Thieme

Response Evaluation of Malignant Liver Lesions After TACE/SIRT: Comparison of Manual and Semi-Automatic Measurement of Different Response Criteria in Multislice CT

Evaluation des Therapieansprechens maligner Leberläsionen nach TACE/SIRT: Vergleich von manuellen und semiautomatischen Messverfahren unterschiedlicher Response-Kriterien im Multislice-CT

Authors
Anna Janina Höink[1], Christoph Schülke[2], Raphael Koch[3], Annika Löhnert[2], Sara Kammerer[2], Rasmus Fortkamp[2], Walter Heindel[2], Boris Buerke[2]

Affiliations
1 Diagnostic and Interventional Radiology, University Hospital Cologne, Germany
2 Department of Clinical Radiology, University Hospital Münster (UKM), Münster, Germany
3 Institute of Biostatistics and Clinical Research (IBKF), University Hospital Münster (UKM), Münster, Germany

Key words
CT, semi-automatic, tumor response, interventional oncology, interobserver variability

received 27.01.2017
accepted 06.06.2017

Bibliography
DOI https://doi.org/10.1055/s-0043-116220
Published online: 23.8.2017 | Fortschr Röntgenstr 2017; 189: 1067–1075

ISSN 1438-9029

Correspondence
Herr Dr. Christoph Schülke
Institut für Klinische Radiologie, Universitätsklinikum Münster, Albert-Schweitzer-Campus 1, Gebäude A1, 48149 Münster, Germany
Tel.: ++ 49/2 51/8 34 73 10
Fax: ++ 49/2 51/8 34 96 56
schuelke@uni-muenster.de

ZUSAMMENFASSUNG

Ziel Vergleich von Messgenauigkeit und Interobserver-Variabilität in der computertomografischen Beurteilung von hepatozellulären Karzinomen (HCC) und Lebermetastasen vor und nach transarteriellen selektiven Therapien.

Material und Methoden Retrospektive Studie an 72 Patienten mit malignen Leberläsionen (42 Metastasen, 30 HCC) vor und nach Therapie mit SIRT (n = 42) oder TACE (n = 29). Etablierte (LAD, SAD, WHO) und Vitalitäts-assoziierte Größenparameter (mRECIST, mLAD, mSAD, EASL) wurden manuell und semiautomatisch von zwei Auswertern bestimmt. Die relative Interobserverdifferenz (RID) und der Intraclass Korrelationskoeffizient (ICC) wurden berechnet.

Ergebnisse Die mediane RID der Vitalitäts-assoziierten Parameter war für die semiautomatischen niedriger als für die manuellen Messverfahren, im Einzelnen: für mLAD 3,4 % gegenüber 12,5 %; für mSAD 5,7 % gegenüber 12,7 %; für EASL 1,8 % gegenüber 10,4 %. Statistisch signifikante Unterschiede zwischen den etablierten Messverfahren bestanden nicht ($p > 0.05$). Der ICC für LAD (manuell 0,984; semiautomatisch 0,982), SAD (manuell 0,975; semiautomatisch 0,958) und WHO (manuell 0,984; semiautomatisch 0,978) ist für manuelle und semiautomatische Messungen gleichermaßen hoch. Der ICC für manuelle Messungen von mLAD (0,897), mSAD (0,844) und EASL (0,875) ist im Vergleich hierzu niedriger. Diese Reduktion bestand jedoch nicht für die semiautomatischen Messungen von mLAD (0,997), mSAD (0,992) und EASL (0,998).

Schlussfolgerung Die Bestimmung Vitalitäts-assoziierter Größenparameter von HCC und Metastasen nach transarterieller selektiver Therapie ist mit semiautomatischen Messverfahren präziser durchzuführen als mit manuellen Messverfahren. Die hieraus resultierende höhere Reproduzierbarkeit kann die Verlässlichkeit der therapeutischen Entscheidungen verbessern.

Kernaussagen

- Die Größenbestimmung von Leberläsionen nach EASL und mRECIST ist semiautomatisch präziser als manuell.
- Die höhere Präzision ermöglicht eine verlässlichere Klassifikation des Therapieansprechens.
- Die Größenbestimmung nach RECIST und WHO ist semiautomatisch und manuell vergleichbar präzise.

ABSTRACT

Purpose To compare measurement precision and interobserver variability in the evaluation of hepatocellular carcinoma (HCC) and liver metastases in MSCT before and after transarterial local ablative therapies.

Materials and Methods Retrospective study of 72 patients with malignant liver lesions (42 metastases; 30 HCCs) before and after therapy (43 SIRT procedures; 29 TACE procedures). Established (LAD; SAD; WHO) and vitality-based parameters (mRECIST; mLAD; mSAD; EASL) were assessed manually and semi-automatically by two readers. The relative interobserver difference (RID) and intraclass correlation coefficient (ICC) were calculated.

Results The median RID for vitality-based parameters was lower from semi-automatic than from manual measurement of mLAD (manual 12.5 %; semi-automatic 3.4 %), mSAD (manual 12.7 %; semi-automatic 5.7 %) and EASL (manual 10.4 %; semi-automatic 1.8 %). The difference in established parameters was not statistically noticeable ($p > 0.05$). The ICCs of LAD (manual 0.984; semi-automatic 0.982), SAD (manual 0.975; semi-automatic 0.958) and WHO (manual 0.984; semi-automatic 0.978) are high, both in manual and semi-automatic measurements. The ICCs of manual measurements of mLAD (0.897), mSAD (0.844) and EASL (0.875) are lower. This decrease cannot be found in semi-automatic measurements of mLAD (0.997), mSAD (0.992) and EASL (0.998).

Conclusion Vitality-based tumor measurements of HCC and metastases after transarterial local therapies should be performed semi-automatically due to greater measurement precision, thus increasing the reproducibility and in turn the reliability of therapeutic decisions.

