Contents lists available at ScienceDirect

Zeitschrift für Medizinische Physik

journal homepage: www.elsevier.com/locate/zemedi

Original Paper

Performance evaluation of deformable image registration algorithms: target registration error and its correlation to Dice similarity coefficient

Yun Ming Wong [a,b,1], Wen Siang Lew [a], James Cheow Lei Lee [a,b], Hong Qi Tan [a,b,c,*]

[a] Division of Physics and Applied Physics, Nanyang Technological University, Singapore
[b] Division of Radiation Oncology, National Cancer Centre Singapore, Singapore
[c] Oncology Academic Clinical Programme, Duke-NUS Medical School, Singapore

ARTICLE INFO

Keywords:
Deformable image registration
Dose accumulation
Quality assurance
Dice similarity coefficient
Target registration error

ABSTRACT

Objective: The wide usability of deformable image registration (DIR) deems the process of quality assurance important for a reliable clinical translation. Our work mainly aimed to compare the performances of four DIR software, in terms of voxel mapping accuracy quantified through target registration error (TRE), and its organ-wise correlation with Dice similarity coefficient (DSC), a widely used segmentation metric.
Methods: CT scans were taken for one static scenario and four deformation scenarios simulated using an in-house deformable anthropomorphic pelvis phantom. Their CT numbers were overridden based on actual patient scan, and these overridden scans were used as input images in this study. Four DIR software were tested: RayStation v10B, Velocity v4.1, Slicer, and Plastimatch. Multiple DIRs were performed for each software, using different algorithm options or parameters. The TRE was quantified by calculating the difference between the true and mapped marker positions. Subsequently, Pearson correlation tests were done to examine the correlation between DSC and mean TRE, separately for bladder, prostate, rectum and all organs combined. Similar analyses were conducted for prostate alone, to gain more insights regarding a homogeneous medium. Additionally, DSC was used to predict whether the mean TRE exceeded 3 mm. The classification performance was assessed using accuracy, precision, recall, F1-score, specificity and area under the Receiver Operating Characteristic curve (AUC).
Results: Among the four software tested, RayStation achieved the lowest mean TRE for all deformation scenarios, with values between 1.48 mm and 3.06 mm. Pearson correlation tests revealed an exceptionally strong negative correlation between DSC and mean TRE for SlicerElastix, where the correlation coefficients ranged from −0.901 to −0.987. In line with the strongest correlation found, SlicerElastix achieved the highest classification performance scores overall. For all three organs, the scores at their corresponding best DSC threshold were mostly higher than 0.80, and the AUCs were close to 1.
Conclusion: In short, this work quantified and compared four DIR software based on the voxel mapping accuracy as well as its correlation with DSC, in the major organs in prostate radiotherapy.

Introduction

The importance of deformable image registration (DIR) in the radiotherapy (RT) community is irrefutable, especially due to its high applicability in adaptive radiotherapy (ART). Over the years, DIR algorithms have seen a steady evolution, giving rise to a wide variety available for use in RT and in medical physics field in general. A typical DIR algorithm consists of a transformation model, a similarity metric and an optimiser. Based on the transformation models, two main classes have been defined, namely, parametric and non-parametric [1]. A parametric algorithm uses a limited number of control points for the deformation, whereas a non-parametric algorithm involves every single voxel.

Free-form deformation (FFD) is a parametric method that is often associated with a B-spline transformation model in the context of medical image analysis [2]. In this model, a mesh of control points is overlaid onto the moving image, and each control point is deformed based on the B-spline functions. As the B-spline functions limit the influence of deforming a control point to only its local neighbourhood, this model is able to provide a localized deformation [3]. In such a model,

[*] Corresponding author at: Division of Radiation Oncology, National Cancer Centre Singapore, 30 Hospital Blvd, Singapore 168583.
E-mail address: tan.hong.qi@nccs.com.sg (H.Q. Tan).
[1] First authors

https://doi.org/10.1016/j.zemedi.2025.09.001
Received 25 May 2025; Accepted 3 September 2025

the control points are important parameters, as the mesh resolution (i.e. the control point spacing) determines the extent and degrees of freedom for the deformation [4]. Depending on the situation, regularization could also be done on the FFD method by imposing task-specific constraints, including topology preservation, volume preservation and rigidity constraints [2].

