Interfractional dose accumulation for MR-guided liver SBRT
Variation among algorithms is highly patient- and fraction-dependent

Wahlstedt, Isak; Smith, Abraham George; Andersen, Claus Erik; Behrens, Claus Preibisch; Nørring Bekke, Susanne; Boye, Kristian; van Overeem Felter, Mette; Josipovic, Mirjana; Petersen, Jens; Risumlund, Signe Lenora; Tascón-Vidarte, José David; van Timmeren, Janita Elizabeth; Vogelius, Ivan Richter

Published in:
Radiotherapy and Oncology

DOI:
10.1016/j.radonc.2022.109448

Publication date:
2023

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Interfractional dose accumulation for MR-guided liver SBRT: Variation among algorithms is highly patient- and fraction-dependent

Isak Wahlstedt, Abraham George Smith, Claus Erik Andersen, Claus Preibisch Behrens, Susanne Nørring Bekke, Kristian Boye, Mette van Overeem Felter, Mirjana Josipovic, Jens Petersen, Signe Lenora Risumlund, José David Tascón-Vidarte, Janita Elizabeth van Timmeren, Ivan Richter Vogelius

Background and purpose: Daily plan adaptations could take the dose delivered in previous fractions into account. Due to high dose delivered per fraction, low number of fractions, steep dose gradients, and large interfractional organ deformations, this might be particularly important for liver SBRT. This study investigates inter-algorithm variation of interfractional dose accumulation for MR-guided liver SBRT.

Materials and methods: We assessed 27 consecutive MR-guided liver SBRT treatments of 67.5 Gy in three fractions (n = 15) or 50 Gy in five fractions (n = 12), both prescribed to the GTV. We calculated fraction doses on daily patient anatomy, warped these doses to the simulation MRI using seven different algorithms, and accumulated the warped doses. Thus, we obtained differences in planned doses and warped or accumulated doses for each algorithm. This enabled us to calculate the inter-algorithm variations in warped doses per fraction and in accumulated doses per treatment course.

Results: The four intensity-based algorithms were more consistent with planned PTV dose than affine or contour-based algorithms. The mean (range) variation of the dose difference for PTV D95% due to dose warping by these intensity-based algorithms was 10.4 percentage points (0.3 to 43.7) between fractions and 8.6 (0.3 to 24.9) between accumulated treatment doses. As seen by these ranges, the variation was very dependent on the patient and the fraction being analyzed. Nevertheless, no correlations between patient or plan characteristics on the one hand and inter-algorithm dose warping variation on the other hand was found.

Conclusion: Inter-algorithm dose accumulation variation is highly patient- and fraction-dependent for MR-guided liver SBRT. We advise against trusting a single algorithm for dose accumulation in liver SBRT.

Accurate dose accumulations would allow daily dose planning in online-adaptive radiotherapy to be based on accumulated delivered dose, rather than planned dose on a pre-treatment simulation scan. Such an approach is hypothesized to improve daily plan adaptations. Several commercial vendors and open source algorithms offer dose warping and accumulation based on deformable image registration (DIR). The use of these algorithms has gained international interest [1,2]. Bohoudi et al. [3] investigated the correlation between accumulated dose and outcome for 101 prostate cancer patients treated in five fractions. They found that an increase in International Prostate Symptoms Score correlated better with bladder doses accumulated after three and five fractions than with pre-treatment planned dose. This suggests that accounting for accumulated dose from previous fractions could improve radiotherapy outcome. High dose delivered per fraction, steep dose gradients, and large interfractional organ deformations imply that accounting for dose delivered in previous fractions might be particularly important for online-adaptive stereotactic body radiotherapy (SBRT) in the liver. However, as shown by Wang et al. [4], DIR algorithm settings can considerably affect the accumulated dose for patients with liver metastases treated with SBRT.

