Journal Pre-proof

Interobserver variability of Gross Tumour Volume delineation for colorectal liver metastases using CT and MRI

Cora Marshall MSc , Pierre Thirion MD , Alina Mihai MD , John G. Armstrong Prof MD , Seán Cournane PhD , Darina Hickey MSc , Brendan McClean PhD , John Quinn PhD

PII: S2452-1094(22)00126-9
DOI: https://doi.org/10.1016/j.adro.2022.101020
Reference: ADRO 101020

To appear in: *Advances in Radiation Oncology*

Received date: 27 October 2021
Accepted date: 28 June 2022

Please cite this article as: Cora Marshall MSc , Pierre Thirion MD , Alina Mihai MD , John G. Armstrong Prof MD , Seán Cournane PhD , Darina Hickey MSc , Brendan McClean PhD , John Quinn PhD , Interobserver variability of Gross Tumour Volume delineation for colorectal liver metastases using CT and MRI, *Advances in Radiation Oncology* (2022), doi: https://doi.org/10.1016/j.adro.2022.101020

Interobserver variability of Gross Tumour Volume delineation for colorectal liver metastases using CT and MRI

Cora Marshall MSc(1)(2), Pierre Thirion MD(2)(3), Alina Mihai MD(2), John G Armstrong Prof MD(2), Seán Cournane PhD(1) , Darina Hickey MSc(2), Brendan McClean PhD (1)(3),John Quinn PhD(1)
(1)School of Physics, University College Dublin
(2)Beacon Hospital, Dublin
(3)St Luke's Radiation Oncology Network
Corresponding Author Name & Email Address
Cora Marshall email: marshallcora@gmail.com
Funding Statement
Funding: None
Data Availability Statement for this Work

Research data are not available at this time
Author responsible for Statistical analysis
Cora Marshall **email: marshallcora@gmail.com**

Conflicts of interest for authors:
Conflicts : None

Abstract

Purpose: The purpose of this study was to evaluate the interobserver variability in the contouring of the gross tumour volume (GTV) on magnetic resonance imaging (MRI) and computed tomography (CT) for colorectal liver metastases in the setting of stereotactic ablative radiotherapy (SABR).

Methods and Materials:

Three expert radiation oncologists contoured 10 GTV volumes on 3 MRI sequences and on the CT image dataset. Three metrics were chosen to evaluate the interobserver variability: the conformity index, the DICE coefficient and the maximum Hausdorff Distance (HDmax). Statistical analysis of the results was performed using a one sided permutation test.

Results:

For all three metrics, the MR LAVA showed the lowest interobserver variability. Analysis showed a significant difference ($p<0.01$) in the mean DICE, an overlap metric, for MR LAVA (0.82) and CT (0.74). The HDmax which highlights boundary errors also showed a significant difference ($p=0.04$) with MR LAVA having a lower mean HDmax (7.2mm) compared to CT (5.7mm). The mean HDmax for both MR SSFSE (19.3mm) and DWI (9.5mm) showed large interobserver variability with MR SSFSE having a mean HDmax of 19.3mm. A volume comparison between MR LAVA and CT showed a significantly higher volume for small GTVs (<5cc) when using MR LAVA for contouring in comparison to CT.

Conclusions: This study reported the lowest interobserver variability for the MR LAVA, thus indicating the benefit of using MR to complement CT when contouring GTV for colorectal liver metastases.

Introduction

Stereotactic ablative radiotherapy (SABR) is an external beam radiotherapy technique which uses precise targeting to deliver high doses of radiation capable of ablating tumours directly [1]. Treating primary or secondary liver malignancies with these ablative doses has become possible with the emergence of image-guided radiotherapy and respiratory management. The delivery of radiation to reduced planning target volumes (PTVs) allows for functional liver, away from the target area, to be spared [2].

As a result, SABR is increasingly used in the management of liver metastases, with clinical series reporting promising 2 year local control rates, of approximately 90% [3]. Studies have shown that Liver SABR could have a major role in treating colorectal cancer patients, for whom the liver is the dominant metastatic site. In some cases, particularly patients with oligometastatic disease [4, 5] when there are a limited number of tumours, up to five in the liver, the aim is to eradicate the disease completely in liver.

Due to the steep dose gradients in SABR treatments, the accurate determination of the gross tumour volume (GTV) is a crucial step. However, it is widely accepted that this step of delineation of the GTV by the radiation oncologist is subject to interobserver variability [6]**.** While numerous studies have evaluated interobserver variability, a recent review of 119 studies [7] has identified only one which has examined interobserver variability in liver cancer.

In liver SABR the precise delineation of the GTV is challenging due to the poor soft tissue contrast of Computed Tomography (CT) and the limited literature identifying pathological correlation with radiological features. Despite these limitations, CT remains the clinical standard for volume delineation in radiotherapy; however, other modalities are increasingly being utilised and showing promise. Magnetic Resonance imaging (MRI) is now considered the gold standard for delineation of brain tumours [2] for stereotactic treatments, offering superior soft-tissue contrast to that of CT imaging. Furthermore, the use of MRI for the delineation of abdominal tumours has also been reported to be increasing [6].