Demons algorithm, which is commonly used due to its remarkable speed [5], falls into the non-parametric category. The idea of this algorithm originated from the concept of Maxwell's demons, which was introduced to address a paradox in thermodynamics [6]. From the perspective of image registration, an object boundary in a static image is analogous to a semi-permeable membrane, where "demons" are scattered on it. These demons act as effectors to "filter" the voxels of another image, by classifying them as either inside or outside the object boundary. The image to be deformed is considered a deformable grid, whose grid vertices correspond to the image voxels. This model is coined the diffusing model, as the deformable grid is seen to be diffusing through the object boundary in the static image, by the force of demons. For regularization purpose, a Gaussian filter is applied on the resulting deformation vector field (DVF) at each iteration. This filter serves to smoothen the DVF and preserve the geometric continuity of the deformed image [7].

Apart from the two algorithms described above, there are many more out there being constantly introduced and adapted to meet the demands of the community from various aspects (e.g. deformation complexity needed and computational efficiency). Different algorithms perform well in different aspects, and it is unlikely to find a one-size-fits-all algorithm that is suitable for use in all clinical situations.

For the application of DIR in dose accumulation, i.e., deformable dose accumulation (DDA), assessment of voxel mapping accuracy is necessary. This could be done by evaluating the target registration error (TRE) based on landmarks. Comparison of performance among several algorithms using TRE has been widely reported by the extensive body of work found in the literature. These include multiple sites, such as head and neck [8–11], liver [12–14], thoracic [14–17] and prostate [14,18] sites. Thus far, however, none has attempted to assess the organ-wise correlation between segmentation metrics and voxel mapping accuracy for different algorithms.

Considering the tedious nature of evaluating voxel mapping accuracy, not to mention that it is not always possible to be carried out (subject to the availability of anatomical landmarks or phantoms), it would be advantageous to discover any previously unknown link between the more easily obtainable segmentation metrics and the voxel mapping accuracy. One of our past work [19] made use of a deformable