Recently, online-adaptive magnetic resonance-guided radiotherapy (MRgRT) with automated beam-gating based on intrafractional organ motion has been implemented [5]. MR-guided SBRT in
the liver has achieved local control rates consistent with literature with minimal toxicity for both metastatic and primary lesions [6]. MRgRT is offering a non-invasive, marker-free option for high precision SBRT of liver malignancies and allows for online plan adaptations to account for interfractional anatomical variations. MRgRT provides daily volumetric magnetic resonance images (MRIs) and fraction-wise dose calculations. Hence interfractional dose accumulation is readily available. The superior soft tissue contrast of MRIs compared to computed tomography (CT) images makes DIR-based dose accumulation using MRIs promising for patients with liver malignancies; both with regards to precision and accuracy. However, the lack of clinical landmarks makes assessment of DIRs in the liver challenging. A “ground truth” can be achieved by annotating anatomical landmarks, e.g. Brock et al. [7], or by comparing DIRs to known displacement vector fields (DVFs), e.g. Bosma et al. [8]. In the absence of a ground truth deformation it is not possible to quantify absolute errors in registration on real world patient cases. Thus, the purpose of this study was to assess the impact of daily anatomical variation on the delivered dose per fraction. Subsequently, we warped the recalculated fraction doses onto the MRsim (Step 3 in Fig. 1) using seven image registration algorithms. Finally, the warped fraction doses (from Step 3, in Fig. 1) were accumulated on the MRsim (Step 4 in Fig. 1). It is important to note that: a) the target structures were rigidly transferred to the daily MRIs according to the online match performed during treatment, and b) the position of the isocenter of the treatment plan is fixed in relation to the transferred target structure. Thus, the relationship between the beam- and target geometry stays the same but may move together in the daily image according to the online match. This means that if a setup-error pushes the target structure away from its actual position this will not be correctly reflected in the calculated daily dose. Therefore, setup errors are not included when we evaluate the impact of daily anatomical variation on the delivered dose per fraction (black arrow in Fig. 1). Since the target mainly is in a dosimetricaly homogenous environment, as the liver, the daily target coverage may still be calculated as good even though the target structure is placed incorrectly due to a setup error. However, when this dose is mapped back to the MRsim with DIR, it is mapped to the position in the MRsim corresponding to the target structure position in the fraction scan. Consequently, warped dose contains setup error with respect to the target on MRsim while daily recalculated dose does not.

Assessment of Inter-algorithm dose warping variation

First, we calculated the planned static dose on the simulation MRI (MRsim) (Step 1 in Fig. 1). Subsequently, we recalculated these base plans on the daily patient anatomy (Step 2 in Fig. 1) as described in the supplement. Thus, we were able to estimate the impact of daily anatomical variation on the delivered dose per fraction. Subsequently, we warped the recalculated fraction doses onto the MRsim (Step 3 in Fig. 1) using seven image registration algorithms. Finally, the warped fraction doses (from Step 3, in Fig. 1) were accumulated on the MRsim (Step 4 in Fig. 1). It is important to note that: a) the target structures were rigidly transferred to the daily MRIs according to the online match performed during treatment, and b) the position of the isocenter of the treatment plan is fixed in relation to the transferred target structure. Thus, the relationship between the beam- and target geometry stays the same but may move together in the daily image according to the online match. This means that if a setup-error pushes the target structure away from its actual position this will not be correctly reflected in the calculated daily dose. Therefore, setup errors are not included when we evaluate the impact of daily anatomical variation on the delivered dose per fraction (black arrow in Fig. 1). Since the target mainly is in a dosimetricaly homogenous environment, as the liver, the daily target coverage may still be calculated as good even though the target structure is placed incorrectly due to a setup error. However, when this dose is mapped back to the MRsim with DIR, it is mapped to the position in the MRsim corresponding to the target structure position in the fraction scan. Consequently, warped dose contains setup error with respect to the target on MRsim while daily recalculated dose does not.

Materials and methods

Patients and plans

We assessed MRgRT treatment plans (n = 35) with dose prescription of 67.5 Gy in three fractions and 50 Gy in 5 fractions to the GTV for liver SBRT at our institutions (RH and HGH) from the day of commissioning of the first (of two) MR-linac in January 2019 until April 1, 2021. A flowchart of the inclusion of 27 of those 35 treatment plans is presented in Fig. S1 in the supplement. The PTV coverage aim was D99% ≥ 67 %. All patients were treated in inspiration breathhold (IBH) and MR-guided gating was used in all fractions. Two patients were treated twice. For one of these patients treatments occurred nine months apart and for the other patient treatments were performed on different days. Thus, all four treatments represented individual inter-fractional changes, dose distributions, and locations in the liver. Hence 27 MRgRT plans and 105 fractions from 25 patients were included. As per Fig. S1 we would have analyzed (15 × 3) + (12 × 5) = 105 fractions. However, due to treatment interruptions some fractions were split, to yield a total of 139 fractions and partial fraction MRIs.