According to ICRU 83 [8] , a clinical margin is added to the GTV to determine the PTV. Random and systematic uncertainties do not have an equal effect on the dose distribution. Random errors cause a blurring of the dose distribution where systematic errors cause a shift of the cumulative dose distribution. Interobserver variability is considered a systematic error. The reduction in such errors should be optimized to prevent inadvertent irradiation of normal tissues, particularly in high-dose treatments.

The primary objective of this study was to evaluate the interobserver delineation variation for colorectal liver metastases for SABR when using CT-based GTV delineation and MR-based delineation for a number of MR sequences. In addition, we aimed to establish which MR sequence yielded the lowest interobserver variability.

Materials and Methods

The study was approved by the institutional clinical audit committee of the institution.

Patient database / eligibility

An anonymized database was created from 7 patients with metastatic colorectal cancer having attended our institution for liver SABR, representing a total of 10 lesions. Eligible cases had to have completed both CT simulation and MRI simulation for a number of sequences outlined below. Information on the GTVs delineations are contained in table 1 below. The location of each GTV is given in reference to the Couinaud classification of liver anatomy, commonly used in radiology reporting.

(**Table 1** here)

MRI and CT acquisition and characteristics

The MRI imaging was carried out using a 1.5T GE Signa HDxT in the radiology department. The MRI protocol included a T1 contrast-enhanced sequence called Liver Acquisition Volume Acquisition (LAVA), a non-contrast enhanced Single Shot Fast Spin Echo (SSFSE) and a Diffusion Weighted Image (DWI). The LAVA and SSFSE sequences were taken on a voluntary end expiration breathhold. The MRI, for planning purposes, is typically acquired immediately after the simulation CT with both acquired at end-expiration breathhold in order to improve image registration. The DWI was a respiratory-gated sequence rather than breathhold. The end phase of expiration was chosen for the gate. Due to irregularity in some patients breathing, only six patients had DWI sequences.

The volume of contrast administered for the LAVA sequence was determined according to 0.1 mL/kg body weight (0.1 mmol/kg) for each patient and images were acquired at four phases of contrast enhancement, (i) non contrast, (ii) arterial enhancement at 20 seconds post injection (iii) portal-venous enhancement is approximately 70 seconds post injection and (iv) a delayed contrast phase. The target appearance on a contrast enhanced T1 sequence such as LAVA includes a central hypoattenuating portion that corresponds to the central necrosis often surrounded by an ill-defined

enhancing rim, which corresponds to the proliferative tumoral border. Delayed enhancement may also be present due to the desmoplastic reaction.

The LAVA sequence is a T1 fat-saturated 3d acquisition. This is a fast sequence with the aim of acquiring the whole liver within one breathhold. The LAVA sequence had a slice thickness of 2.5mm. The DWI was acquired with b values of 50 and 800. The SSFSE and the DWI sequences were low resolution scans with slice thicknesses of 8 mm, and would not be used in isolation for GTV delineation. An example of the appearance of each image set can be seen in figure 1 below.

(Figure 1 here**)**

The CT simulation was acquired on a GE Lightspeed RT. The scans were taken at 60 seconds post contrast in end-expiration breathhold. The contrast was Omnipaque™ with a concentration of 70-80 mls and a flow rate of 1.5 to 1.7 ml/sec. Contrast was not varied with patient□s weight. Seven of the scans have 2.5 mm slice thickness, two have 5mm slice thickness and one had 1.25 mm slice thickness.

Delineations

The contouring process included 2 steps:

Firstly, each case was reviewed by a senior radiologist (>10 èŜI·№ă□ ŜÈŁJŜ№ŜyŎŜ) who chose the most appropriate contrast-enhanced sequence for the delineation. Delineation instructions were provided for each GTV. The instructions included (i) slice visible (ii) estimate of tumour volume dimension (iii) appearance on the image e.g. dark in respect to surrounding parenchyma.

Contour analysis

Due to the irregular shapes of tumours, evaluating both the overlap and the boundary differences between the GTV delineations are important [9]. Three metrics were chosen: the conformity index, the DICE coefficient and the maximum Hausdorff Distance (HDmax) [10]. All analyses were conducted using SlicerRT 4.10.2 [11].

The conformity index is the ratio of the common volume of all three GTVs to the encompassing volume of all three radiation oncologists GTVs [12].

The DICE coefficient is also an overlap-based metric, a pairwise comparison of each 2 delineation was performed (i.e. interobserver 1 to interobserver 2, interobserver 2 to 3 and interobserver 1 to 3). The DICE ratio is the ratio of the common volume to the encompassing volume and varies from 0 (no overlap) to 1 (complete overlap).

The maximum Hausdorff Distance (HDmax) is a spatial distance metric to take into account boundary errors in the delineation [10]. The undirected is measured as the HDmax distance from boundary X to Y or from boundary Y to X. Slicer 4.10.2 segment comparison Y 2 gives the undirected HDmax, which is considered in 3D for the delineations. A pairwise HDmax was performed for each GTV delineated.

