One study evaluated the change in ADC as well as arterial (AE) and portal venous phase enhancement (VE) on a voxel-by-voxel basis and determined that a change of ADC of greater than 160 μ m/s or a change of 10% or more in AE and VE constituted a real change induced by treatment and not a measurement attributable to background fluctuation [4]. Another study described the selection of an optimal threshold for increase in mean ADC (25%) and decrease in mean VE (65%) for the prediction of favorable response to therapy in patients with HCC [8]. Another study in patients with primary liver cancer showed high

correlation between responders according to these functional, volumetric MRI criteria and overall patient survival after the initial session of IAT [9].

Knowing the reproducibility of a potential biomarker of treatment response is critical because it creates confidence that a change in the measured value is real and reflects treatment or disease progression induced changes. Two recent studies have found a good reproducibility of ADC in liver cancer with intraclass correlation coefficients (ICC) ranging from 0.77 to 0.83 and a coefficient of reproducibility of $0.20 \times 10^{-3} \text{ mm}^2/\text{s}$ [10,11]. Measurement reproducibility depends on variations of the specific hardware or instrument used, the imaging technique (i.e. free-breathing versus breath-hold acquisition, choice of b values for DWI), tumor biology, and, especially in a clinical setting, observer variability. Our aim was to establish the interobserver agreement of functional, volumetric MRI for the assessment of response to treatment in patients with liver cancer undergoing IAT. To be able to contrast the functional, volumetric MRI parameters with currently used methods we assessed the interobserver agreement of three methods (two semi-automated, volumetric methods and one manual, ROI-based approach) in the evaluation of the following parameters:

1. Functional: enhancement in the arterial (AE) and portal venous phase (VE), and ADC.
2. Morphologic: tumor volume, RECIST, EASL.

We hypothesize that the selection of the axial slice used to obtain manual measurements of tumor size or functional parameters will differ between observers and a comprehensive, semi-automated, volumetric tumor assessment would provide more reproducible results.

2. Methods

This Health Insurance Portability and Accountability Act - compliant study was approved by the institutional ethics research board and the requirement for patient informed consent was waived. The software and hardware used in this study were provided free of charge by Siemens Corporate Research (Princeton, NJ) as part of a research agreement with our institution. The authors had full control of the data and the information submitted for publication.

2.1. Study cohort

A database search identified 723 patients with newly diagnosed HCC who had undergone IAT (Transarterial Chemo Embolization [TACE] or drug-eluting bead TACE [DEB-TACE]) as well as pre- and post-treatment MRI on the same MRI scanner between October 2005 and February 2011. The diagnosis of HCC was based on imaging criteria, patient history (i.e. chronic liver disease), and AFP levels. Patient selection was performed in our departmental electronic radiology report database. Exclusion criteria were as follows: (a) 51 patients had received systemic therapy (Sorafenib, Bevacizumab, and Doxorubicin), (b) 21 patients underwent radioembolization, (c) 98 patients had undergone MR imaging outside our institution, (d) 286 patients had undergone pre- or post-IAT imaging on a different scanner within our institution, (e) 57 patients had no follow-up MRI before retreatment, (f) 26

patients had not undergone diffusion-weighted imaging, (g) 33 patients had not received any contrast agent due to preexisting conditions, and (h) in 8 patients image artifacts led to exclusion of the data. Among the remaining 143 patients we selected 50 patients for this study using random number generation. A sample size of $N = 50$ was selected to reflect the power of typical studies in the literature [4,12]. All analyzed lesions were larger than 2 cm in diameter.

2.2. MR image acquisition

All patients included in this study underwent a standardized imaging protocol. MR imaging was performed on a 1.5-T MRI scanner (Siemens Magnetom Avanto) using a phased array torso coil. The protocol consisted of T2-weighted turbo spin-echo images (matrix size, 256×256 ; slice thickness, 8 mm; interslice gap, 2 mm; repetition time/echo time, 4500/92) and breath-hold diffusion-weighted echoplanar images (matrix, 128×128 ; slice thickness, 8 mm; interslice gap, 2 mm; b -value, 0 s/mm^2 , 750 s/mm^2 ; repetition time, 3000 ms; echo time, 69 ms). Breath-hold unenhanced and contrast-enhanced (0.1 mmol/kg intravenous gadopentetate; Magnevist; Bayer, Wayne, NJ) T1-weighted three-dimensional fat suppressed spoiled gradient-echo images (field of view, 320–400 mm; matrix, 192×160 ; slice thickness, 2.5 mm; repetition time, 5.77 ms; echo time, 2.77 ms; flip angle, 10°) in the hepatic arterial phase (20 s), portal venous phase (70 s), and delayed phase (3 min) were also obtained.

2.3. MR image analysis

Three readers (D.B. radiologist with 10 years of experience with image data processing and analysis, S.B. research faculty with 8 years of experience, V.G.H. post doctoral fellow with 2 years of experience) analyzed the pre- and post-IAT MRI data of the largest, treated index lesion in 50 patients with HCC using three different methods. Methods 1 and 2 utilized the same semi-automated, proprietary software (MROncotreat, Siemens Corporate Research, NJ) while measurements for Method 3 were performed manually on our institutional DICOM system (Advanced Visualization, Emageon and Enterprise Visual Medical System, Merge Healthcare, Chicago, Illinois). To avoid recall the order of the studies was scrambled between methods and measurements using the next method were performed after a break of at least 4 weeks. Fig. 1 gives an overview of the measurements obtained with each method. All readers analyzed the cases independently.