All treatment plans were delivered with step-and-shoot intensity-modulated radiotherapy. Patients were treated with MRIdian MR-linacs (ViewRay Inc., Mountain View, CA) [5]. All treatment fractions were delivered non-adaptively and target structures were transferred rigidly from the pre-treatment MRI to subsequent daily MRIs.

The GTV coverage constraints used in treatment planning varied as follows: D99% ≥ 95 % and mean dose ≥ 100 % (24 of 27 plans), D99% ≥ 95 % (2 of 27, no mean dose constraint), and mean dose ≥ 100 % in the last plan (no D99% constraint). The remaining dose constraints are given in Table S1 in the supplement. One treatment plan included three GTVs, three plans included two GTVs, and 23 plans included one GTV. One treatment plan including two GTVs had an isotropic GTV to planning target volume (PTV) margin of 8 mm. The remaining 26 plans had a GTV-to-PTV margin of 5 mm in the left–right and anterior-posterior directions and 10 mm (n = 15), 7 mm (n = 7), 5 mm (n = 2), or a mix of 10 mm and 7 mm (n = 2) in the craniocaudal direction. The GTV-to-PTV margin used clinically was chosen based on the number of targets in the treatment plan, inter-breathhold reproducibility of the tumor position, and distance to organs at risk. The base plans (planned dose calculated on simulation MRI) had (mean ± sd) GTV D99% of 96.2 % ± 2.0 %, GTV D2% of 107.7 % ± 5.3 %, and PTV D99% of 67.6 % ± 6.4 % of the prescription dose. The conformity indices of the base plans, defined as the volume covered by 95 % of the prescription dose to the structure divided by the structure volume, were median (interquartile range) 1.6 (1.4, 2.1) for the GTV and 1.6 (1.4, 1.9) for the PTV. See supplement for further details.

Imaging

Pre-treatment CT scans were acquired in IBH and, when relevant, also as four-dimensional binned CTs for each patient. Three CT scanner models were used: a Siemens SOMATOM Definition AS (n = 24), a Siemens Somatom go.Open Pro (n = 2) (Siemens Healthcare, Forchheim, Germany), and a Philips Brilliance CT Big Bore, version 3.5.17001 (Philips Medical Systems, Cleveland, OH) (n = 1). All CT scans had a slice spacing of 2 mm and in-plane resolutions of 1.0 × 1.0 mm² (n = 6), 1.3 × 1.3 mm² (n = 3), 1.4 × 1.4 mm² (n = 17), or 1.5 × 1.5 mm² (n = 1). The field of view (FOV) of the CT scans ranged from 51 cm to 77 cm in the left–right and anterior-posterior directions and 26 cm to 36 cm in the cranial-caudal direction. We acquired volumetric MRIs in IBH before treatment planning (MRsim) and before each treatment fraction using a True Fast Imaging with Steady State Free Precession (TRUF) sequence, described by Green et al. [9]. The MRI acquisition times were either 17 s or 25 s, depending on the patient’s breath hold compliance, and maintained during all fractions. A 3–5 mm gating window margin was used in the sagittal plane during each treatment MRI acquisition, which is expected to minimize breath hold variation. The MRI resolution was 1.5 × 1.5 × 3.0 mm³ (n = 23) or 1.6 × 1.6 × 3.0 mm³ (n = 4). The FOVs of the MRIs were 50 × 45 × 43 cm³ (n = 21), 45 × 45 × 24 cm³ (n = 4), 54 × 47 × 43 cm³ (n = 1), and 35 × 35 × 43 cm³ (n = 1). Twenty-four plans were delivered at RH and three at HGH. For information on quality assurance of the MRIs, see supplement.
Image registration algorithms