Statistical analysis

Both the Conformity index and the DICE Coefficient range from 0 to 1, with less interobserver variability as the metric approaches 1. The resultant data, where no manipulation of the data is carried out, is not normally distributed. A -test was therefore not appropriate.

The Hausdorff Distance is a distance metric where lower values demonstrate lower interobserver variability, yielding data which is not normally distributed. Thus, significance of the difference in means of the DICE, HDmax and the conformality index were analysed using a one sided non-parametric permutation test, following Ernst [13].

In this one-sided test, the observed datasets were resampled and the difference in the parameter to be tested (in this case the mean) of the resampled sets was calculated. As the number of combinations can be large, (30 MR LAVA and 27 CT amounted to 1.4×10^{16} combinations) a Monte Carlo approach was used to evaluate n permutations. An n of 100,000 was used for the DICE and

HDmax. The p-value of the test is the number combinations where the difference in the mean is equal to or greater than the measured mean difference, divided by the number of samples.

A p-value of <0.05 was considered statistically significant.

Comparison of CT to MR LAVA

The ratio of the volume of the GTV delineated by each observer on the MR LAVA and the CT was evaluated. To compare the delineations, a registration between the CT and MR was performed. A rigid registration using Eclipse version 15.5 was used to register the images in the area of the GTV. Surrounding vessels were used as a guide for the registration. Each registration was checked by a second experienced physicist, by checking the anatomy in proximity to the tumour, most commonly using vessels. In one case, where a large deformation was observed, a deformable registration was required. Varian Medical Systems Velocity 4.1 program was used for deformable image registration.

Margin

The PTV in ICRU 83 is a geometric concept, where by adding a margin on the GTV/ CTV we are delivering a clinically accepted probability adequate dose to the GTV. All geometric uncertainties are included, including respiratory motion. Our liver SABR treatments are conducted in end-expiration breathhold, eliminating the impact of respiratory motion.

Several mathematical formulae have been recommended for generating the GTV-PTV margins. In this study we used the van Herk recipe [14] to demonstrate the difference in the margin required based on the interobserver variability seen with MR LAVA and CT. To ensure that the minimum dose of 95% to the GTV to 90% of the patients, the Van Herk margin recipe (н♦ρ1 ь л♦т`) is used, which requires a margin that is 2.5 times the total standard deviation of the systematic errors (1) and 0.7 times the standard deviation of the random errors (`).

Using Varian Systems Velocity 4.1 software package, the mean distance between the boundary of the GTVs for the MR LAVA and the contrast-enhanced CT was evaluated. The package computes the mean value of the closest point from one boundary to the closest point on the second boundary volume. To determine the margin difference, 2.5 times the total standard deviation of this boundary distance was determined.

Results

Graphical representations of the pairwise DICE similarity coefficient and the pairwise HDmax are shown in Figure 2 and Figure 3. The conformity index is summarised in Table 2 below. The MR LAVA showed less interobserver variation than the CT, MR SSFSE or the DWI. The overall mean DICE coefficients for the MR LAVA, CT, MR SSFSE and DWI were 0.82, 0.74, 0.55 and 0.76 respectively (Table 2). The overall mean HDmax for the MR LAVA, CT, MR SSFSE and DWI were 5.68mm, 7.25mm, 19.34mm and 9.51mm respectively. Similarly, the overall mean conformity indices for MR LAVA, CT, MR SSFSE and DWI were 0.58, 0.47, 0.29 and 0.46.

For all three metrics, the MR LAVA shows the lowest interobserver variability. The CT with contrast has a slightly lower mean DICE than the DWI, but the mean HDmax and mean conformity index was lower for the CT with contrast. A summary of this data is available in table 3.

From figure 3 and 4, large variability in contouring on the non-contrast SSFSE was evident, with GTV 5 and GTV 7 having no overlap in the contouring, giving DICE values of 0. In addition, the average of the HDmax for MR SSFSE was 19.34mm, with values ranging from 2.7mm to 47mm. From the limited number of DWI datasets, the mean DICE was slightly higher than CT at 0.76, but the HDmax (9.51mm) and conformity index (0.46) indicated more variability in contouring.

Interobserver variability can be accounted for in the planning margin on the GTV as a systematic error. The pairwise mean distance between the boundary of the GTVs delineated on CT and MR LAVA was 1.8mm and 1.3mm, respectively. With a standard deviation on the mean of 1.6mm for CT

and 1.2mm for MR LAVA, the resulting margins, following the Van Herk formula [14], required to account for interobserver variability would be 4mm (CT) and 3.1mm (MR LAVA).

Permutation Test

The permutation test results are show in table 4. A statistically significant difference ($p < 0.01$) was found between the mean DICE for CT (0.74) and MR LAVA (0.82). The mean HDmax for CT (7.25mm) and mean HDmax MR LAVA (5.68mm) were also found to be significantly different ($p = 0.04$). The difference in mean conformity index of CT (0.47) and MR LAVA (0.58) was not found to be statistically significant ($p = 0.08$).