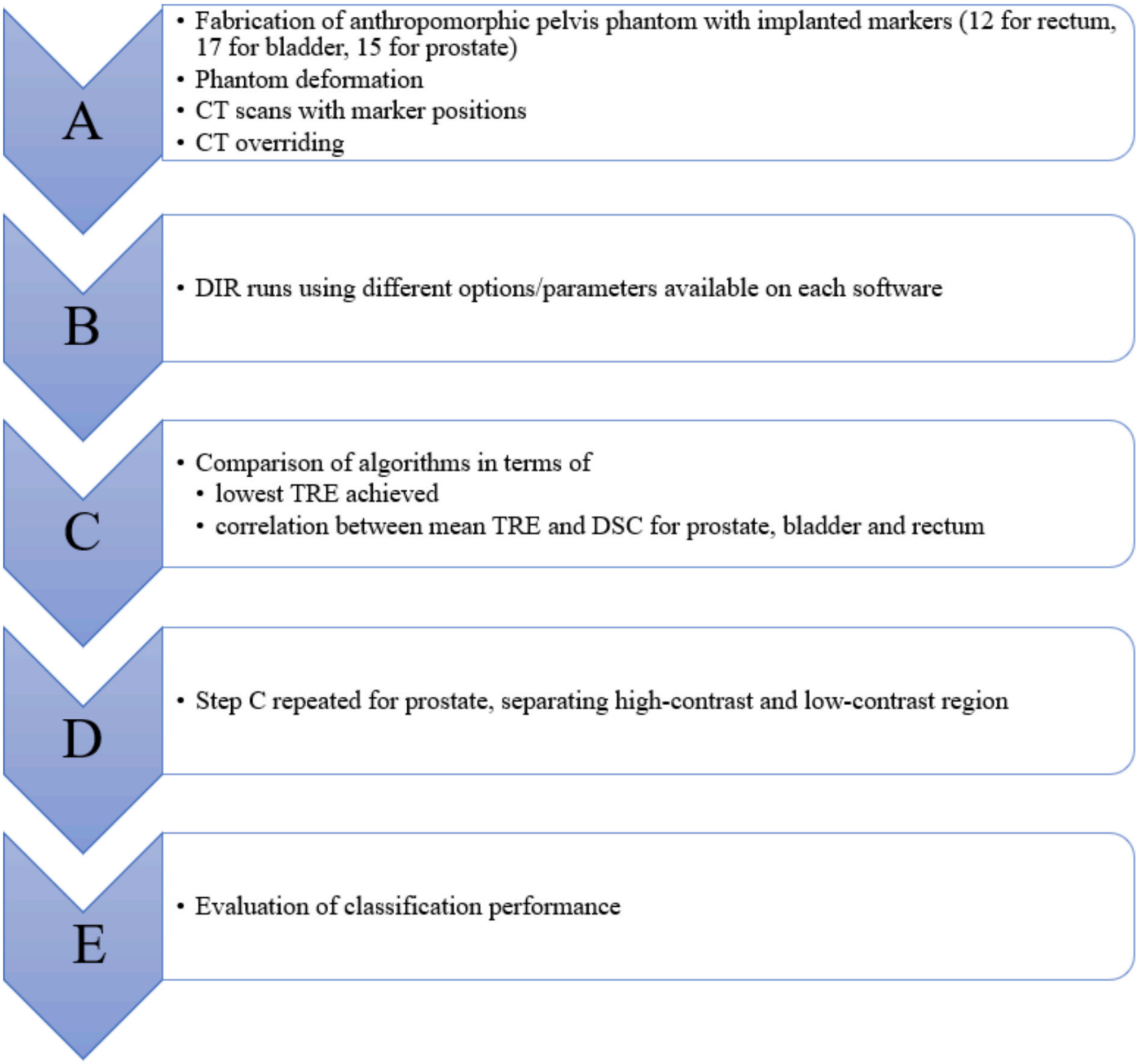

A
• Fabrication of anthropomorphic pelvis phantom with implanted markers (12 for rectum, 17 for bladder, 15 for prostate)
• Phantom deformation
• CT scans with marker positions
• CT overriding
B
• DIR runs using different options/parameters available on each software
C
• Comparison of algorithms in terms of
• lowest TRE achieved
• correlation between mean TRE and DSC for prostate, bladder and rectum
D
• Step C repeated for prostate, separating high-contrast and low-contrast region
E
• Evaluation of classification performance

Fig. 1. Summary of the study design, where A was covered in our past work [19] while B-E were done in this current work.

anthropomorphic pelvis phantom implanted with markers and demonstrated the feasibility of such an analysis on the hybrid intensity and structure based deformable registration in RayStation v10B (RaySearch Laboratories AB, Stockholm, Sweden). Using a similar approach, our current work set out to examine and compare the performance of three additional DIR algorithms, in terms of TRE and its correlation with Dice similarity coefficient (DSC), a widely used segmentation metric. This metric indicates the overlap of a certain region on two images, and is calculated from the contour delineated around the region. The ultimate goal of this work is twofold: 1) to identify algorithms with satisfactory TRE based on the tolerance set by TG132 [20], implying their suitability for dose mapping, and 2) to identify algorithms with a high correlation between DSC and TRE, indicating their potential for an easy DDA quality assurance (QA).