Six DIR algorithms and one affine algorithm were employed for MRI-based dose warping (Step 3 in Fig. 1). Four of the DIR algorithms are commercially available and two are open source. For the four commercially available algorithms, the pre-DIR rigid registration (RIR) was focused on the GTV and the volumes of interest for the DIRs were encompassing the whole body in the transaxial view and the whole liver in the craniocaudal direction. Three of the commercially available algorithms are from MIM (MIM Software Inc, Cleveland, OH). The MIM algorithms used the same RIRs as the starting transformation for theDIRs. The MIM algorithms were intensity-based (MIM-IB), contour-based (MIM-CB), and hybrid (MIM-Hybrid) [10]. In short, MIM-IB uses Multi-Modality Deformable Registration with a feature similarity scoring metric. The MIM-CB algorithm iteratively minimizes the signed distances between the liver contours on the fixed and the moving image. The MIM-Hybrid algorithm utilizes both intensity-based and contour-based DIR. The fourth commercially available algorithm was the intensity-based Deformable Multipass algorithm from Velocity (Varian Medical Systems, Inc, Palo Alto, CA) (from now on called Velocity). The open source algorithms were scripted in GNU bash 5.0.17 and Python 3.8.10 to compute dose accumulation automatically. Before DIR, both algorithms performed an affine registration. The open source DIR algorithms employed were the Symmetric Image Normalization method [11] in Advanced Normalization Tools (ANTS) using the default input parameters and a b-spline DIR from the Elastix [12,13] toolbox (from now on called Elastix). Two patients were used for visually optimizing the Elastix deformation settings, one with small anatomical differences between the moving and the fixed image and another with large anatomical differences. Both previously published parameter files1 and in-house-developed parameter files were tested, and the final parameter files can be found in the supplement. The code employed for dose accumulation with ANTs and Elastix is available on GitHub [14]. All the algorithms employed used trilinear interpolation as described by Chetty and Rosu-Bubulac [15] for dose warping after DIR. Further information about the DIR algorithms can be found in Table S2 in the supplement. Apart from DIR algorithms, we employed the affine algorithm available in ANTs (Affine) for image-based dose accumulation. This affine transform was also used as a starting point for the ANTs DIR.

The MIM-IB algorithm was employed twice. This yielded two sets of RIRs, DIRs, and dose accumulation results for MIM-IB. One set had common RIRs with MIM-CB and MIM-Hybrid. This set was used in the inter-algorithm variation assessment. The two results for accumulated treatment dose obtained with MIM-IB were compared for PTV D95% in order to assess the impact of the pre-DIR RIR on the accumulated dose.

1 From the model zoo, (https://elastix.lumc.nl/modelzoo/): Par0003, Par0022, Par0032, Par0040.
Liver contour delineation

Accurate liver delineations were needed for all 139 fraction MRIs in order to perform contour-based DIR as implemented in the MIM-CB and MIM-Hybrid algorithms. However, the daily MRIs only had an auto-segmented liver contour of suboptimal correspondence with anatomical boundaries. Thus, one observer (AGS) delineated the liver on all 27 MR sim and on all 139 fraction MRIs using RootPainter3D [16], an open-source corrective-annotation software application. The output from RootPainter3D is a mask of the liver structure. These masks were converted to DICOM format using an inhouse-developed script [17] before import to MIM. All delineated livers were checked, and if necessary corrected, in MIM by another observer (IW) before application in DIR. The focus of the corrections in MIM was that the liver structure on the MR sim and on the fraction MRIs should be as similar as possible and correspond to the same patient anatomy.

Dosimetric analysis

All dose volume histograms (DVHs) were generated in the open source image computation software Plastimatch [18]. Target structures rigidly transferred to the daily MRIs were used for fraction-wise DVH computations. The positions of the target structures on the daily MRIs were checked in the clinical workflow as part of the radiotherapy workflow. We analyzed the dosimetric effect of anatomical differences and dose warping variation per fraction separately, as suggested by Nenoff et al. [19]. We compared the planned dose distribution with the doses calculated on daily patient anatomy (leftmost arrow in Fig. 1) for GTV D_{98\%} and PTV D_{95\%}. The same metrics were chosen for evaluation of the variation introduced by dose warping per fraction (Comparison A, blue arrow in Fig. 1) and over whole treatment courses (Comparison B, green arrow in Fig. 1). In Comparison B, we also investigated the dose differences between planned and accumulated dose for