Key points

- Liver lesion measurements according to EASL and mRECIST are more precise when performed semi-automatically.
- The higher reproducibility may facilitate a more reliable classification of therapy response.
- Measurements according to RECIST and WHO offer equivalent precision semi-automatically and manually.

Citation Format

- Höink AJ, Schülke C, Koch R et al. Response Evaluation of Malignant Liver Lesions After TACE/SIRT: Comparison of Manual and Semi-Automatic Measurement of Different Response Criteria in Multislice CT. Fortschr Röntgenstr 2017; 189: 1067–1075

Introduction

Liver metastases and advanced primary liver tumors, such as hepatocellular carcinomas (HCCs), are associated with a poor prognosis. Surgical therapeutic options entail tumor or metastasis resection or liver transplantation (in non-metastasized HCCs). Local non-surgical therapies, such as transarterial chemoembolization (TACE), selective internal radiation therapy (SIRT) and radiofrequency ablation (RFA), can be used in hepatic metastases and HCCs [1 – 4]. Systemic therapy could constitute classic cytotoxic chemotherapy or targeted treatment (e. g., sorafenib).

Therapy-induced changes in HCCs and hepatic metastases are classically determined using the firmly established Response Evaluation Criteria in Solid Tumors (RECIST 1.1) [5] or the criteria of the World Health Organization (WHO) [6]. These are based on uni- and bidimensional measurements of the entire lesion which are usually assessed by analyzing transversely oriented images acquired with computed tomography (CT) or magnetic resonance imaging (MRI). However, the mere quantification of the size of a lesion has significant limitations, since treatment such as TACE or SIRT initially results in modified tumor vascularization and not in a size reduction at the initial stage of tumor response [4, 7, 8]. Therefore, the therapeutic effect could be either undetected or underestimated and lead to inappropriate therapeutic decisions (e. g., unnecessary modification of the therapeutic regime).

These limitations led to the development of criteria that also account for vascularization and the extent of possible necrosis. The response criteria according to the European Association of the Study of the Liver (EASL) are based on bidimensional measurements of the vital parts of the tumor, i. e., those revealing arterial contrast uptake [9]. The RECIST guidelines were also adapted to include and quantify tumor necrosis. These criteria are known as modified RECIST (mRECIST) and are still based on unidimensional measurements [10]. Both guidelines primarily aimed to assess HCCs but have already been used in the evaluation of metastases in the context of TACE, SIRT or targeted systemic treatments.

In the clinical routine, radiological assessment of tumor size is usually performed manually. A major disadvantage of manual measurement is the high intraobserver and interobserver variability, which may lead to misinterpretation of tumor response [11]. Previous studies have shown that the use of semi-automatic measuring techniques could reduce this variability [12], leading to increased precision [13] and thus to a more accurate classification of therapeutic response.

The aim of this study was to determine the measurement precision of established and vitality-based response criteria depending on the method (manual versus semi-automatic) used to measure hepatic metastases and HCCs under endovascular therapy.

Materials and Methods

Patients

72 consecutive (January 2008 to December 2013) patients (46 male [64 %], 26 female [36 %]; mean age: 60 years [17 – 83 years]) with HCCs (n = 30 [42 %]) or hepatic metastases (n = 42 [58 %]) from other tumors (19 colorectal adenocarcinomas, 6 breast adenocarcinomas, 4 malignant melanomas, 3 pancreatic adenocarcinomas, 3 respiratory adenocarcinomas, 3 gastrointestinal neu-

roendocrine tumors, 4 other tumor types) were included in this retrospective study.

The inclusion criteria were: (a) transarterial chemoembolization (TACE, n = 29 [40 %]) or (b) selective internal radiation therapy (SIRT, n = 43 [60 %]) of at least one of the liver lesions. Patients who received LIPIODOL® (Guerbet LLC, Bloomington, Indiana, USA) as part of the TACE regimen were excluded due to its high attenuation and interference with the evaluation of contrast enhancement. TACE was conducted by injection of doxorubicin-loaded (50 mg) polyvinyl alcohol particles (100 µm) into the tumor-feeding segmental liver artery. During SIRT loaded spheres (30 – 40 µm) with activities between 0.7 and 1.8 GBq were applied.

All patients underwent contrast-enhanced multislice CT (MSCT) before and after local ablative therapy. Depending on the survival time, the number of procedures (some patients underwent two or more TACE and/or SIRT) and follow-up CT scans varied. A maximum of two lesions were included per patient to avoid bias due to a great number of genetically and presumably phenotypically similar lesions. In the case of more than two lesions, the two largest ones were chosen.

Written informed consent was obtained from every patient. The study was approved by the local ethics committee and conducted in accordance with the guidelines of the institutional review board.

Data acquisition

Images were taken with a 64-slice dual-source CT scanner (SOMATOM® Definition, Siemens AG, Medical Solutions, Forchheim, Germany). The tube voltage was 120 kV and the collimation was 32 × 0.6 mm. Dose modulation (CARE Dose4 D™, Siemens AG, Medical Solutions, Forchheim, Germany) was undertaken to reduce radiation exposure. Iodine-containing contrast agent (Ultravist®-370, Bayer Schering Pharma, Leverkusen, Germany) was injected intravenously at a constant flow rate of 5 mL/s. The arterial contrast phase was determined dynamically by means of bolus tracking. The venous contrast phase was defined by a fixed delay of 75 seconds after i. v. injection. All CT data sets were reconstructed at a slice thickness of 1.5 mm, using a reconstruction interval of 0.6 mm.