(Table 4 here)

MR LAVA to CT comparison

Figure 4 is a graphical representation of the ratio of the volume of GTV delineated on MR LAVA to CT for each observer in order of GTV volume. Each of the observers GTV delineations on CT were compared to MR LAVA, 68% of volumes drawn on MR LAVA are larger than on CT ($p<0.01$). By dividing the volumes into those with a cc of less than 5, it was shown that the effect is more significant for small GTVs. In this case, 87% of GTVs with a volume of 5cc or less were smaller on CT than on MR LAVA ($p<=0.01$), while 53% of those greater than 5cc were smaller on CT ($p=0.57$). All of the MR LAVA scans are 2.5mm slice thickness, seven of the CT scans are 2.5mm, however GTV 4 and GTV 5 are 5mm slice thickness. Given the size of GTV5, reported by radiology as 2cm, the resolution along the Z axis (superior/inferior), a finer slice thickness would be appropriate.

Discussion

Interobserver variability in delineation of the GTV is a widely accepted source of uncertainty in radiotherapy and has a direct effect on the GTV to PTV margin. In this study we examined the interobserver variability on a range of image sets with the aim of determining the most appropriate

image set for GTV delineation. A secondary aim was to compare the GTVs delineated on MR to those on CT.

A thorough analysis of the interobserver variability in delineation was achieved by using a range of metrics which consider both the overlap ratio and the boundary differences. The analysis showed MR LAVA had the lowest interobserver variability when compared to CT, MR SSFSE and MR DWI. Two of the metrics used, the HDmax and the DICE coefficient showed a statistically significant improvement in the interobserver variability on MR LAVA when compared to CT.

SSFSE is a very fast imaging sequence and is used in body imaging where bowel and respiratory motion are an issue. However, this results in images with lower signal to noise, blurring and reduced image contrast. The large interobserver variability found in this study for SSFSE is not unexpected and while useful for diagnostic purposes, this study finds that the variability renders it unsuitable for use in radiotherapy as a delineation image set.

There are few studies which examine interobserver variability of GTV delineation in the liver. One such study, by Jensen et al.[7], included patients with hepatocellular carcinoma (n=6) and metastatic liver tumours (n=6), while the observers included two radiation oncologists, two radiation therapists and a radiology resident. The volumes were delineated on a dynamic contrast enhanced CT and a 4D-CT with the analysis including the DICE coefficient but with no boundary difference metrics. As such, the results presented by Jenson et al. [7] were not directly comparable to this study as it used different image sets, along with a more varied patient group and observer set.

The results of this study allow for the accurate estimate of the systematic error introduced by the interobserver variability, which is added to the margin recipe for calculation of the planning target volume (PTV). The margin adds a buffer to account for the uncertainties in the delineation of the GTV (ICRU 83 [8]). This study yielded a reduction of the interobserver variability from 1.6cm (SD) for CT to 1.2cm (SD) for MR LAVA.

Steenbakkers et al. [12] studied the impact on interobserver variability for lung cancer delineation using PET-CT in comparison to CT alone. The overall interobserver variability was reduced from 1cm (SD) to 0.4cm (SD) when using CT vs PET-CT alone. This much lower interobserver variability in lung than liver can be expected considering the less well-defined boundaries and artifacts due to bowel and respiratory motion in liver. PET-CT can be useful in highlighting a Biological Target Volume in liver SBRT. However, Riou et al. [15] in their study of the benefit of 4D-PET CT in volume delineation for liver SBRT, found that non-respiratory gated PET in the liver can result in a possible underestimation or a complete miss of the target volume.

By introducing MRI as an image set for delineation, the interobserver variability is reduced but this study also saw a significant difference in the volume of the GTV delineated on MRI in comparison to CT for small tumours. For the LAVA sequence, when GTVs delineated were 5cc or less, the volume delineated on MRI was larger in 87% of cases, with a mean ratio of MRI volume to CT volume of 2.52. Previous studies have investigated the differences in CT and MR delineation. Pech et al.[16] studied 25 patients with 43 colorectal liver metastases. Similar to our study they reported that the volume on contrast enhanced CT (mean volume = 20ml) was less than that on the T1 weighted contrast enhanced MRI sequence (mean volume = 65ml). The PV phase of CT contrast enhancement was used in this study.

A limitation of these studies is the lack of literature currently available which compares imaging to histopathology. These studies are technically difficult, specifically in the preparation of the specimen. The histopathology correlation of T1 weighted images was studied by Outwater et al in 1991 [17]. This study reported low intensity regions corresponded to histologic findings of coagulative necrosis and desmoplasia within the tumor. The study also found that peripheral hyperintense halos around central hypointense areas encompassed the growing tumor margin and variable degrees of cell necrosis. Another matter for consideration is whether microscopic tumour beyond the macroscopic tumour can be depicted with imaging [18]. Traditionally, in stereotactic

radiotherapy a clinical target volume (CTV) margin for microscopic extension is not used. However, there is debate in the case of the liver, with some clinical groups adding up to 8mm CTV margin [19]. Pech et al. [16] proposed that the contrast enhancing tissue is more at risk of carrying tumour cells and by including this area on contrast enhancement on the MRI in the GTV, the CTV is included.