![Box plots of difference between accumulated and planned dose for four DVH metrics: GTV D_{98\%} (A), PTV D_{95\%} (B), 3 cm Ring D_{50\%} (C), and Healthy liver V_{<15Gy} (D). All plots are ordered from lowest to highest median PTV D_{95\%} difference to planned dose. pps = percentage points.](image-url)
the PTV DVHs. Furthermore, in Comparison B, we evaluated differences in $D_{95\%}$ to the 3 cm Ring (a 3 cm expansion of the PTV) and $V_{\text{75Gy}}$ to the healthy liver (liver minus GTV). We also investigated correlations between plan characteristics (conformity indexes for GTV and PTV) or patient characteristics (GTV size, distance from GTV structure to liver structure, and distance from GTV structure to lung structure) on the one hand and inter-algorithm variations in dose warping on the other hand. We performed Wilcoxon signed-rank tests in the evaluation of these dose differences.

Quality assurance of image registrations for a relative comparison

We assessed the quality of the registrations underlying the dose accumulations by visually comparing the deformed moving image to the fixed image. This was performed as a crude quality assurance (QA) prior to the quantitative QA in order to ensure that the algorithms did not fail completely for any deformation. Such failures could have resulted in algorithm tuning where possible and discardment of the algorithm if tuning was not an option. Furthermore, we conducted a quantitative analysis of the registrations. All metrics used for QA of the registrations are defined in the supplement. This analysis was conducted to get a better understanding of the algorithms included and to obtain a relative comparison of the algorithms in different aspects. Additionally, when possible, we compared the results to recommended absolute tolerances by American Association of Physicists in Medicine Task Group 132 (TG 132) [20]. First, we computed the segment metrics for the liver contour by converting the structures from DICOM to numpy arrays using an inhouse script [17]. Then we used MedPy [21] to compute the dice similarity coefficient (DSC), the 95th percentile Hausdorff distance (HD95), precision, and recall between the liver contour delineated on the MR sim (reference contour) and the liver contour delineated on the daily MRI and subsequently transferred to the MR sim (experimental contour). We also assessed the image similarity metrics normalized mutual information (NMI) and structural similarity index (SSI). First, we exported the DVFs from the algorithms in order to reproduce all deformations in Plastimatch. This enabled us to compute NMI and SSI both for the full images and within 3 cm Ring using in-house scripts based on implementations of NMI and SSI from scikit-image [22]. All segment metrics and image similarity metrics were computed for all 139 registrations for all algorithms. We were able to reproduce the deformations in Plastimatch for all algorithms except Velocity by using DVFs from the evaluated algorithms. This yielded a consistent estimation of the Jacobian of the transformation (JD), which could otherwise be sensitive to the implementation of the calculation. DVFs exported from Velocity to Plastimatch did not yield self-consistent results and we therefore excluded Velocity from this comparison. JDs for the remaining algorithms were computed with consistent results and we therefore excluded Velocity from this comparison. JDs for the remaining algorithms were computed with consistent results and we therefore excluded Velocity from this comparison. JDs for the remaining algorithms were computed with consistent results and we therefore excluded Velocity from this comparison. JDs for the remaining algorithms were computed with consistent results and we therefore excluded Velocity from this comparison. JDs for the remaining algorithms were computed with consistent results and we therefore excluded Velocity from this comparison.

Results

The mean differences between planned and accumulated doses for full treatment courses (Comparison B) for the target metrics (GTV $D_{98\%}$ and PTV $D_{95\%}$) were smallest for ANTs and largest for Affine. Generally, Affine and MIM-CB showed the largest differences to planned dose for the target metrics (Fig. 2 A/B and supplement Table S3). $D_{95\%}$ to 3 cm Ring and $V_{\text{75Gy}}$ to the healthy liver were less sensitive to DIR variation than the target DVH metrics (Fig. 2 C/D and supplement Table S3). Thus, the remaining analysis of inter-algorithm dose warping variation was performed for the target metrics only. A negligible impact of the pre-DIR manual RIR on the PTV $D_{95\%}$ was found for the MIM-IB algorithm (Supplementary Table S3). Differences in $D_{50\%}$ to the 3 cm Ring were reported.