Methods

Fig. 1 summarises the study design, including key steps done in the previous work [19] and current work. The process of phantom fabrication, image acquisition and preprocessing steps have been detailed in the previous work [19], and hence will only be described briefly here. A deformable anthropomorphic phantom (Fig. 2**A**), consisting of prostate, bladder, rectum and pelvic bone, was fabricated through 3D printing. Markers were attached uniformly on the walls of prostate (8), bladder (17) and rectum (12) (Fig. 2**B**), as well as within the prostate (7). By simulating deformation of bladder and rectum, five CT scans (pixel spacing 0.885 mm, slice thickness 1 mm; example shown in Fig. 2**C**) were obtained from one static and four deformation scenarios. The four deformation scenarios comprise CT 2 – 5, where the rectum/bladder deformation for CT 2, CT 3, CT 4 and CT 5 were 25/30 ml, 25/60 ml, 50/30 ml and 50/60 ml, respectively. All the CT scans were then imported into 3D Slicer [21] to override the CT number according to the mean CT number obtained from the actual patient CT scan. This was done by creating "segments" in Slicer corresponding to each structure of interest (including the organs, organ walls and external region surrounding the organs), and subsequently replacing the voxel values within the segments. During this process, all the markers were erased from the images. These overridden images (example shown in Fig. 2**D**) were used as input images for the DIR algorithms. On top of RayStation which was studied earlier, three additional algorithms were tested: one on commercial software: Velocity v4.1 (Varian Medical Systems, Palo Alto, USA), and two on open-source software: Slicer [21] and Plastimatch [22]. The DIR details differ for each algorithm and will be described separately as follows.

Velocity

Velocity uses a B-spline transformation model and mutual information as the similarity metric. For each DIR, CT 1 was selected as the primary image (reference image to be deformed into) and CT 2 - 5 were selected in turn as the secondary image (image to be deformed). Five options of DIR algorithms available on Velocity were used: 1) Rigid Registration (RR) + Deformable Multi Pass (RR included by default), 2) Deformable, 3) Extended Deformable Multi Pass, 4) Structure Guided