The AAPM [2] and UK SABR [20] consortium recommend CT and MRI for delineation of tumour volumes. We routinely employ MR imaging for tumour delineation in our clinic and, indeed, a range of MR sequences had been presented for radiation oncologist delineation until the completion of this study. With evidence from this work, the number of acquired MR sequences has been significantly reduced, eliminating the use of SSFSE in most cases while focussing on the MR LAVA sequence, which returned the lowest interobserver variability. As a result, the abridged imaging protocols have led to time savings on the MRI scanner with a resultant increased efficiency within the radiology department. Further work is required to investigate the interobserver variability when using the DWI as we had a limited number of datasets available. However, this study highlighted the potential for improvements in the MR DWI resolution, an investigation which, in collaboration with the radiology department, is ongoing.

When using MRI in conjunction with CT for treatment planning, registration of the images is required which may introduce delineation errors, especially in the case of the liver. It is, thus, imperative to employ deformable registration. Varoney et al [19] showed the need for deformable registration, demonstrating how the error can be magnified for smaller tumours in cases where the deformable registration it is not used. According to AAPM TG 132 [21], an estimation of this error should be taken into account in margin recipes.

Reducing the interobserver variability in liver stereotactic radiosurgery is desirable to reduce margins and allow a therapeutic ratio necessary for tumour ablation. MR LAVA provided the lowest interobserver variability of the image-sets studied. There may be a systematic error introduced for smaller tumours where MR is not used for delineation. The limited sample size of this study means

that the investigation is exploratory in nature. Further work would be required to assess any systematic difference in the delineation of small tumours on MR LAVA images as compared to CT. Nevertheless, studying the interobserver variability informed on the target margin necessary for accounting for such variability, and may help in determining improvements in treatment precision and standardisation. The addition of automatic segmentation techniques may further assist in standardising tumour delineation. Indeed, the recent literature indicates that there has been significant advances in tumour delineation using of neural networks [22, 23].

Conclusion

The use of magnetic resonance imaging to complement CT in the delineation of the target in the treatment of colorectal liver metastases with SABR gives an advantage by significantly reducing the interobserver variability. The magnetic resonance sequence which shows the least variability in delineation of the target was the MR LAVA.

References

1. Jaffray, D.A., *Image-guided radiotherapy: from current concept to future perspectives.* Nat Rev Clin Oncol, 2012. **9**(12): p. 688-99.
2. Benedict, S.H., et al., *Stereotactic body radiation therapy: The report of AAPM Task Group 101.* Medical Physics, 2010. **37**(8): p. 4078-4101.
3. Rusthoven, K.E., et al., *Multi-institutional phase I/II trial of stereotactic body radiation therapy for liver metastases.* J Clin Oncol, 2009. **27**(10): p. 1572-8.
4. Palma, D.A., et al., *Stereotactic ablative radiotherapy versus standard of care palliative treatment in patients with oligometastatic cancers (SABR-COMET): a randomised, phase 2, open-label trial.* Lancet, 2019. **393**(10185): p. 2051-2058.
5. Palma, D.A., et al., *Stereotactic Ablative Radiotherapy for the Comprehensive Treatment of Oligometastatic Cancers: Long-Term Results of the SABR-COMET Phase II Randomized Trial.* J Clin Oncol, 2020. **38**(25): p. 2830-2838.
6. Vinod, S.K., et al., *Uncertainties in volume delineation in radiation oncology: A systematic review and recommendations for future studies.* Radiother Oncol, 2016. **121**(2): p. 169-179.
7. Jensen, N.K., et al., *Dynamic contrast enhanced CT aiding gross tumor volume delineation of liver tumors: an interobserver variability study.* Radiother Oncol, 2014. **111**(1): p. 153-7.
8. Hodapp, N., *[The ICRU Report 83: prescribing, recording and reporting photon-beam intensity-modulated radiation therapy (IMRT)].* Strahlenther Onkol, 2012. **188**(1): p. 97-9.
9. Taha, A.A. and A. Hanbury, *Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool.* BMC Med Imaging, 2015. **15**: p. 29.
10. Taha, A.A. and A. Hanbury, *An efficient algorithm for calculating the exact Hausdorff distance.* IEEE Trans Pattern Anal Mach Intell, 2015. **37**(11): p. 2153-63.