![Fig. 3. Differences between planned and accumulated treatment doses of the PTV DVHs. A: Population-based median and interquartile ranges of cumulative DVHs. B: Population-based difference to planned dose cumulative DVHs, where solid lines are the median DVH per method (only intensity-based algorithms shown) and shaded areas are interquartile ranges (only Affine and ANTs shown). C: Individual plan data show area under the difference to planned dose PTV DVHs. The boxplots in C are ordered from top to bottom from smallest to largest median area per method. Planned dose (in A) is plotted with black, the rest of the colors for each method are given in C.](image-url)

| Table 1 |
| --- | --- | --- |
| Mean [Range] difference in percentage points | GTV $D_{95\%}$ | PTV $D_{95\%}$ |
| Comparison A: Variation per Fraction | 12.0 [0.8 to 61.8] | 10.4 [0.3 to 43.7] |
| Comparison B: Accumulated Variation per Treatment Course | 9.1 [1.0 to 24.9] | 8.6 [0.3 to 6.8] |
| Differences due to Anatomical Variation | $-0.9 [-10.9]$ | $-1.1 [-6.8]$ |
| to 3.0 | to 2.3 |
ment (Fig. S2). The standard deviation of the differences in accumulated dose between the first and the second attempt was 0.63 percentage points (pps).

The difference to planned dose cumulative PTV DVHs (Fig. 3 B) show inter-algorithm variation more clearly than median DVHs (Fig. 3 A). Affine, MIM-CB, and MIM-Hybrid obtained accumulated dose less consistent with planned dose than the other algorithms did (large areas in Fig. 3 C). ANTs obtained accumulated doses closest to the planned dose (Fig. 3 C). Considering the large inter-algorithm variation in accumulated treatment dose in Fig. 3, we decided to present the analysis of fraction-wise inter-algorithm variation for a subset of similarly performing algorithms. The intensity-based algorithms (ANTS, MIM-IB, Velocity, and Elastix) are appealing since they do not demand any contouring. Wilcoxon signed rank tests between ANTs and the other algorithms for the PTV area plot (Fig. 3 C) showed that the other intensity-based algorithms obtained values closest to ANTs (see p-values in Fig. 3 C). Thus, we decided to present the rest of the analysis of inter-algorithm variation for the intensity-based algorithms algorithms only.

Table 1 shows that the mean variation in dose differences due to dose warping per fraction (Comparison A) and per treatment course (Comparison B) were both larger in magnitude than dose differences due to anatomical variation (leftmost arrow in Fig. 1). Fig. 4 and Fig. S3 show the coverage difference to planned dose per fraction due to dose warping for PTV D95% and GTV D98%, respectively. The variation of the dose difference due to dose warping by intensity-based algorithms varied considerably both between treatment courses (length of black lines in Fig. 4 and Fig. S3) and between fractions (length of blue lines in Fig. 4 and Fig. S3). Nevertheless, no correlations between patient or plan characteristics on the one hand and inter-algorithm dose warping variation on the other hand were found in this study (data not shown). Treatment 23 (bottom row, third column from right in Fig. 4), fraction 3 had the largest inter-algorithm variation among the intensity-based algorithms (43.7 pps). If we also include the MIM-Hybrid algorithm in this comparison, the inter-algorithm variation is even larger (see DVHs in Fig. 5). MIM-CB deforms the liver boundary on the third fraction scan to match the liver boundary on the MR sim while the other algorithms are more preserving of the third fraction liver boundary (Fig. S5 in supplement). For this fraction the caudal part of the liver boundary looks very different on the third fraction scan than on the MR sim (Fig. 5 A). This pushes the dose in different directions relative to the GTV on the MR sim depending on how well the caudal boundary of the liver is aligned on the two scans (Fig. 5).