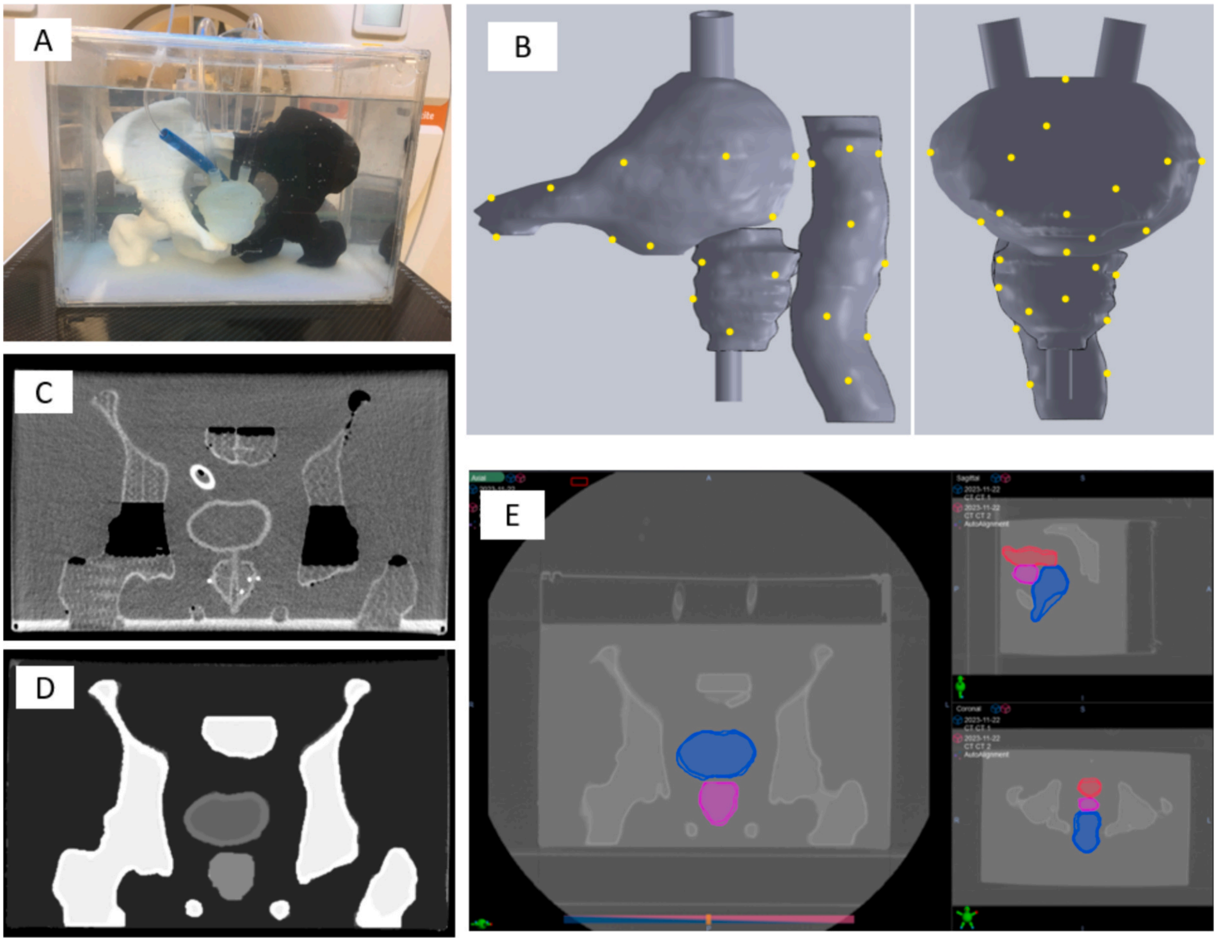


Fig. 2. (A) Pelvis phantom consisting of prostate, bladder, rectum and pelvic bone, fabricated via 3D printing. (B) Position of markers attached on the wall of prostate, bladder and rectum. (C) Example of phantom CT scan before overriding. The black regions are caused by air trapped within the pelvic bone. (D) Example of phantom CT scan after overriding. (E) Example of fused image with original contours and deformed contours.

Deformable, and 5) Extended Deformable Multi Pass + Structure Guided Deformable. For algorithm options 2 to 5, RR was done manually before performing DIR.

Slicer

There are several DIR packages in Slicer, including BRAINS, Advanced Normalization Tools (ANTs), and Elastix. Both ANTs and Elastix are available as extensions upon installation in Slicer, as SlicerANTs and SlicerElastix respectively. Since BRAINS and ANTs were both developed on the basis of registering brain images [23,24], Elastix was chosen to be the main focus of this study. Before the registrations, CT 1 was assigned as the fixed volume while CT 2 – 5 were successively assigned as the moving volumes.

Elastix offers a wide selection of registration presets, and the parameters in each preset are freely customizable by modifying the corresponding text file in the database folder. For this study, the generic preset, which utilizes a B-spline transformation, was used. To obtain a diverse range of DIR performances, a number of parameters were varied (Table 1**A)**, according to the guidelines stated in the manual. These parameters were varied in an "optimization" manner, using the mean TRE as the "objective function" to be minimized. In other words, the parameter setting which gave the lowest mean TRE would be considered the optimal setting. All the parameters were kept to the default settings at first, and modified one by one (from left to right on Table 1**A**), with the previous parameter being set to the optimal setting. This not only provided multiple DIRs with different mean TRE and DSC values (for subsequent correlation analysis), but also allowed for the best possible performance of the software.

Plastimatch

Plastimatch provides DIR via Demons, B-spline and landmark-based methods. As B-spline method has been tested in Velocity, and the CT images are devoid of landmarks (the markers were erased to prevent registration bias), only Demons method was examined in this case. Similar to SlicerElastix, CT 1 was the fixed image while CT 2 – 5 were the moving images. The parameters are also freely editable on the command file, and were varied in an "optimization" manner. Table 1**B** summarizes the parameters varied.

Table 1
Parameters varied for (A) the generic preset on SlicerElastix and (B) Plastimatch Demons.* signifies the default setting for each parameter. The use of mask for SlicerElastix is illustrated in Fig. S1 in the supplementary material.