11. Pinter, C., et al., *SlicerRT: radiation therapy research toolkit for 3D Slicer.* Med Phys, 2012. **39**(10): p. 6332-8.
12. Steenbakkers, R.J., et al., *Reduction of observer variation using matched CT-PET for lung cancer delineation: a three-dimensional analysis.* Int J Radiat Oncol Biol Phys, 2006. **64**(2): p. 435-48.
13. Ernst, M.D., *Permutation Methods: A Basis for Exact Inference.* Statistical Science, 2004. **19**(4): p. 676-685.
14. van Herk, M., et al., *The probability of correct target dosage: dose-population histograms for deriving treatment margins in radiotherapy.* Int J Radiat Oncol Biol Phys, 2000. **47**(4): p. 1121-35.
15. Riou, O., et al., *Integrating respiratory-gated PET-based target volume delineation in liver SBRT planning, a pilot study.* Radiation Oncology, 2014. **9**(1): p. 127.
16. Pech, M., et al., *Radiotherapy of liver metastases. Comparison of target volumes and dose-volume histograms employing CT- or MRI-based treatment planning.* Strahlenther Onkol, 2008. **184**(5): p. 256-61.
17. Outwater, E., et al., *Hepatic colorectal metastases: correlation of MR imaging and pathologic appearance.* Radiology, 1991. **180**(2): p. 327-32.
18. Okano, K., et al., *Fibrous pseudocapsule of metastatic liver tumors from colorectal carcinoma. Clinicopathologic study of 152 first resection cases.* Cancer, 2000. **89**(2): p. 267-75.
19. Voroney, J.P., et al., *Prospective comparison of computed tomography and magnetic resonance imaging for liver cancer delineation using deformable image registration.* Int J Radiat Oncol Biol Phys, 2006. **66**(3): p. 780-91.
20. Consortium, U. *Stereotactic Ablative Body Radiotherapy (SBRT): A Resource*. 2020 2019.
21. Brock, K.K., et al., *Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132.* Medical Physics, 2017. **44**(7): p. e43-e76.
22. Lundervold, A.S. and A. Lundervold, *An overview of deep learning in medical imaging focusing on MRI.* Zeitschrift für Medizinische Physik, 2019. **29**(2): p. 102-127.
23. Bousabarah, K., et al., *Deep convolutional neural networks for automated segmentation of brain metastases trained on clinical data.* Radiation Oncology, 2020. **15**(1): p. 87.

List of Figures:

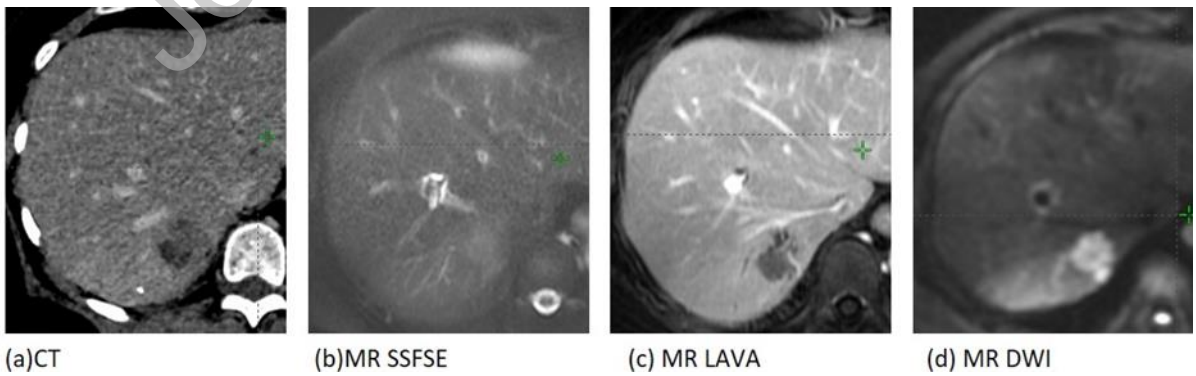

Figure 1 The appearance of the GTV for delineation on the (a) CT&contrast, (b) MR SSFSE (c) MR LAVA (d) MR DWI