Since the purpose of all acquired MRI measurements was to determine the response to IAT we also calculated change in mean pre- and post IAT as percent change using the following formula:

$$\text{Percent Change} = [(\text{Mean Post-IAT value} - \text{Mean Pre-IAT value}) / \text{Mean Pre-IAT value}] * 100$$

2.4. Method 1 - co-registered volumetric analysis

Digital Imaging and Communications in Medicine data was imported into the software (MROncotreat) and three-dimensional *intra*- and inter-study co-registration was performed, using a deformable non-rigid registration algorithm which has been optimized for 3D

multimodal image registration [13]. *Intra*-study registration non-rigidly coregistered unenhanced, arterial phase, venous phase and diffusion-weighted images. *Inter*-study registration registered the portal venous data between studies after *intra*-study registration. The *intra*-study registration step corrects for non-rigid deformations of the liver between sequences, as the depth of inspiration may vary and shape variations of the liver between sequences may result. After coregistration, a voxelwise comparison between imaging volumes becomes possible and a single ROI can be used to evaluate the same tissue on different sequences. After coregistration, index lesions were segmented semi-automatically, using a graph-based interactive segmentation technique which is based on random walks in multimodal image space, and which achieves a quick and accurate segmentation based on a small number of user-defined labels [14,15]. Systematic evaluations of segmentation performance of this algorithm are available in the literature [15]. For segmentation, the user labels voxels within the tumor and in the vicinity of the tumor border (in normal appearing liver parenchyma) on a few slices. A full 3D labeling on all slices is not necessary, and labeled slices do not need to be directly adjacent. On the high resolution datasets labeling of between 3 and 5 slices of up to 80 slices encompassing the entire tumor was usually sufficient to achieve a visually correct segmentation. Labeling is performed using a “brush” tool, such that many voxels can be labeled at the same time. Brush size is adjustable for most efficient labeling of the lesion. A few central brush strokes are sufficient and a delineation of the boundary is not needed. The algorithm then calculates the segmentation. Segmentation was accurate upon completion of the first segmentation in most cases. All segmentation results were visually inspected for accuracy. The labeling can be refined and the segmentation repeated as needed until a visually acceptable segmentation is achieved. After segmentation, the software automatically performs a comparison of the volumetric ADC, AE, and VE measurements before and after IAT. Parameter changes are calculated at a voxel level between coregistered images pre- and post-IAT and summarized in a histogram. In addition, descriptive statistics of the entire tumor volume are calculated. Three regions of the histogram are defined to represent parameter decrease, parameter increase and no significant parameter change. Voxels belonging to the three histogram regions are displayed as pseudo-colored parametric maps overlaid on the anatomic image for visual assessment the response of different tumor regions: red voxels (V_r) represent significant elevation of parameters above the background threshold of ADC ($+160 \mu\text{m}^2/\text{s}$) or enhancement (AE and VE; $+10\%$), blue voxels (V_b) represent significant decrease of ADC ($-160 \mu\text{m}^2/\text{s}$) or enhancement (-10%) parameters, and green voxels (V_g) represent no significant change. For interobserver comparison only two numeric representations were utilized: Percent of tumor volume with significant elevation of ADC ($+160 \mu\text{m}^2/\text{s}$) and percent of tumor volume with significant decrease in enhancement (-10%). The threshold selection for this “significant change” in ADC or enhancement values has been previously described [4]. Fig. 2 shows an example of the semi-automated, volumetric, co-registered MRI data analysis using method 1. Method 1 can only be used to obtain functional volumetric ADC and enhancement, as well as pre-treatment size. The tumor size is assumed constant in method 1. This assumption is reasonable as tumor volume changes after treatment are expected to be insignificant in comparison to the large size of the assessed tumors, e.g. a recent study of Bonekamp et al. reported an average volumetric change of only -3.9% [8]. The ROI generated from the segmentation of the pre-IAT volume is applied

also to the post-IAT volume. This is based on our observation that size changes in the immediate follow-up interval are often negligible. At the same time, the generation of pre- and post-IAT ROIs yields a significant operator processing time reduction of approximately one half, as the segmentation is the most time consuming part of the analysis. If tumor size reduction occurs it will thus lead to the inclusion of a peripheral rim of normal appearing liver tissue in the ROI data of the post-IAT study and influence the average and standard deviation of the parameters (ADC, VE, AE) commensurate to the volume partition of additionally included liver parenchyma. Method 2 (below) specifically addresses this potential inaccuracy by performing a separate segmentation on both volumes. Furthermore, neither method 1 nor method 2 (see below) provide a measurement of EASL or RECIST. These measurements would require a separate segmentation of the contrast-enhancing part of the tumor pre- and post-treatment.

2.5. Method 2 - unregistered volumetric analysis

For method 2 the same data was imported into the same software compared to method 1. The difference in the analysis is twofold. First, segmentation is performed separately on the pre-IAT and post-IAT volumes. Second, a coregistration is not performed, as two separate ROIs are available to extract the tumor volume, and a direct voxelwise comparison and color-coding according to parameter increase or decrease is not possible with this approach. Method 2 required approximately twice the time needed for Method 1, as the semi-automatic segmentation was carried out on both volumes. Method 2 was performed at a different time and independent of Method 1, i.e. the segmentation from Method 1 was not used for the pre-IAT ROI definition in Method 2. Subsequently, the software automatically calculated descriptive statistics of the volumetric data including ADC and enhancement values, and it also automatically calculated tumor volume and RECIST diameter before and after IAT. Fig. 3 shows an example of the semi-automated, volumetric MRI data analysis using method 2.