Data preparation

The CT examinations were transferred to a commercially available, dedicated oncological software suite (mint Lesion™, Mint Medical GmbH, Heidelberg, Germany), working on a server-client principle. The client, as shown in figure 1, was installed as additional software on dedicated RIS-PACS workstations. A radiologist (15 years of oncological imaging experience), who was not involved in further measurements or data processing, identified and tagged the malignant liver lesions treated with TACE and SIRT on the basis of the images obtained from the intervention and CT examinations. These liver lesions were identified and measured independently by two readers (R1 with 4 years and R2 with 1 year of oncological imaging experience) who were not involved in data preparation. The evaluation was performed in a semi-random order with a time period of at least 1 month between manual and semi-automatic evaluation of the same lesion. Both readers were blinded with respect to the patients' diseases (i. e., type of liver lesion) and treatments. The abovementioned software, which supports both manual and semi-automatic radiological measurements, was used for evaluation purposes.

Manual evaluation

The assessment comprised measurement of the long axis diameter (LAD, [mm]) of the entire lesion as well as determination of the modified long axis diameter (mLAD, [mm]), defined as the arterial contrast-enhancing portion and presumably representing the vital part of the tumor. Both diameters were measured in transversely oriented CT images, illustrating the largest dimension of the lesion.

Perpendicular to the LAD and mLAD, the shortest diameter of the lesion was determined and referred to as the short axis diameter (SAD, [mm]) and modified short axis diameter (mSAD, [mm]), respectively.

Based on these diameters, the areas were calculated by multiplication, resulting in the WHO (LAD × SAD [mm^2]) and the EASL (mLAD × mSAD [mm^2]) areas.

Although it is possible to determine the volume manually, this is not feasible in the daily routine and therefore was not performed.

Semi-automatic evaluation

Semi-automatic two-dimensional (area-based) and three-dimensional (volume-based) segmentation was performed on all tagged lesions.

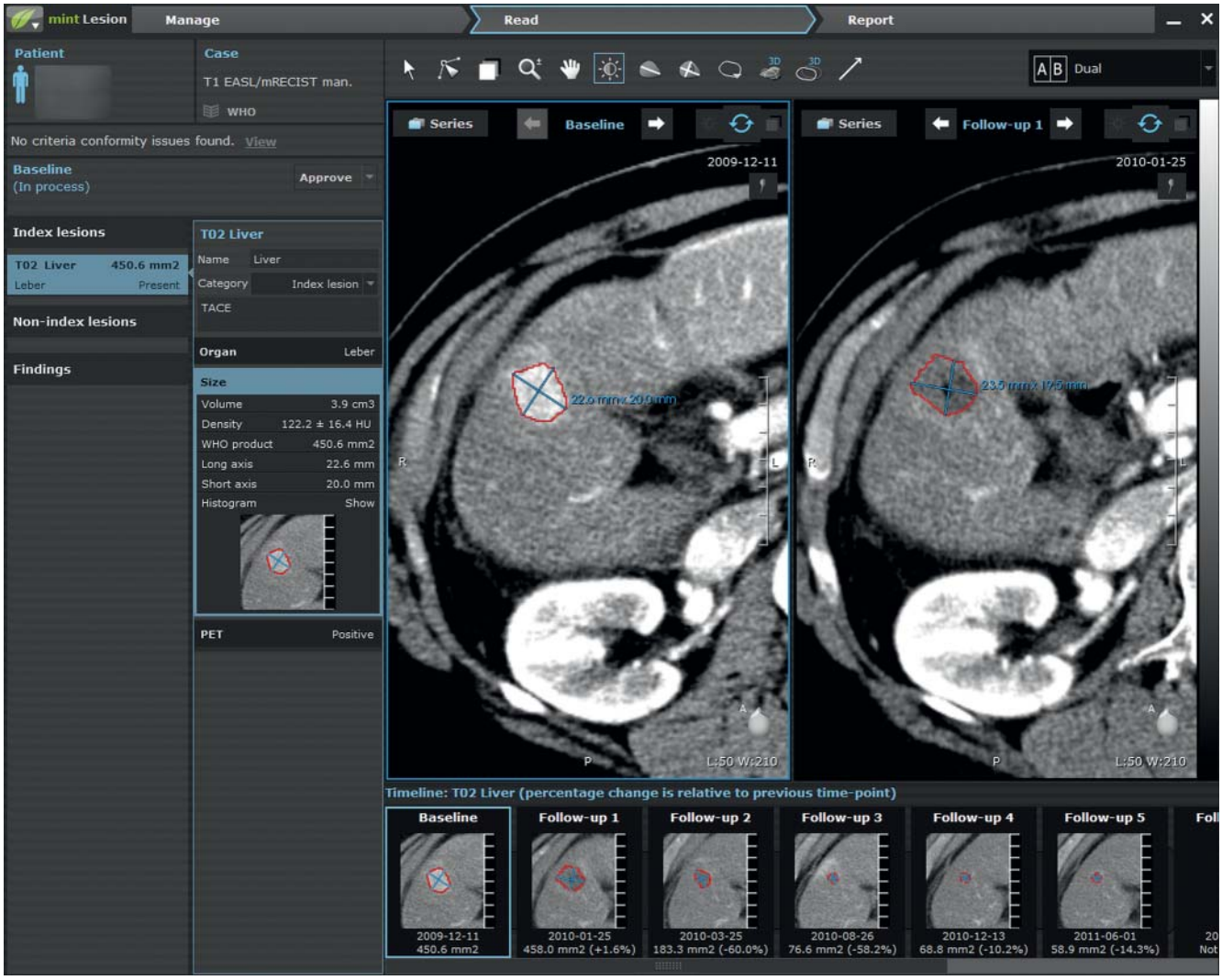
The area-based segmentation process was initiated by drawing a circle around the rough margins of the lesion, preferably in transverse reconstructions. The correct contour was then approximated based on threshold- and contour-based algorithms (▶ **Fig. 1**). Correction tools could be used without restraint to modify any insufficient segmentation results within a maximum time of 120 seconds per measurement. The area-based LAD, SAD, WHO, mLAD, mSAD and EASL were computed from these segmentations.

For volume-based segmentation, additional contours had to be defined from adjacent transverse slices or any perpendicular reconstruction planes [14]. The volume-based LAD, SAD and WHO were derived from the result. Volume-based EASL analysis is not supported by the software and therefore was not performed.