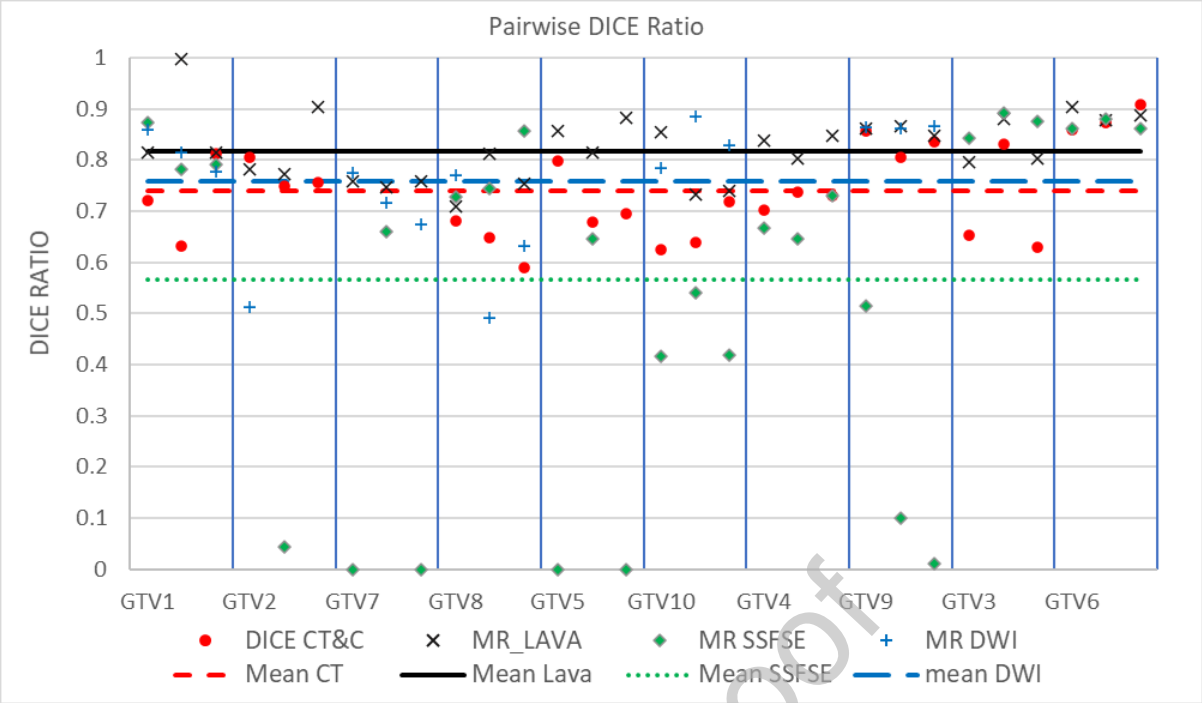

Figure 2 Pairwise Dice Ratio comparison, comparing interobserver 1 & 2, interobserver 1&3 and interobserver 2&3 for each of the 10 GTVs in order of GTV size.

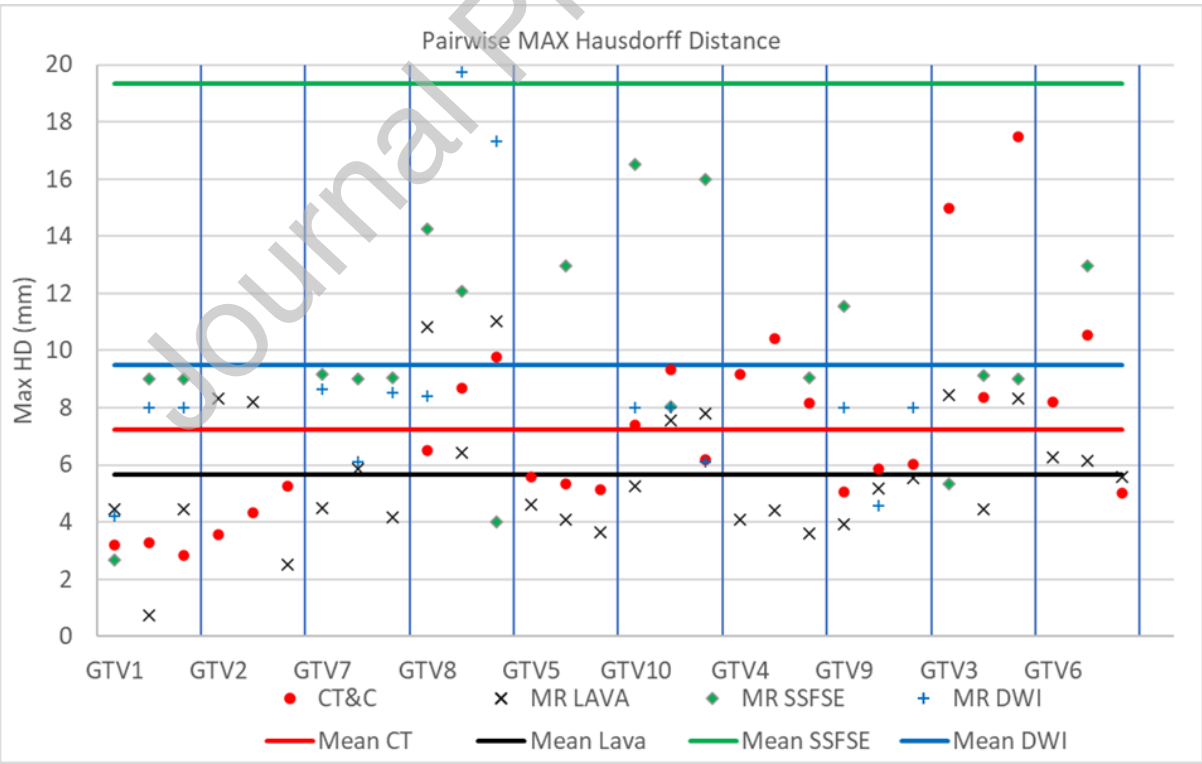

Figure 3 Pairwise HDmax distance of Interobservers 1&2, Interobservers 1&3 and Interobservers 2&3 in order of GTV size. Large HDmax of over 20mm were seen in MR SSFSE, they are not included in this graph.