2.6. Method 3 - manual ROI-based analysis

MR images were analyzed manually using our institutional image viewing system. Each reader matched the post-IAT slice as closely as possible to the pre-IAT slice for these measurements. First, RECIST diameter was measured before and after IAT on an axial slice which showed the largest tumor diameter on post-contrast images. Next, EASL diameter before and after IAT was measured on the axial image containing the largest enhancing tumor diameter. Finally, mean ADC, AE, and VE was measured within the tumor, placing a single ROI of at least 1 cm diameter on the axial slice that provided the largest tumor area encompassing the entire tumor area as far as possible. T1-weighted pre- and post-contrast images were utilized to guide the ROI placement on ADC maps. Enhancement (AE and VE) was determined as the absolute signal intensity within the ROI. No subtraction images were generated or utilized. Fig. 4 shows an example of the manual MRI data analysis using method 3.

2.7. Statistical analysis

All analyses were conducted using Stata 12.0 (StataCorp, College Station, Texas, USA). First, mean values and standard deviation (SD) of the HCC index lesion before and after IAT were calculated for each measurement.

Interobserver comparison—To determine interobserver agreement we calculated between-subject as well as within-subject SD of each variable. The between subject SD is a measure of the overall variability between all subjects. The within-subject SD is a measurement reflecting the agreement between readers analyzing each subject. Intraclass correlation coefficients (ICCs), i.e. the ratio of the between subject variance to the total variance of measurements, were calculated based on repeated measures ANOVA [16,17]. ICC results were interpreted according to the following criteria: poor ($ICC < 0.50$), moderate ($0.50 < ICC < 0.75$), good ($0.75 < ICC < 0.90$), and excellent ($ICC > 0.90$). Bland-Altman plots were utilized to graph the data [18].

3. Results

Demographics and clinical characteristics of the 50 patients are summarized in Table 1. Median age was 65 years (range, 50–88). The majority of patients was male (41/50, 82%). Inter-study co-registration of the pre- and post-IAT MRI data was not possible in one case and results were therefore not available for method 1. Furthermore, the software did not provide post-IAT ADC values for 2 patients due to intra-study misregistration between ADC maps and T1-weighted images, resulting in missing data for method 1 and method 2. EASL measurements could not be performed in 7 patients due to heterogeneous tumor enhancement or rim enhancement. An example of a case in which EASL could not be assessed is shown in Fig. 5.

Using Method 1 the interobserver variability measuring mean ADC and mean enhancement was good-to-excellent (Table 2). With Method 2 the interobserver variability was excellent for mean ADC, mean AE, mean VE, RECIST diameter, and tumor volume (Table 2). Method 3 resulted in moderate-to-good interobserver variability for mean ADC, mean AE, mean VE, RECIST diameter, and EASL diameter (Table 2).

3.1. Response to treatment

After calculating changes in mean ADC, mean enhancement, and tumor diameter, the interobserver variability of the treatment induced change in each parameter was determined. Using Method 1, the intraobserver variability remained good to excellent for all measurements ($ICC: 0.830\text{--}0.910$) (Table 3). With Method 2, the interobserver variability was excellent for change in mean ADC ($ICC = 0.903$) and good for change in VE ($ICC = 0.877$) and tumor volume ($ICC = 0.854$) but only moderate for mean AE ($ICC = 0.715$) and changes in RECIST diameter ($ICC = 0.655$) (Table 3). Method 3 resulted in poor-to-moderate interobserver variability for change in mean ADC, mean VE, as well as change in RECIST and EASL diameter ($ICC: 0.157\text{--}0.653$) (Table 3).

The results of our study can be summarized as follows: (1) The two methods (method 1 and method 2) using semi-automated, volumetric analysis of the anatomic and functional MRI

parameters before and after IAT yielded good interobserver agreement for evaluating tumor ADC, AE, VE, RECIST and tumor volume (ICC from 0.864 to 0.996); (2) interobserver agreement for the manual analysis of anatomic and functional MRI parameters (method 3) was moderate to good (ICC from 0.543 to 0.799), the interobserver agreement for the measurement of EASL was good (ICC 0.758 and 0.809); and (3) for all comparisons the interobserver agreement of the two semi-automated methods (method 1 and method 2) was higher than interobserver agreement within the manual measurements (method 3).

4. Discussion

The aim of this study was to describe the interobserver agreement of MR imaging-based criteria for the assessment of treatment response. Three readers used three different methods to obtain conventional (axial, anatomic) as well as novel (volumetric, functional) measurements within the HCC index lesion before and after IAT in 50 patients with unresectable HCC.

As hypothesized, the semi-automated analysis of functional and morphologic MRI metrics resulted in a higher interobserver agreement compared with the manual approach. The reproducibility for change in tumor diameter was overall lower compared with functional measurements. Among the morphologic measurements the interobserver agreement was poor for the 1-dimensional, manual assessment of RECIST. Prior studies have reported results ranging from no significant difference to substantial inter- and intraobserver variability in lesion size measurements [19,20]. Comparing inter- and intraobserver agreement of linear tumor measurements using a caliper measurement versus an edge tracing method, Monsky et al. [19] found no statistically significant differences between the measures and no greater or lesser interobserver reproducibility. On the other hand, a CT-based study by Rothe et al. [21] found significantly higher intraobserver differences within manual RECIST measurements compared with three different, simple 3D-methods (threshold based segmentation, manual slice segmentation, seeded region growing method). Another CT-based study in 32 patients with colorectal cancer and liver metastases found comparable results in interobserver agreement between manual and semi-automated methods [22]. However, this study evaluated RECIST measurements only and did not assess any volumetric size criteria, hindering a direct comparison of studies.