All measurement results from every time point were transferred to a dedicated spreadsheet for further statistical analysis.

Data management

To ensure the comparability of the different parameters, the measurements had to be converted into standard units as described by James et al. [15]. Therefore, all volume and bidimensional measurements were converted into separate diameters as previously published by different groups [16 – 18]. These effective diameters were measured in mm and defined as "volume-equivalent and area-equivalent diameters." The volume-equivalent diameter (D_{vol}) was calculated by inverting the sphere volume for-



▶ **Fig. 1** Example of a semi-automatically determined WHO product with automatically derived LAD and SAD.

▶ **Abb. 1** Beispiel einer semiautomatischen Messung nach WHO mit automatisch ermitteltem LAD und SAD.

mula: $D_{vol} = (6 \times V/\pi)^{1/3}$, whereby V = volume measurement (mm^3) and D_{vol} = diameter (mm). To convert bidimensional measurements into unidimensional measurements, the surface area of a sphere was assumed. By inverting the area formula $A = (\pi \times D_s^2)/4$, whereby A = bidimensional measurement (mm^2) and D_s = diameter (mm), the area-equivalent diameters were calculated using the formula $D_s = 2\sqrt{((1/\pi) \times A)}$.

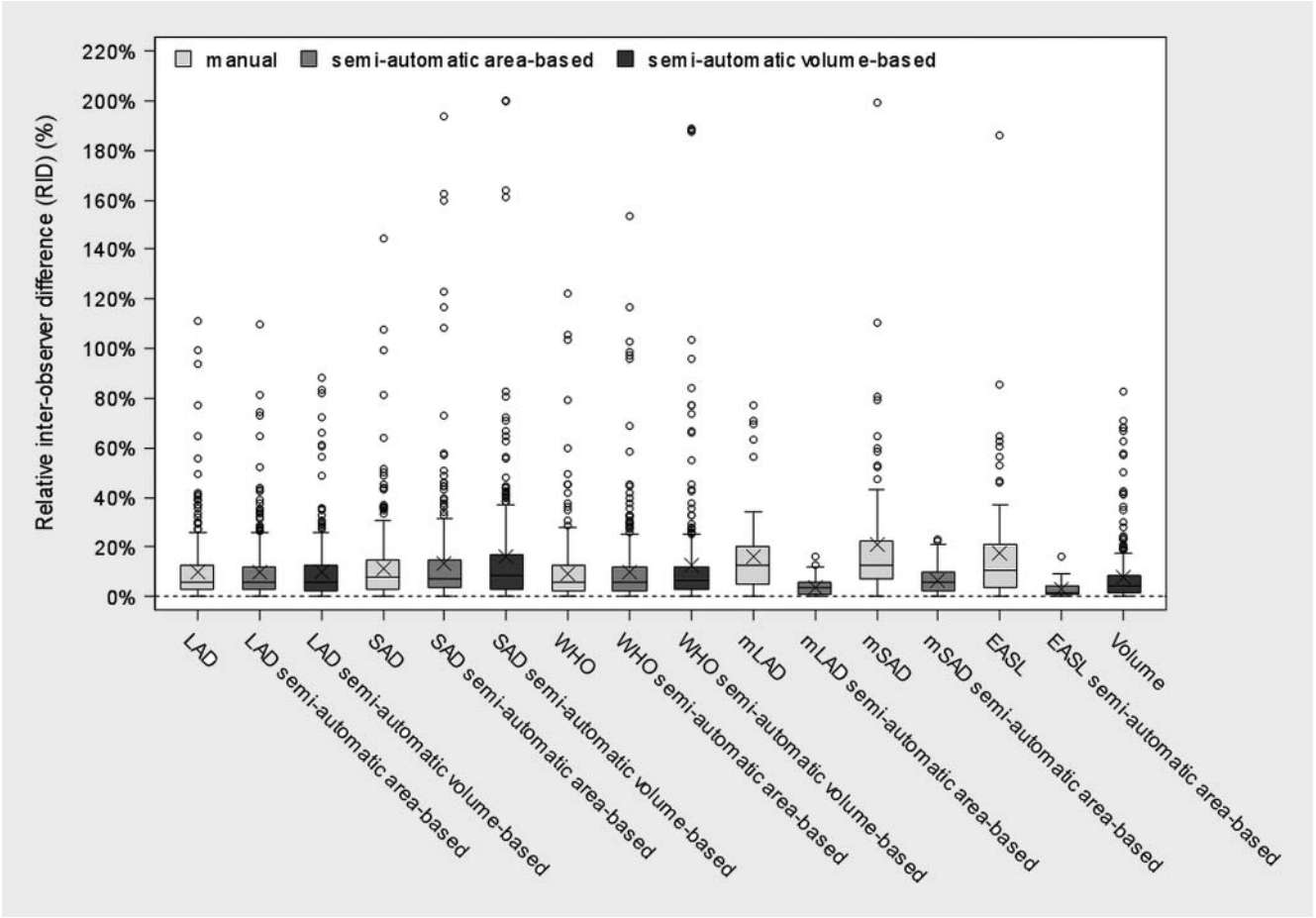
Statistical analysis

Statistical analyses were performed using SAS® software, version 9.4, for Windows (SAS Institute, Cary, NC, USA) and IBM SPSS® Statistics 22 for Windows (IBM Corporation, Somers, NY, USA). Inferential statistics were intended to be exploratory rather than confirmatory. P-values were used to generate new hypotheses and represent only a metric measure of evidence against the respective null hypothesis. Thus, neither a global significance level nor local levels were determined, and no adjustment for multiplicity was made. P-values ≤ 0.05 were considered statistically significant.