The visual assessment of the registrations did not result in any additional tuning of the tunable algorithms (ANTS and Elastix) nor any discardment of one of the other algorithms. MIM-CB and MIM-Hybrid performed the best for the segment metrics, while MIM-Hybrid, MIM-IB, and ANTs were top performers for the image similarity metrics (Table S4 and Fig. S4 supplement). TG 132 [20] recommends that DSC values should be above 0.8–0.9 and that JDs should be positive. All registrations with DIR algorithms, except for one registration in Velocity, obtained DSC scores above 0.8 whereas all algorithms obtained some transformations with

![Fig. 4. PTV D95% coverage differences compared to planned dose due to warping of the daily doses to the MR sim for intensity-based algorithms. One panel represents one treatment course. Median coverage differences per fraction (blue squares) and variation per fraction (lengths of blue lines). The treatment courses are arranged from upper left to lower right according to the variation in accumulated dose among the algorithms (lengths of black lines). Difference between planned dose and calculated dose on daily anatomy (red dots) is given as comparison of magnitude. F1 to F5 on the horizontal axis are fraction numbers and pps is percentage points.](image-url)
negative JDs, mainly in areas with missing or appearing tissue. Affine performed poorly for both types of metrics. ANTs and MIM-IB had highest median JD 1% and lowest median JD 99%, indicating that these were the most well-regularized algorithms (Table S4 and Fig. S4 supplement). The transformation between MRsim and the third fraction for treatment 27 in Elastix had the lowest JD1% (−5.1). Heat maps of the JD values for Elastix, ANTs, and MIM-IB in slices with lowest JDs in 3 cm Ring reveal that deformations with ANTs and MIM-IB resulted in less extreme JD values than the deformation with Elastix for this fraction (Fig. S6), most visible where air or arms are present on one image and not on the other.

Discussion

Assessing inter-algorithm variation in dose warping based on differences in GTV coverage is complicated by the different PTV margins applied in this study. Larger PTV margins lead to less steep dose falloff outside of the GTV, making the GTV coverage more robust against inter-algorithm variation. Thus, we focus this discussion on inter-algorithm variations obtained in PTV coverage. Affine, MIM-CB, and MIM-Hybrid produced warped doses further from the planned dose than the other algorithms evaluated. The inter-algorithm variation among the intensity-based algorithms was substantial between treatment courses; with mean (range) variations of 8.6 (0.3 to 24.9) pps for PTV D95%. This means that for some treatment courses the inter-algorithm variation was as low as 0.3 pps for the sensitive PTV D95% metric, whereas dose accumulation for other treatment courses yielded variation for the same metric up to 25 pps. Thus, the inter-algorithm variation in accumulated dose was highly patient-dependent (length of black lines in Fig 0.5). Nevertheless, no correlations between patient or plan characteristics on the one hand and inter-algorithm dose warping variation on the other hand was found in this study. Furthermore, in the fraction-wise assessment of the warped doses, we found that inter-algorithm variation was not only patient-dependent but also highly fraction-dependent.

Wang et al. [4] investigated DIR uncertainty for 16 patients with 17 liver metastases treated with three fractions of 12.5 Gy to 16.75 Gy (prescribed to the 67 % isodose surface surrounding the PTV) of CT-based SBRT with steep dose gradients. Each daily CT was registered 132 times to the pre-treatment CT using a range of parameter settings for the same DIR algorithm. They evaluated how DIR uncertainty translated to dose uncertainty for three hollow organs at risk. Among the registrations selected as realistic, 12 of 144 (16/130) evaluated combinations of fractions and doses showed a 99th percentile dose difference over 3 Gy. Despite the superior soft tissue contrast of MRIs compared to CT images, our results agree with Wang et al. [4] that DIR-based dose warping can be highly dependent on the algorithm applied; in their case evaluated based on several settings for the same algorithm and in our case evaluated for several commercial or open source algorithms. Bohoudi et al. [24] employed one contour-based and one intensity-based DIR algorithm for accumulating dose to bladder and rectum in six MR-guided prostate SBRT treatments. They found that doses to rectum and bladder surfaces on average differed from film measurements in an anthropomorphic phantom by 0.6 % and 0.3 %, respectively, for the contour-based algorithm and 7.2 % and 2.5 %, respectively, for the intensity-based algorithm. These results also seem to agree with ours in showing a large inter-algorithm variation in MRI-based dose warping for SBRT treatments. As shown by Saleh-Sayah et al. [25], DVF accuracy demands are the highest in steep dose gradient regions. In this study we observed large differences between planned and accumulated doses for GTV D95% and PTV D95%, as well as high inter-algorithm variation for the same metrics, which are probably partly due to the steep dose gradients outside of the GTV. Also lack of contrast in the liver, despite the high soft tissue contrast of MRI, probably contributes to DIR uncertainty. The large differences between recalculated dose on daily anatomy (red dots in Fig. 4 and
Fig. S3) and median of warped doses among intensity-based algorithms (blue squares in Fig. 4 and Fig. S3) are amplified by the fact that warped doses also contain setup-uncertainties whereas daily calculated doses do not.