Change in EASL size measured manually displayed a fairly high interobserver agreement (ICC = 0.653). This confirms results of prior studies [23,24]. Galizia et al. [24] determined a slightly higher agreement between observers for EASL in a study of 29 consecutive patients with HCC treated with ⁹⁰Y radioembolization (ICC = 0.7232). The higher ICC values determined by Galizia et al. [24] could be explained by the varying experience of the readers in our study. Another possible cause for this discrepancy in interobserver precision is that the study by Galizia et al. [24] had two readers, whereas our analysis had three readers, which likely introduced some additional variability. However, one limitation of EASL is that it cannot always be assessed and, indeed, could not be determined in 7 subjects (14%) in our study. Furthermore, the correct method used to evaluate viable tumor tissue is a matter of debate. Most investigators have quantified necrosis using the cross-product of the whole tumor on the axial image which had the largest cross-sectional area, modified from WHO

guidelines [3,24,25], others summed up all of the necrotic areas in one lesion [23] or used a combination of all targeted lesions [26]; complicating the standardization and comparison of EASL response assessment.

There are numerous studies of the reproducibility and interobserver agreement of functional measures including contrast-enhancement and ADC [12,24,27–29]. However, while the variability of values obtained in healthy volunteers or normal tissue is usually small, i.e. with a coefficient of variation around 15% for ADC values [27,30,31], the variability of ADC values measured by different observers in lesions or lymph nodes has a reported range from substantial to good interobserver agreement [32,33] to non-significant with excellent interobserver agreement [29]. Similarly, studies investigating the interobserver agreement of contrast-enhanced MRI measurements have reported interobserver agreements ranging from very good to only moderate [12,34].

In our study, the interobserver agreement of measurements of functional MRI parameters, namely enhancement and ADC, was very good when the semi-automated methods were employed (ICC from 0.830–0.910), but were lower for manual assessment of VE (ICC = 0.648) and very low for the manual assessment of AE (ICC = 0.202) and ADC (ICC = 0.157). These results confirm the finding of previous studies which had reported higher inter- or intraobserver agreement when using volumetric, semi-automated measurements compared to manual, ROI-based measurements [24,35,36]. In addition, the very low interobserver agreement of manual ADC and AE measurements may be due to the high variability in ADC and AE values within the large and heterogeneous HCC lesions. Selection of the axial slice and placement of the ROI can influence the obtained result greatly. While a previous study found no correlation between the heterogeneity of the lesion and the interobserver difference of ADC measurements [28], the authors also admitted that interobserver agreement was higher if only the areas of most restricted diffusion were assessed instead of the whole lesion. Since we sought to simulate clinical image interpretation where the radiologist decides the location of the ROI, the observers in our study were instructed to include as much of the tumor as possible into an ROI, but had no additional prior consensus for the placement of ROIs (i.e. ROI selection based on predominantly on ADC maps versus ROI selection based predominantly on T1 weighted images). Therefore, agreement may have been influenced this way. This reinforces the need for rigorous formal training for manual measurements and for development of automated segmentation algorithms to obtain highly reproducible measurements.

This study has limitations. First, our study was performed in a single, major referral center and therefore we included a number of large unresectable locally advanced tumors (> 2 cm). Future studies should investigate volumetric change in early stage, small tumors (<2 cm) to show the reliability of this technique in such cases. Additionally, the inclusion and exclusion criteria used may have lead to a selection bias, since we only included patients who underwent IAT and MR imaging at one specific scanner. We did not have pathologic correlation of tumor size or viability so we did not measure accuracy of the results but rather reproducibility of the measurements. In addition we did not evaluate the correlation between imaging response and patient survival, because it was addressed in prior studies and was beyond the scope of this work. Moreover, we did not investigate the interobserver agreement

between different response categories stratified according to EASL or RECIST. We omitted this step because we felt that comparing the continuous values instead of categorical variables would lead to a more precise assessment of the interobserver agreement. Furthermore, a large number of lesions are generally classified as Stable Disease and this classification will not change even if the actual tumor size measurements vary a lot between readers because the range for the classification “Stable Disease” for RECIST (<30% decrease to 20% increase in maximum diameter of the index lesion) and EASL (<50% decrease to 25% increase in amount of enhancing tissue in the index lesion) are fairly wide. Finally, we did not assess intraobserver agreement, because intraobserver agreement is generally higher than interobserver agreement. Therefore, we considered interobserver agreement as the upper limit of intraobserver agreement. Additionally, we assumed that the influence of the observer on measurements will be more important in clinical follow-up as well as long term studies of treatment response.

The potential of diffusion-weighted imaging and contrast-enhanced MRI for early response monitoring of IAT has been demonstrated. However, validation is hampered by the lack of reproducibility and standardization. Objective evaluation parameters of tumor response after IAT are needed. Reliability studies are an important step in the establishment of new biomarkers in oncologic imaging. Our results suggest that volumetric measurements of tumor size and function using semi-automated software are highly reproducible. We believe that further research into the potential of semi-automated, volumetric assessment of tumor size, as well as tumor diffusion and perfusion, may result in a new powerful tool for response evaluation. A major advantage for the individual patient would be the availability of early response evaluation, which may spare non-responding patients the toxicity and expenses and allow for a modification of their treatment plan.

5. Conclusion

Current standardized MRI assessment schemes for treatment response have only limited ability to assess the heterogenous composition of HCC lesions before and after treatment. Semi-automated measurements of functional changes assessed by ADC and VE based on whole-lesion segmentation demonstrate better reproducibility than ROI-based axial measurements, or RECIST or EASL measurements.