Standard descriptive statistical analyses were performed for the parameters LAD, SAD, bidimensional WHO, mLAD, mSAD, bidimensional EASL (all manual and semi-automatic) and volume. Categorical variables are reported as absolute and relative frequencies. Normally distributed continuous variables are reported as mean ± standard deviation, and non-normally distributed continuous variables as median (10 %, 90 % quantile).

To assess interobserver variability between readers R1 and R2 for each parameter, the relative interobserver difference (RID) was determined as RID = | R1 – R2 |/mean (R1, R2) × 100 %.

To determine absolute agreement between the two readers, intraclass correlation coefficients (ICC, two-way random single measure) were calculated for manual and semi-automatic measurements [19]. The ICCs ranged from 0 to 1, whereby values from 0.61 to 0.80 indicate substantial agreement and values from 0.81 to 1 almost perfect agreement.



▶ **Fig. 2** Box plots of the relative interobserver difference (RID) between reader 1 and reader 2 for each manual and semi-automatic area-based and volume-based parameter (long axis diameter [LAD], short axis diameter [SAD], WHO, volume, modified LAD [mLAD], modified SAD [mSAD], EASL). A larger RID and more outliers are found in the vitality-based parameters (mLAD, mSAD, EASL) when determined manually; this can be counteracted by using a semi-automatic area-based approach.

▶ **Abb. 2** Boxplots der relativen Interobserverdifferenz (RID) zwischen Reader 1 und 2 für jeden manuell und semiautomatisch bestimmten flächen- und volumenabgeleiteten Parameter (Längsachsendurchmesser [LAD], Kurzachsendurchmesser [SAD], WHO, Volumen, modifizierter LAD [mLAD], modifizierter SAD [mSAD], EASL). Eine größere RID und mehr Ausreißer können bei den manuell bestimmten Vitalitäts-assoziierten Parametern (mLAD, mSAD, EASL) gefunden werden; dies kann durch ein semiautomatisches Messverfahren reduziert werden.

Results

Lesion characteristics

137 lesions (57 HCCs, 80 metastases) were measured both manually and semi-automatically on baseline and follow-up CT scans in 72 patients, resulting in a total of 691 observations. As EASL, mLAD and mSAD are only applicable to lesions with a hypervascularized portion, fewer lesions were measured in line with these vitality-based criteria.

The medians of the measurements resulting from the manual and semi-automatic methods differ only slightly from each other, regardless of the number of dimensions taken into account (▶ **Table 1**). Slightly higher deviations can be found in the modified RECIST and EASL criteria.

Relative interobserver difference (RID)

The RID (▶ **Table 2,** ▶ **Fig. 2**) – as a measure of divergence between readers R1 and R2 – reveals no statistically significant difference in the established parameters (LAD, SAD, WHO), regardless of the measurement technique, i. e., manual LAD 6.0 % and semi-automatic area-based LAD 5.9 %, manual SAD 7.7 % and semi-automatic area-based SAD 6.9 %.

In contrast, the deviation in the vitality-based criteria (mLAD, mSAD and EASL) is lower in the semi-automatic area-based measurements compared to manual measurements, i. e., manual mLAD 12.5 % and semi-automatic area-based mLAD 3.4 %, manual EASL 10.4 % and semi-automatic area-based EASL 1.8 %. Moreover, the number of outliers is drastically reduced using the semi-automated area-based method of measurement (▶ **Fig. 1**).

The volume can only be determined semi-automatically and has no manually derived equivalent. Its median deviation of 4.1 % is relatively low compared to the other parameters.

▶ **Table 1** Lesion characteristics and number of measurements for manually and semi-automatically derived established (long axis diameter [LAD], short axis diameter [SAD], WHO, volume) and vitality-based (modified LAD [mLAD], modified SAD [mSAD], EASL) parameters. Multidimensional parameters (WHO, EASL and volume) are given as unidimensional-equivalent diameters (mm) for better comparability. Measurements are pooled from both readers and all examinations. Different numbers of measurements result from fewer HCC lesions (established versus vitality-based parameters) and technical limitations in segmentation (area-based versus volume-based parameters).

▶ **Tab. 1** Läsionscharakteristiken und Anzahl der Messungen für manuell und semiautomatisch bestimmte, etablierte (Längsachsendurchmesser [LAD], Kurzachsendurchmesser [SAD], WHO, Volumen) und Vitalitäts-assoziierte (modifizierter LAD [mLAD], modifizierter SAD [mSAD], EASL) Parameter. Mehrdimensionale Parameter (WHO, EASL und Volumen) sind zwecks Vergleichbarkeit als unidimensionales Äquivalent (mm) angegeben. Messungen über beide Reader und alle Untersuchungen. Unterschiedliche Anzahl an Messungen durch geringere Anzahl von HCCs (etablierte versus Vitalitäts-assoziierte Parameter) und technische Limitationen der Segmentierung (flächenbasiert versus volumenbasierte Parameter).

	Lesion size in mm (mm-equivalent for WHO, EASL and volume)		
	Median (10 %, 90 % quantile), n = number of measurements		
Parameter	**Manual**	**Semi-automatic Area-based**	**Semi-automatic Volume-based**
LAD (mm)	33.8 (15.0, 90.0), n = 691	35.3 (16.4, 90.1), n = 691	34.8 (16.4, 86.5), n = 669
SAD (mm)	26.6 (12.3, 69.2), n = 690	27.4 (12.7, 70.3), n = 691	25.9 (12.2, 65.2), n = 669
WHO (mm)	33.8 (15.2, 88.4), n = 690	35.0 (16.2, 89.3), n = 691	33.9 (15.5, 84.2), n = 669
mLAD (mm)	27.2 (12.0, 59.7), n = 206	29.4 (13.9, 61.1), n = 212	
mSAD (mm)	18.6 (7.7, 40.3), n = 206	21.3 (8.6, 43.1), n = 212	
EASL (mm)	25.1 (10.7, 51.7), n = 206	27.7 (12.5, 55.7), n = 212	
Volume (mm)			27.8 (12.8, 64.6), n = 626

▶ **Table 2** Relative interobserver difference (RID) between reader 1 and reader 2 for each manual and semi-automatic area-based parameter (long axis diameter [LAD], short axis diameter [SAD], WHO, volume, modified LAD [mLAD], modified SAD [mSAD], EASL). The RID in the established parameters (LAD, SAD, WHO) reveals no statistically noticeable difference, whereas the RID for vitality-based parameters is lower in the case of semi-automatic measurement.