One limitation of this study is the number of treatment courses included. This might be why we fell short of finding any patient or plan characteristics correlating with inter-algorithm dose warping variation. However, despite the sample size limitation and the lack of possibility to test all possible associations, the existence of a metric that can foresee dose accumulation variability with sufficient accuracy for clinical use without a secondary and confirmative algorithm seems unlikely. Thus, our conclusion that inter-algorithm dose warping variation is both patient- and fraction-dependent is valid despite the limited sample size. This is, for example, supported by the diversity of inter-algorithm variation even between fractions in the same treatment course (Fig. 4).

Another limitation is that we failed to include Velocity in the quantitative analysis of the JDs. Despite help from colleagues at other institutions, we were not able to reproduce deformations from Velocity externally. Thus, the displacement vector fields from Velocity were evaluated only qualitatively.

Due to the lack of ground truth deformations, this study was not designed to find the most suitable algorithm for interfractional dose accumulation for daily planning in online-adaptive MR-guided liver SBRT. Rather, a small inter-algorithm dose accumulation variation would have yielded an increased confidence in the algorithms employed. However, considering the high inter-algorithm dose accumulation variation found in this study, interfractional dose accumulation for daily planning in online-adaptive MR-guided liver SBRT cannot be considered safe without assessment of uncertainty of the involved algorithms on a per-patient basis. However, for reporting delivered dose after a course of MR-guided liver SBRT, accumulated dose based on a range of DIR algorithms might still be appropriate. A set of algorithms, properly quality-assured for MRI-based dose accumulation in the liver, preferably algorithms with different similarity metrics and/or regularization, could be employed on a set of patients. If we were to conduct such a study at our departments, we would, of the algorithms evaluated in this study, include ANTs and MIM-IB for the task. These algorithms are most regularized (Fig. S4 I and J) and still among the top performers for the image similarity metrics (Fig. S4 E-H). Such an approach could be hypothesized to correlate better with outcome than planned static dose and thus increase understanding of dose–response relationships. Especially for online-adapted treatment courses where the daily delivered dose to the target structure is varied from day to day due to nearby organs at risk. In such cases the planned static dose calculated at the simulation scan will most likely differ from the dose that the patient actually received, at least for the target structures.

Conclusion

Inter-algorithm dose accumulation variation is highly patient- and fraction dependent for MR-guided liver SBRT. In the absence of a ground truth, we recommend multiple algorithms employed to provide an estimate of the precision and caution in applying dose accumulation for online adaptation.

Conflict of interest statement

Isak Wahlstedt reports funding from ViewRay and the Danish Comprehensive Cancer Center as well as institutional research and teaching contracts with Varian Medical Systems, Brainlab, and Viewray. Abraham George Smith, Claus E. Andersen, Susanne Norring Bekke, Kristian Boyne, Mette van Overeem Felter, Mirjana Josipovic, Jens Petersen, Signe Lenora Risumlund, and Janita E. van Timmeren report no conflicts of interest.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Isak Wahlstedt reports funding from ViewRay and the Danish Comprehensive Cancer Center as well as institutional research and teaching contracts with Varian Medical Systems, Brainlab, and Viewray. Thomas Carlslund at Rigshospitalet is acknowledged for solving several IT- and software-related issues. Rasmus Hvass Hansen at Rigshospitalet is acknowledged for proofreading and for acquiring results for quality assurance of MRIs. Angela Trask and Jorn Verweij at ViewRay are acknowledged for providing information about how to perform calculation of delivered doses with new Monte Carlo dose calculation settings in the MRIdian treatment planning system. Maxim de Smidt, Stefano Sobkowiak, and Michael Duchateau at MIM are acknowledged for their assistance in designing workflows for dose accumulation and for providing information about the MIM algorithms used in this study. Stefan Zepter at Varian is acknowledged for support in building the dose accumulation workflow in Velocity. Jamie McClelland and Bjoern Eiben from University College London are acknowledged for helping us understand the DICOM registrations exported from Velocity.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2022.109448.

References