Acknowledgments

Financial disclosures

Dr. Susanne Bonekamp is receiving salary support from Siemens Medical.

Dr. Kamel is receiving grant support from Siemens Medical.

Abbreviations

IAT	intra-arterial therapy
ADC	apparent diffusion coefficient

AE	arterial phase enhancement
VE	portal venous phase enhancement
ROI	region of interest
RECIST	Response Evaluation Criteria in Solid Tumors
EASL	European Association for the Study of the Liver
MRI	magnetic resonance imaging
HCC	hepatocellular carcinoma
ICC	intraclass correlation coefficients
TACE	transarterial chemoembolization
DEB-TACE	drug-eluting bead transarterial chemoembolization
AFP	alpha fetoprotein
SD	standard deviation

References

1. Therasse P, Arbuck SG, Eisenhauer EA, et al. New guidelines to evaluate the response to treatment in solid tumors European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. *Journal of the National Cancer Institute*. 2000; 92(3):205–216. [PubMed: 10655437]
2. Suzuki C, Jacobsson H, Hatschek T, et al. Radiologic measurements of tumor response to treatment: practical approaches and limitations. *Radiographics: A Review Publication of the Radiological Society of North America, Inc.* 2008; 28(2):329–344.
3. Bruix J, Sherman M, Llovet JM, et al. Clinical management of hepatocellular carcinoma. Conclusions of the Barcelona-2000 EASL conference European Association for the Study of the Liver. *Journal of Hepatology*. 2001; 35(3):421–430. [PubMed: 11592607]
4. Bonekamp S, Jolepalem P, Lazo M, Gulsun MA, Kiraly AP, Kamel IR. Hepatocellular carcinoma: response to TACE assessed with semiautomated volumetric and functional analysis of diffusion-weighted and contrast-enhanced MR imaging data. *Radiology*. 2011; 260(3):752–761. [PubMed: 21771960]
5. Gowdra Halappa V, Corona-Villalobos CP, Bonekamp S, et al. Neuroendocrine liver metastasis treated by using intraarterial therapy: volumetric functional imaging biomarkers of early tumor response and survival. *Radiology*. 2013; 266(2):502–513. [PubMed: 23192780]
6. Li Z, Bonekamp S, Halappa VG, et al. Islet cell liver metastases: assessment of volumetric early response with functional MR imaging after transarterial chemoembolization. *Radiology*. 2012; 264(1):97–109. [PubMed: 22627602]
7. Rhee TK, Lewandowski RJ, Liu DM, et al. 90Y Radioembolization for metastatic neuroendocrine liver tumors: preliminary results from a multi-institutional experience. *Annals of Surgery*. 2008; 247(6):1029–1035. [PubMed: 18520231]
8. Bonekamp S, Li Z, Geschwind JF, et al. Unresectable hepatocellular carcinoma: MR imaging after intraarterial therapy. Part I Identification and validation of volumetric functional response criteria. *Radiology*. 2013; 268(2):420–430. [PubMed: 23616631]
9. Bonekamp S, Halappa VG, Geschwind JF, et al. Unresectable hepatocellular carcinoma: MR imaging after intraarterial therapy. Part II. Response stratification using volumetric functional criteria after intraarterial therapy. *Radiology*. 2013; 268(2):431–439. [PubMed: 23616632]