▶ **Tab. 2** Relative Interobserverdifferenz (RID) zwischen Reader 1 und 2 für jeden manuell und semiautomatisch bestimmten flächengeleiteten Parameter (Längsachsendurchmesser [LAD], Kurzachsendurchmesser [SAD], WHO, Volumen, modifizierter LAD [mLAD], modifizierter SAD [mSAD], EASL). Die RID der semiautomatisch bestimmten, etablierten Parameter (LAD, SAD, WHO) unterscheidet sich von den manuellen Messungen nicht statistisch signifikant, wohingegen die RID der Vitalitäts-assoziierten Parameter (mLAD, mSAD, EASL) deutlich niedriger ist als die der jeweilig manuellen Messungen.

	Relative interobserver difference in % (reader 1 vs. reader 2)	
	Median (10 %, 90 % quantile)	
Parameter	**Manual**	**Semi-automatic Area-based**
LAD	6.0 (1.1, 21.4)	5.9 (1.0, 20.8)
SAD	7.7 (0.9, 24.7)	6.9 (1.0, 27.8)
WHO	5.7 (0.8, 19.1)	5.4 (0.8, 20.2)
mLAD	12.5 (2.1, 31.8)	3.4 (0.4, 8.5)
mSAD	12.7 (1.9, 52.5)	5.7 (0.7, 15.6)
EASL	10.4 (1.4, 45.9)	1.8 (0.4, 6.4)
Volume		4.1 (0.5, 17.0)

▶ **Table 3** Intraclass correlation coefficients (ICC) (two-way random single measure) and 95 % confidence intervals for agreement between reader 1 and 2 in manual and semi-automatic measurements (long axis diameter [LAD], short axis diameter [SAD], WHO, volume, modified LAD [mLAD], modified SAD [mSAD], EASL). Semi-automatic, area-based determination of vitality-based parameters (mLAD, mSAD, EASL) leads to substantially higher agreement (ICC) between reader 1 and 2 compared to manual measurements of the same parameters.

▶ **Tab. 3** Intraclass Korrelationskoeffizient (ICC) (two-way random single measure) und 95 % Konfidenzintervalle der Übereinstimmung zwischen Reader 1 und 2 bezüglich manueller und semiautomatischer Messungen (Längsachsendurchmesser [LAD], Kurzachsendurchmesser [SAD], WHO, Volumen, modifizierter LAD [mLAD], modifizierter SAD [mSAD], EASL). Semiautomatische, flächenabgeleitete Bestimmung Vitalitäts-assoziierter Parameter (mLAD, mSAD, EASL) führt zu einer substanziell höheren Übereinstimmung (ICC) zwischen Reader 1 und 2 im Vergleich zu einer manuellen Bestimmung der selben Parameter.

	Intraclass correlation coefficient (reader 1 vs. reader 2)		
	ICC (95 % CI), n = number of measurements		
Parameter	**Manual**	**Semi-automatic Area-based**	**Semi-automatic Volume-based**
LAD	0.984 (0.980, 0.987), n = 324	0.982 (0.976, 0.986), n = 324	0.976 (0.969, 0.982), n = 303
SAD	0.975 (0.969, 0.984), n = 323	0.958 (0.948, 0.966), n = 324	0.758 (0.706, 0.802), n = 303
WHO	0.984 (0.980, 0.987), n = 323	0.978 (0.973, 0.983), n = 324	0.903 (0.878, 0.923), n = 303
mLAD	0.897 (0.846, 0.932), n = 85	0.997 (0.996, 0.998), n = 105	
mSAD	0.844 (0.770, 0.896), n = 85	0.992 (0.988, 0.995), n = 105	
EASL	0.875 (0.815, 0.917), n = 85	0.998 (0.997, 0.999), n = 105	
Volume			0.987 (0.984, 0.990), n = 284

Intraclass correlation coefficient (ICC)

The ICC – as an indicator of interobserver agreement – is consistently high for the established parameters LAD, SAD and WHO, with no relevant difference between manual and semi-automatic area-based measurements (▶ **Table 3**).

The ICCs from manual measurement of mLAD, mSAD and EASL are lower. The manual parameter with the best correlation, mLAD, has an ICC of 0.897, for example.

Taking the type of lesion – HCCs versus metastases – into account (▶ **Table 4**), there are only small differences regarding the LAD (all ICCs above 0.95). A lower ICC can be found in the SAD of HCCs, especially in the semi-automatic 3 D measurements (3 D SAD 0.780).

Discussion

Manual radiological measurements in CT examinations are an established clinical approach and form the basis for any evaluation of imaging in oncology. Nevertheless, numerous recent studies have demonstrated lower interobserver variability and higher reproducibility with semi-automatic measurements [12, 13, 17, 20 – 24]. These advantages permit more reliable and accurate classification of the therapeutic response and directly influence treatment decisions. These studies are limited in that they mostly focused on the relatively easy task of lung nodule [17, 22, 23] or lymph node segmentation in CT examinations [12, 13].

The segmentation of liver lesions in MRI examinations is also firmly established and usually involves a semi-automatic, volume-based approach [25, 26]. On the other hand, reliable segmentation of liver lesions in CT – with its lower soft-tissue-contrast – is a more demanding task and has been addressed only recently [24, 27]. Special challenges are posed by the initially variable morphology which changes over the course of new targeted und endovascular therapies due to decreased tumor vascularization with subsequently reduced contrast enhancement or even necrosis. In light of these unavoidable hindrances, the mode of measurement (manual or semi-automatic) should not add any further uncertainty.