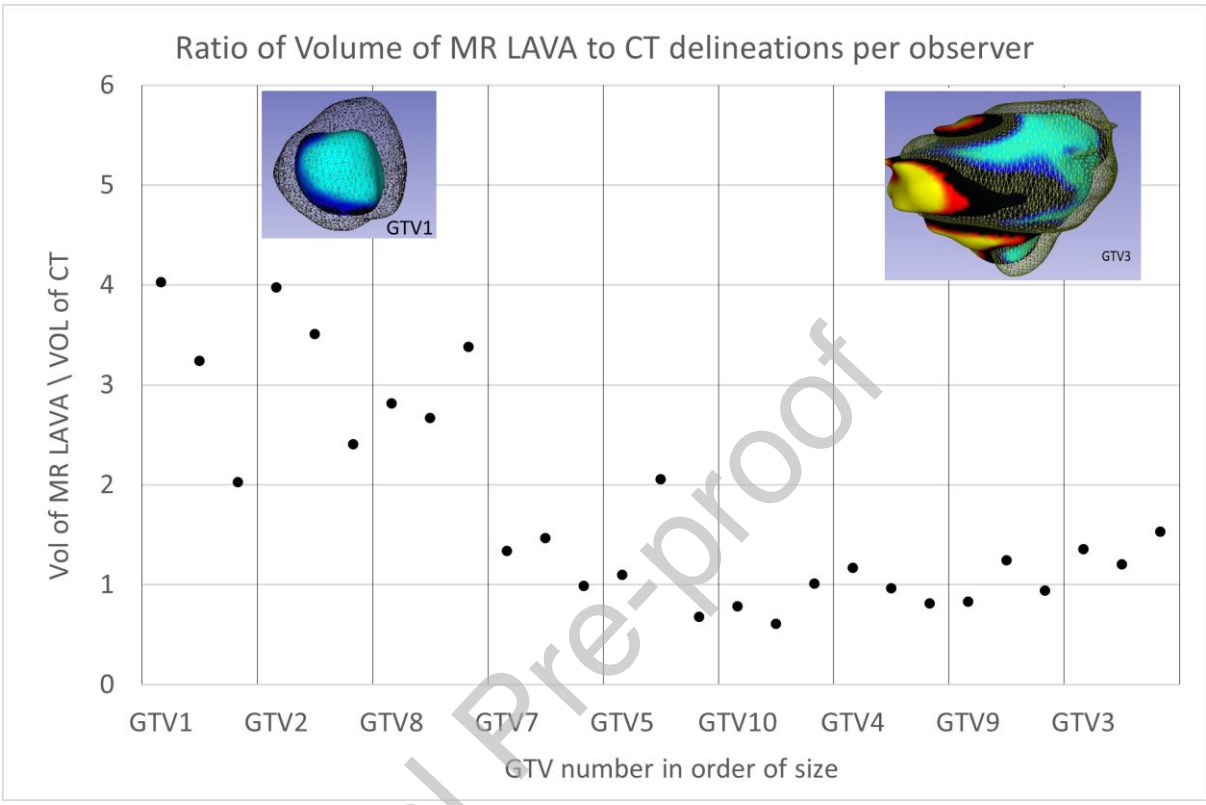

Figure 4 Ratio of the volume of the GTV drawn on MR LAVA to CT by each observer, inset GTV1 and GTV3 are 3D models, the wireframe is the MR LAVA and solid structure is the CT

Tables

Table 1 Information on the GTVs delineated, the segment of the liver, the estimated size of the tumour by the radiologist, the timing of the image post contrast injection, whether a DWI was available and if a contrast enhanced CT was possible

	Liver Segment	Size(cm)	MR LAVA Contrast timing (s)	MR DWI
GTV1	2	1.3	70	Yes
GTV2	7	2	130	Yes
GTV3	6	4	70	No

GTV4	5	3.5	70	No
GTV5	6	2	70	No
GTV6	8	4.5	130	No
GTV7	6	1.4	70	Yes
GTV8	7	2	70	Yes
GTV9	7	2	70	Yes
GTV10	7	2	70	Yes

Table 2 Conformity index, the overlap volume of all three GTVs divided by the encompassing volume of all 3 GTVs for CT&C, MR LAVA, MR SSFSE and MR DWI.

	CT&C	**MR LAVA**	**MR SSFSE**	**MR DWI**

Table 3 A comparison of CT, MR LAVA, MR SSFSE and MR DWI mean and standard deviation data for each metric

	CT&C		MR LAVA		MR SSFSE		MR DWI	
Metric	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
DICE	0.74	0.09	0.82	0.06	0.55	0.34	0.76	0.12
HDmax (mm)	7.25	3.45	5.68	2.31	19.34	15.5	9.51	5.01
Conformity index	0.47	0.12	0.58	0.12	0.29	0.28	0.46	0.16
GTV10	0.83		0.84		0.10		0.82	

Table 4 Permutation test p value results of each image set mean metric value compared to MR LAVA

	CT	SSFSE	DWI
DICE	<0.01	<0.01	0.01
HDmax	0.04	<0.01	<0.01
CI	0.08	0.02	0.09