10. Heijmen L, TerVoert EE, Nagtegaal ID, et al. Diffusion-weighted MR imaging in liver metastases of colorectal cancer: reproducibility and biological validation. *European Radiology*. 2013; 23(3): 748–756. [PubMed: 23001604]
11. Kim SY, Lee SS, Byun JH, et al. Malignant hepatic tumors: short-term reproducibility of apparent diffusion coefficients with breath-hold and respiratory-triggered diffusion-weighted MR imaging. *Radiology*. 2010; 255(3):815–823. [PubMed: 20501719]
12. Kim S, Mannelli L, Hajdu CH, et al. Hepatocellular carcinoma: assessment of response to transarterial chemoembolization with image subtraction. *Journal of Magnetic Resonance Imaging*. 2010; 31(2):348–355. [PubMed: 20099348]
13. Chef'd'Hotel C, Hermosillo G, Faugeras O. Flows of diffeomorphisms for multimodal image registration. In: *Proceedings 2002 IEEE International Symposium on Biomedical Imaging*. 2002:753–756.
14. Grady L. Random walks for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2006; 28(11):1768–1783. [PubMed: 17063682]
15. Moschidis E, Graham J. A systematic performance evaluation of interactive image segmentation methods based on simulated user interaction. In: *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. 2010:928–931.
16. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*. 1966; 19(1):3–11. [PubMed: 5942109]
17. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength & Conditioning Research*. 2005; 19(1):231–240. [PubMed: 15705040]
18. Martin Bland J, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986; 327(8476):307–310.
19. Monsky WL, Raptopoulos V, Keogan MT, et al. Reproducibility of linear tumor measurements using PACS: comparison of caliper method with edge-tracing method. *European Radiology*. 2004; 14(3):519–525. [PubMed: 14658000]
20. Suzuki C, Torkzad MR, Jacobsson H, et al. Interobserver and intraobserver variability in the response evaluation of cancer therapy according to RECIST and WHO-criteria. *Acta Oncologica*. 2010; 49(4):509–514. [PubMed: 20397778]
21. Rothe JH, Grieser C, Lehmkuhl L, et al. Size determination and response assessment of liver metastases with computed tomography—comparison of RECIST and volumetric algorithms. *European Journal of Radiology*. 2013; 82(11):1831–1839. [PubMed: 22717124]
22. Kalkmann J, Ladd SC, de Greiff A, Forsting M, Statta J. Suitability of semi-automated tumor response assessment of liver metastases using a dedicated software package. *RöFo: Fortschritte auf dem Gebiete der Röntgenstrahlen und der Nuklearmedizin*. 2010; 182(7):581–588. [PubMed: 20419616]
23. Duke E, Deng J, Ibrahim SM, et al. Agreement between competing imaging measures of response of hepatocellular carcinoma to yttrium-90 radioembolization. *Journal of Vascular and Interventional Radiology*. 2010; 21(4):515–521. [PubMed: 20172741]
24. Galizia MS, Tore HG, Chalian H, McCarthy R, Salem R, Yaghmai V. MDCT necrosis quantification in the assessment of hepatocellular carcinoma response to yttrium 90 radioembolization therapy: comparison of two-dimensional and volumetric techniques. *Academic Radiology*. 2012; 19(1):48–54. [PubMed: 22054801]
25. Forner A, Ayuso C, Varela M, et al. Evaluation of tumor response after locoregional therapies in hepatocellular carcinoma: are response evaluation criteria in solid tumors reliable? *Cancer*. 2009; 115(3):616–623. [PubMed: 19117042]
26. Shim JH, Lee HC, Kim SO, et al. Which response criteria best help predict survival of patients with hepatocellular carcinoma following chemoembolization? A validation study of old and new models. *Radiology*. 2012; 262(2):708–718. [PubMed: 22187634]
27. Braithwaite AC, Dale BM, Boll DT, Merkle EM. Short- and midterm reproducibility of apparent diffusion coefficient measurements at 3.0-T diffusion-weighted imaging of the abdomen. *Radiology*. 2009; 250(2):459–465. [PubMed: 19095786]
28. Lu TL, Meuli RA, Marques-Vidal PM, Bize P, Denys A, Schmidt S. Interobserver and intraobserver variability of the apparent diffusion coefficient in treated malignant hepatic lesions

- on a 3.0T machine: measurements in the whole lesion versus in the area with the most restricted diffusion Journal of magnetic resonance imaging. Journal of Magnetic Resonance Imaging. 2010; 32(3):647–653. [PubMed: 20815063]
29. Petralia G, Bonello L, Summers P, et al. Intraobserver and interobserver variability in the calculation of apparent diffusion coefficient (ADC) from diffusion-weighted magnetic resonance imaging (DW-MRI) of breast tumours. La Radiologia Medica. 2011; 116(3):466–476. [PubMed: 21225368]
30. Rosenkrantz AB, Oei M, Babb JS, Niver BE, Taouli B. Diffusion-weighted imaging of the abdomen at 3.0Tesla: image quality and apparent diffusion coefficient reproducibility compared with 1.5Tesla. Journal of Magnetic Resonance Imaging. 2011; 33(1):128–135. [PubMed: 21182130]
31. Sasaki M, Yamada K, Watanabe Y, et al. Variability in absolute apparent diffusion coefficient values across different platforms may be substantial: a multivendor, multi-institutional comparison study. Radiology. 2008; 249(2):624–630. [PubMed: 18936317]
32. Kwee TC, Takahara T, Luijten PR, Nievelstein RA. ADC measurements of lymph nodes: inter- and intraobserver reproducibility study and an overview of the literature. European Journal of Radiology. 2010; 75(2):215–220. [PubMed: 19375259]
33. Lambregts DM, Maas M, Riedl RG, et al. Value of ADC measurements for nodal staging after chemoradiation in locally advanced rectal cancer—a per lesion validation study. European Radiology. 2011; 21(2):265–273. [PubMed: 20730540]
34. Lambregts DM, Beets GL, Maas M, et al. Accuracy of gadofosveset-enhanced MRI for nodal staging and restaging in rectal cancer. Annals of Surgery. 2011; 253(3):539–545. [PubMed: 21239980]
35. Chalian H, Tochetto SM, Tore HG, Rezai P, Yaghmai V. Hepatic tumors: region-of-interest versus volumetric analysis for quantification of attenuation at CT. Radiology. 2012; 262(3):853–861. [PubMed: 22357887]
36. Heye T, Merkle EM, Reiner CS, et al. Reproducibility of dynamic contrast-enhanced MR imaging. Part II Comparison of intra- and interobserver variability with manual region of interest placement versus semiautomatic lesion segmentation and histogram analysis. Radiology. 2013; 266(3):812–821. [PubMed: 23220891]

*Caveat: Co-registration is not always possible since index lesion may change in size or form. Co-registration of lesions that changed size or form after therapy will lead to misregistration.

Fig. 1.

Diagram depicting the three methods employed to analyze the MRI data of 50 patients with HCC before and after IAT. Cutoffs used in Method 1 are based on a prior study of the scan-rescan variability of the apparent diffusion coefficient and venous enhancement in the liver [4].

Fig. 2.