Our data reveal a consistently high level of measurement precision (reflected by the ICC) for any semi-automatically derived measurements, including the vitality-based parameters mLAD, mSAD and EASL. In contrast, the precision of the manual measurements of these vitality-based parameters is considerably lower. As the ICC does not mainly depend on the number of cases, this could be explained at least in part by the smaller area to be measured with a consecutively higher variation.

One possible explanation for the higher ICCs of the semi-automatically derived measurements is that the standardized semi-automatic workflow offers guidance (e. g. by proposing reconstruction planes or boundaries) in difficult situations, counteracting the lesion- and therapy-dependent variations and leading to less variation.

This advantage is not expected to come to the fore in the relatively easy task of generally determining lesion size. Our data ultimately reveal no relevant differences in precision between manual and semi-automatic measurements for the established parameters LAD, SAD and WHO, regardless of the lesion type (HCCs versus metastases).

In this regard our results are consistent with previous studies that report a higher ICC for semi-automatic CT measurements of lymph nodes [13] and pulmonary nodules [22, 23] and extend the

▶ **Table 4** Intraclass correlation coefficients (ICC) by tumor entity (two-way random single measure) and 95 % confidence limits for agreement between reader 1 and 2 in manual and semi-automatic measurements (long axis diameter [LAD], short axis diameter [SAD], WHO, volume, modified LAD [mLAD], modified SAD [mSAD], EASL). The mLAD, mSAD and bidimensional EASL were not applicable to metastases and have therefore been omitted. The ICC is equally high for LAD regardless of the measurement method.

▶ **Tab. 4** Intraclass Korrelationskoeffizient (ICC) nach Tumorentität (two-way random single measure) und 95 % Konfidenzintervalle der Übereinstimmung zwischen Reader 1 und 2 bezüglich manueller und semiautomatischer Messungen (Längsachsendurchmesser [LAD], Kurzachsendurchmesser [SAD], WHO, Volumen, modifizierter LAD [mLAD], modifizierter SAD [mSAD], EASL). Die Parameter mLAD, mSAD, und bidimensionaler EASL sind per Definition nicht auf Metastasen anwendbar und entfallen daher in der Betrachtung. Der ICC für LAD ist über alle Messverfahren konstant hoch.

	Intraclass correlation coefficient (reader 1 vs. reader 2)					
	ICC (95 % CI)					
		HCC			Metastases	
Parameter	Manual (n = 131)	Semi-automatic Area-based (n = 132)	Semi-automatic Volume-based (n = 129)	Manual (n = 192)	Semi-automatic Area-based (n = 192)	Semi-automatic Volume-based (n = 174)
LAD	0.962 (0.946, 0.973)	0.957 (0.938, 0.970)	0.962 (0.945, 0.974)	0.992 (0.989, 0.994)	0.991 (0.988, 0.993)	0.988 (0.983, 0.991)
SAD	0.856 (0.937, 0.969)	0.898 (0.860, 0.927)	0.780 (0.693, 0.843)	0.984 (0.979, 0.988)	0.989 (0.985, 0.992)	0.747 (0.673, 0.806)
WHO	0.965 (0.950, 0.975)	0.943 (0.921, 0.960)	0.817 (0.741, 0.871)	0.992 (0.990, 0.994)	0.994 (0.992, 0.995)	0.971 (0.961, 0.978)

applicability to the semi-automatic evaluation of liver lesions in CT. Analogous results were published recently with a focus on MRI evaluation of liver lesions after intra-arterial therapy [28, 29].

For the probably more demanding CT segmentation and measurement, a current publication evaluated HCCs under systemic molecular-targeted therapy [30]. In contrast to this study, we applied therapy (TACE or SIRT) selectively to the liver arteries that could make a difference in the homogeneity and intensity of the therapy effects, making measurements even more difficult.

As an additional benefit, a semi-automatic workflow facilitates standardized and complete documentation [31]. This helps reduce measurement time in follow-up examinations by a third, compensating for the slightly longer, initial segmentation time [32]. Furthermore, it offers a systematic overview and guidance in patients with multiple examinations, possibly at different sites, thereby permitting monitoring of a diversified therapeutic spectrum.

Limitations

This study is limited to the extent that the measured liver lesions were not excised. Thus, the actual size and, depending on this, the accuracy could not be determined. However, a surgical intervention would not have been justified, and the precision – as a relative measure – is not influenced by our approach, which is accepted for in-vivo studies [17].

Furthermore, we chose a single-center, retrospective study design. Because the focus of interest was the measurement agreement between the readers, each measurement was regarded as independent, thus potentially disregarding correlations between lesions in the same patient and between different time points.

We did not evaluate the mean segmentation time which could pose a bias due to over-accurate editing of contours. To prevent this we restricted the maximum segmentation time to 120 seconds per lesion [24].

The CT scanner and reconstruction protocols were kept constant to the disadvantage of limited generalizability.

Conclusion

We conclude that vitality-based tumor measurements of hepatocellular carcinomas and metastases after transarterial local therapies should be performed semi-automatically due to greater measurement precision, thus increasing the reproducibility and, in turn, the reliability of therapeutic decisions. Manual and semi-automatic measurements of established parameters offer the same level of precision, but preference should be given to the semi-automatic approach due to the possibility of generating systematic documentation.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] Primrose JN. Surgery for colorectal liver metastases. Br J Cancer 2010; 27: 1313 – 1318

[2] Raza A, Sood GK. Hepatocellular carcinoma review: current treatment, and evidence-based medicine. World J Gastroenterol 2014; 20: 4115 – 4127

[3] Theysohn JM, Ertle J, Müller S et al. Hepatic volume changes after lobar selective internal radiation therapy (SIRT) of hepatocellular carcinoma. Clin Radiol 2014; 69: 172 – 178

[4] Van de Wiele C, Maes A, Brugman E et al. SIRT of liver metastases: physiological and pathophysiological considerations. Eur J Nucl Med Mol Imaging 2012; 39: 1646 – 1655

[5] Eisenhauer EA, Therasse P, Bogaerts J et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer 2009; 45: 228 – 247

[6] World Health Organization. WHO handbook for reporting results for cancer treatment. Geneva, 1979 http://www.whqlibdoc.who.int/offset/WHO_OFFSET_48.pdf

[7] Tirkes T, Hollar MA, Tann M et al. Response criteria in oncologic imaging: review of traditional and new criteria. Radiographics 2013; 33: 1323 – 1341

[8] Layer G, Stahl T, Hoffend J. Bildgebende Beurteilung des Therapieansprechens unter Chemotherapie. Radiologie up2date 2013; 13: 221 – 239

[9] 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. J Hepatol 2001; 35: 421 – 430

[10] Lencioni R, Llovet JM. Modified RECIST (mRECIST) assessment for hepatocellular carcinoma. Semin Liver Dis 2010; 30: 52 – 60

[11] Weßling J, Puesken M, Koch R et al. MSCT follow-up in malignant lymphoma: comparison of manual linear measurements with semi-automated lymph node analysis for therapy response classification. Rofo 2012; 184: 795 – 804

[12] Buerke B, Gerss J, Puesken M et al. Usefulness of semi-automatic volumetry compared to established linear measurements in predicting lymph node metastases in MSCT. Acta Radiol 2011; 52: 540 – 546

[13] Buerke B, Puesken M, Müter S et al. Measurement accuracy and reproducibility of semiautomated metric and volumetric lymph node analysis in MDCT. Am J Roentgenol 2010; 195: 979 – 985

[14] Fetzer A, Meinzer HP, Heimann T. Interaktive 3D Segmentierung auf Basis einer optimierten Oberflächeninterpolation mittels radialer Basisfunktionen. In: , (eds) Tolxdorff T, et al. Bildverarbeitung für die Medizin 2012, Informatik aktuell. Berlin Heidelberg: Springer, 183 – 188

[15] James K, Eisenhauer E, Christian M et al. Measuring response in solid tumours: unidimensional versus bidimensional measurement. J Natl Cancer Inst 1999; 91: 523 – 528

[16] Fabel M, von Tengg-Kobligk H, Giesel FL et al. Semi-automated volumetric analysis of lymph node metastases in patients with malignant melanoma stage III/IV – a feasibility study. Eur Radiol 2008; 18: 1114 – 1122

[17] Zhao B, Schwartz LH, Moskowitz CS et al. Lung cancer: computerised quantification of tumour response – initial results. Radiology 2006; 241: 892 – 898

[18] Höink AJ, Weßling J, Koch R et al. Comparison of manual and semi-automatic measuring techniques in MSCT scans of patients with lymphoma: a multicentre study. Eur Radiol 2014; 24: 2709 – 2718

[19] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 1979; 86: 3420 – 3428

[20] Wulff AM, Bolte H, Fischer S et al. Lung, liver and lymph node metastases in follow-up MSCT: comprehensive volumetric assessment of lesion size changes. Rofo 2012; 184: 820 – 828

[21] Höink AJ, Heindel W, Buerke B. Radiological Evaluation of the Therapeutic Response of Malignant Diseases: Status Quo, Innovative Developments and Requirements for Radiology. Rofo 2014; 186: 927 – 936

[22] Bolte H, Jahnke T, Schäfer FK et al. Interobserver-variability of lung nodule volumetry considering different segmentation algorithms and observer training levels. Eur J Radiol 2007; 64: 285 – 295

[23] Dinkel J, Khalilzadeh O, Hintze C et al. Inter-observer reproducibility of semi-automatic tumour diameter measurement and volumetric analysis in patients with lung cancer. Lung Cancer 2013; 82: 76 – 82

[24] Wang Z, Chapiro J, Schernthaner R et al. Multimodality 3D Tumour Segmentation in HCC Patients Treated with TACE. Acad Radiol 2015; 22: 840 – 845

[25] Tacher V, Lin M, Duran R et al. Comparison of Existing Response Criteria in Patients with Hepatocellular Carcinoma Treated with Transarterial Chemoembolization Using a 3D Quantitative Approach. Radiology 2016; 278: 275 – 284

[26] Chapiro J, Lin M, Duran R et al. Assessing tumour response after locoregional liver cancer therapies: the role of 3D MRI. Expert Rev Anticancer Ther 2015; 15: 199 – 205

[27] Yan J, Schwartz LH, Zhao B. Semiautomatic segmentation of liver metastases onumetric CT images. Med Phys 2015; 42: 6283

[28] Bonekamp D, Bonekamp S, Halappa VG et al. Interobserver agreement of semi-automated and manual measurements of functional MRI metrics of treatment response in hepatocellular carcinoma. Eur J Radiol 2014; 83: 487 – 496

[29] Budjan J, Sauter EA, Morelli JN et al. Semi-automatic Volumetric Measurement of Treatment Response in Hepatocellular Carcinoma After Trans-arterial Chemoembolization. Anticancer Res 2016; 36: 4353 – 4358

[30] Telegrafo M, Dilorenzo G, Di Giovanni G et al. Follow-up of multicentric HCC according to the mRECIST criteria: role of 320-Row CT with semi-automatic 3D analysis software for evaluating the response to systemic therapy. G Chir 2017; 37: 206 – 210

[31] Dankerl P, Cavallaro A, Uder M et al. [Automatic segmentation and annotation in radiology]. Radiologe 2014; 54: 265 – 270

[32] Moltz JH, D'Anastasi M, Kiessling A et al. Workflow-centred evaluation of an automatic lesion tracking software for chemotherapy monitoring by CT. Eur Radiol 2012; 22: 2759 – 2767