Example of the semi-automated, volumetric, co-registered MRI data analysis (method 1) in a 61 year-old patient with Hepatitis C infection and multiple HCC. The color scale for the segmented tumor lesion and histogram was chosen as follows: red voxels represent elevation of ADC above the background threshold, blue voxels represent decrease of ADC, and green voxels represent no change. (A) Segmentation of the HCC index lesion based on co-registered images (red). (B) Analysis of the tumor pre- and 1 month post-treatment ADC show that 94.0% of the tumor volume increased in ADC above the predetermined threshold

of $160 \times 10^{-5} \text{ mm}^2/\text{s}$ (red voxels, upper and lower right tiles). The blue and orange histograms (left lower tile) depict the distribution of ADC on the pre- (blue) and post-IAT (orange) studies. (C) Volumetric venous enhancement (VE) measurements: Analysis of the tumor pre- and 1 month post-treatment of the venous enhancement (VE) phase show that 85.0% of the tumor volume decreased in VE below the predetermined threshold of 10% (blue voxels in upper and lower right tiles). The blue and orange histograms (left lower tile) depict the distribution of venous enhancement on the pre- (blue) and post-IAT (orange) studies.

Fig. 3.

Example of the semi-automated, volumetric MRI data analysis (method 2) in the same 61-year old patients with Hepatitis C infection and multiple HCC shown in Fig. 2. (A) Segmentation of the HCC index lesion based on pre-treatment images with a volume of 223 cm³, a RECIST diameter of 80 mm (red). (B) Segmentation of the lesion based on post-treatment images, the volume was 271 cm³, RECIST diameter 87 mm (green). (C) The light blue histogram shows the ADC distribution of the tumor before treatment (mean ADC = 157.4×10^{-5} mm²/s). Measurements are based on the tumor segmentation (Segm1, red)

shown in (A). (D) The orange histogram shows the ADC distribution of the tumor after treatment (mean ADC = $269.8 \times 10^{-5} \text{ mm}^2/\text{s}$). Measurements are based on the tumor segmentation (Segm2, green) shown in (B). (E) The light blue histogram shows the venous enhancement (VE) distribution of the tumor before treatment (mean VE = 77.3%). Measurements are based on the tumor segmentation (Segm1, red) shown in (A). (F) The orange histogram shows the VE distribution of the tumor after treatment (mean VE = 38.9%). Measurements are based on the tumor segmentation (Segm2, green) shown in (B).

Fig. 4.

Example of the manual MRI data analysis (method 3) in the same 61-year old patients with Hepatitis C infection and multiple HCC shown in Fig. 2. (A) Unenhanced MRI before treatment (SI = 97.66). (B) Unenhanced MRI after treatment (SI = 82.18). (C) Arterial phase MRI before treatment (SI = 89.50, AE = 0%) and EASL measurement (arrows) before treatment ($7.88 \text{ cm} \times 5.42 \text{ cm} = 42.71 \text{ cm}^2$). (D) Arterial phase MRI after treatment (SI = 90.97, AE = 0%) and EASL measurement after treatment ($7.99 \text{ cm} \times 5.47 \text{ cm} = 43.71 \text{ cm}^2$). (E) Venous phase enhancement before treatment (SI = 181.62, VE = 86.0%). (F) Venous

phase enhancement after treatment (SI = 112.05, VE = 36.3%). (G) Region of interest based ADC measurement before treatment (mean ADC = $149.0 \times 10^{-5} \text{ mm}^2/\text{s}$). (H) Region of interest based ADC measurement before treatment (mean ADC = $320.1 \times 10^{-5} \text{ mm}^2/\text{s}$).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Fig. 5.
Portal venous phase image of a HCC lesion in a 76-year old male with a history of hepatitis B and C before intra-arterial treatment. EASL could not be assessed due to heterogeneous enhancement.

Table 1

Demographics of the 50 patients with hepatocellular carcinoma.

Characteristic		Number of patients (%)
Age: mean $\pm$ standard deviations (range)		66.0 $\pm$ 9.7 (50–88)
Gender	Male	41(82%)
	Female	9(18%)
Ethnic group	Caucasian	29 (58%)
	African American	14(28%)
	Hispanic	3 (6%)
	Asian	3 (6%)
	Other	1 (2%)
Etiology	Alcohol	4 (8%)
	Hepatitis C Virus	18 (36%)
	Hepatitis C and ALD	5 (10%)
	Hepatitis C and HIV	4 (8%)
	Hepatitis B Virus	8 (16%)
	Hepatitis B and C	3 (6%)
	Hepatitis B and HIV	1 (2%)
	Cryptogenic cirrhosis	4 (8%)
	Nonalcoholic steatohepatitis	3 (6%)
Barcelona Clinic Liver	A	12 (24%)
Cancer Stage (BCLC)	B	17 (34%)
	C	13(26%)
	D	8(16%)

Table 2

Mean values, standard deviation and intraclass correlation coefficient (ICC) of all measurements before and after IAT.

Method	Measurement	N	Mean value	Between subject SD	Within subject SD	ICC (95% CI)
Method 1: Co-registered, volumetric analysis	Mean ADC before IAT, $\mu\text{m}^2/\text{s}$	49	1600.27	362.09	129.43	0.886 (0.826, 0.930)
	Mean ADC after IAT, $\mu\text{m}^2/\text{s}$	47	1763.63	422.36	136.69	0.902 (0.848, 0.940)
	Mean AE before IAT, %	49	35.41	20.14	6.94	0.985 (0.974, 0.992)
	Mean AE after IAT, %	49	30.67	23.26	5.65	0.968 (0.943, 0.982)
	Mean VE before IAT, %	49	71.35	33.06	6.82	0.959 (0.936, 0.975)
	Mean VE after IAT, %	49	57.72	25.35	9.95	0.864 (0.794, 0.915)
Method 2: Un-registered, volumetric (except for RECIST) analysis	Mean ADC before IAT, $\mu\text{m}^2/\text{s}$	50	1580.08	352.32	65.20	0.967 (0.948, 0.980)
	Mean ADC after IAT, $\mu\text{m}^2/\text{s}$	48	1678.99	280.44	83.41	0.919 (0.873, 0.951)
	Mean AE before IAT, %	50	37.41	20.78	6.94	0.900 (0.830, 0.942)
	Mean AE after IAT, %	50	32.35	23.26	5.65	0.944 (0.904, 0.968)
	Mean VE before IAT, %	50	71.57	33.73	5.88	0.971 (0.953, 0.982)
	Mean VE after IAT, %	50	52.01	30.26	6.68	0.954 (0.927, 0.972)
	RECIST before IAT, mm	50	88.68	47.00	11.59	0.943 (0.910, 0.965)
	RECIST after IAT, mm	50	84.12	45.61	10.20	0.952 (0.925, 0.971)
	Volume before IAT, mm^3	50	341.12	495.86	30.56	0.996 (0.994, 0.998)
	Volume after IAT, mm^3	50	306.27	463.66	35.20	0.994 (0.991, 0.997)
Method 3: Manual size and, ROI-based analysis in axial plane	Mean ADC before IAT, $\mu\text{m}^2/\text{s}$	50	1389.63	442.13	221.96	0.799 (0.702, 0.872)
	Mean ADC after IAT, $\mu\text{m}^2/\text{s}$	50	1595.10	453.03	227.33	0.799 (0.702, 0.872)
	Mean AE before IAT, %	50	192.72	111.72	51.50	0.786 (0.685, 0.864)
	Mean AE after IAT, %	50	155.97	90.03	52.72	0.364 (0.189, 0.540)
	Mean VE before IAT, %	50	79.35	36.92	34.32	0.536 (0.375, 0.682)
	Mean VE after IAT, %	50	42.62	34.00	22.64	0.693 (0.563, 0.799)
	RECIST before IAT, mm	50	78.97	31.72	29.10	0.543 (0.382, 0.687)
	RECIST after IAT, mm	50	75.13	33.40	27.52	0.596 (0.443, 0.728)
	EASL before IAT, mm	45	72.81	37.92	21.43	0.758 (0.640, 0.849)
	EASL after IAT, mm	46	54.03	39.26	19.07	0.809 (0.712, 0.882)

Abbreviations: ADC, apparent diffusion coefficient; AE, arterial enhancement; EASL, European Association for the Study of the Liver; CI, confidence interval; IAT, intra-arterial therapy; ICC, intraclass correlation coefficient; RECIST, Response Evaluation Criteria in Solid Tumors; ROI, region of interest; SD, standard deviation; VE, venous enhancement.

Table 3

Mean values, standard deviation and intraclass correlation coefficient of measurements of change after IAT assessing response to treatment.

Method	Measurement ratio post-IAT to pre-IAT	N	Mean value	Between subject SD	Within subject SD	ICC (95% CI)
Method 1: Co-registered, volumetric analysis	Mean ADC, % ^a	47	11.54	20.85	9.25	0.830 (0.745, 0.894)
	Mean AE, % ^a		25.77	51.92	17.31	0.987 (0.977, 0.993)
	Mean VE, % ^a	49	-8.31	150.99	15.93	0.910 (0.862, 0.945)
	Tumor volume with increased ADC, % ^a	47	47.84	21.05	9.37	0.829 (0.744, 0.893)
Method 2: Unregistered, volumetric (except for RECIST) analysis	Tumor volume with decreased AE, % ^a	49	43.98	28.25	2.61	0.992 (0.985, 0.995)
	Tumor volume with decreased VE, % ^a	49	53.03	24.30	10.37	0.845 (0.766, 0.903)
	Mean ADC, % ^b	48	7.86	17.91	5.86	0.903 (0.850, 0.941)
	Mean AE, % ^b		15.28	121.56	76.78	0.715 (0.547, 0.828)
Method 3: Manual size and ROI-based analysis in axial plane	Mean VE, % ^b	50	-23.12	37.29	13.95	0.877 (0.813, 0.924)
	RECIST, %	50	-4.28	10.82	7.85	0.655 (0.516, 0.772)
	Volume, %	50	-5.19	43.60	18.01	0.854 (0.780, 0.909)
	Mean ADC, %	50	36.10	68.17	157.77	0.157 (-0.008, 0.347)
	Mean AE, %	50	-16.25945	50.56	22.21	0.202, (0.032, 0.391)
	Mean VE, %	50	-40.54	43.53	35.26	0.648, (0.534, 0.838)
	RECIST, %	50	-6.03	10.03	22.28	0.169 (0.002, 0.359)
	EASL, % ^c	43	-32.89	26.67	19.42	0.653 (0.501, 0.779)

Abbreviations: ADC, apparent diffusion coefficient; AE, arterial enhancement; EASL, European Association for the Study of the Liver; CI, confidence interval; IAT, intraarterial therapy; ICC, intraclass correlation coefficient; RECIST, Response Evaluation Criteria in Solid Tumors; ROI, region of interest; SD, standard deviation; VE, venous enhancement.

^aTumor volume with an increase in ADC after therapy larger than $160 \times 10^{-5} \text{ mm}^2/\text{s}$ or tumor volume with a decrease in arterial venous enhancement larger than 10%.

^bOne lesion could not be co-registered and the patient was excluded from method 1 and in two lesions the ADC maps acquired after IAT could not be co-registered within the study and the data is missing for method 1 and method 2.

^cEASL before IAT could not be measured in 5 patients and EASL after IAT could not be measured in 4 patients, resulting in a total of 7 patients who had to be excluded due to missing measurements.