(A) **Parameters**	Similarity metric	Number of resolutions	Final grid spacing (mm)	Maximum number of iterations	Mask
Values/ Options	MI*, NCC	4*, 5, 6	8, 16*, 32	500*, 1000, 1500, 2000	None*, around water region, around PBR

(B) **Parameters**	Std (mm)	Acc	Hmg	Filter width (voxels)	Maximum number of iterations
Values/ Options	2, 6*, 10, 14	1*, 3, 5, 7	1*, 5, 10	3*, 5, 7, 9, 11	30*, 100, 200, 300, 400, 500

Abbreviations: MI – Mutual Information; NCC – Normalized Cross Correlation; PBR – Prostate, Bladder and Rectum; Std – standard deviation of smoothing kernel; Acc – acceleration, representing the "gain" factor; Hmg – Homogenization, representing the tradeoff between gradient and image difference.

DIR Analysis

Upon completion of DIR, TRE can be determined via two means, as illustrated in Fig. 3. The first way is by calculating the Euclidean distance between the mapped marker positions (e.g. x' in Fig. 3) and the marker positions on the fixed image (e.g. x_F in Fig. 3). The second way is by finding the difference between the deformation vector given by the DIR algorithm ($x' - x_M$) at the marker coordinate x_M and the true deformation vector ($x_F - x_M$).

For SlicerElastix and Plastimatch, the first way was chosen for the analysis. Slicer allows users to apply transform to pre-defined points, either after performing registration in the software itself or using imported transform, as is the case with Plastimatch (in nrrd format). On the other hand, the second approach was used for Velocity. Velocity does not support integrated postprocessing of the DVF. Hence, the registration information was exported as Digital Imaging and Communications in Medicine (DICOM) format metadata and the DVF was read using Matlab R2024a. Interpolation was performed as needed to identify the deformation vector at a certain marker coordinate.

For all three software, the DSC and mean TRE of prostate, bladder and rectum resulting from each DIR were obtained. For SlicerElastix and Plastimatch, contours were drawn for each CT scan on Slicer, and the deformed contours were obtained by applying transform to the drawn contours (similar to how the mapped marker positions were obtained). The original contours on the fixed images and the deformed contours were exported in nrrd format. For Velocity, the contours were also drawn for each CT using the software itself. The original and deformed contours after DIR were exported in DICOM format and converted to nrrd format using Slicer. DSC was then computed using Plastimatch, for all three software. Fig. 2**E** shows an example of a fused image with the original contours on the fixed image and deformed contours overlaid on it.

Subsequently, the mean TRE was calculated from the TRE of all markers corresponding to each organ, i.e., 15 markers for prostate, 17 markers for bladder, and 12 markers for rectum. Pearson correlation analysis was done to test the null hypothesis that there is no correlation between DSC and mean TRE. A two tailed *P*-value of 0.05 marked the significance of the test.

Similar analysis was also done for prostate exclusively, separating the prostate wall (high-contrast region) and inner prostate (low-contrast region). This allowed us to specifically examine the results given by a homogeneous medium with low contrast, as the level of contrast may affect the DIR performance. Out of the 15 markers for prostate, eight markers were on the prostate wall and seven were within the prostate.

Classification Performance

To frame the correlation context into a classification task, DSC was used to classify the quality of voxel mapping into two classes: positive (mean TRE $\leq$ 3 mm) or negative (mean TRE $>$ 3 mm). Four different DSC thresholds were applied, namely 0.75, 0.80, 0.85 and 0.90, as these are close to the commonly used standard for an acceptable DSC range [20]. Several metrics, including accuracy, precision, recall, F1-score, specificity as well as area under the Receiver Operating Characteristic curve (AUC), were evaluated. The confidence intervals for AUC were calculated via bootstrapping with 2000 bootstrap samples.

Results

The results will be presented individually for the following analysis: parameter optimization (for SlicerElastix and Plastimatch only), mean marker movement and lowest mean TRE achieved, correlation analysis between DSC and mean TRE, and prostate analysis. The results for RayStation [19] were included here for comparison.

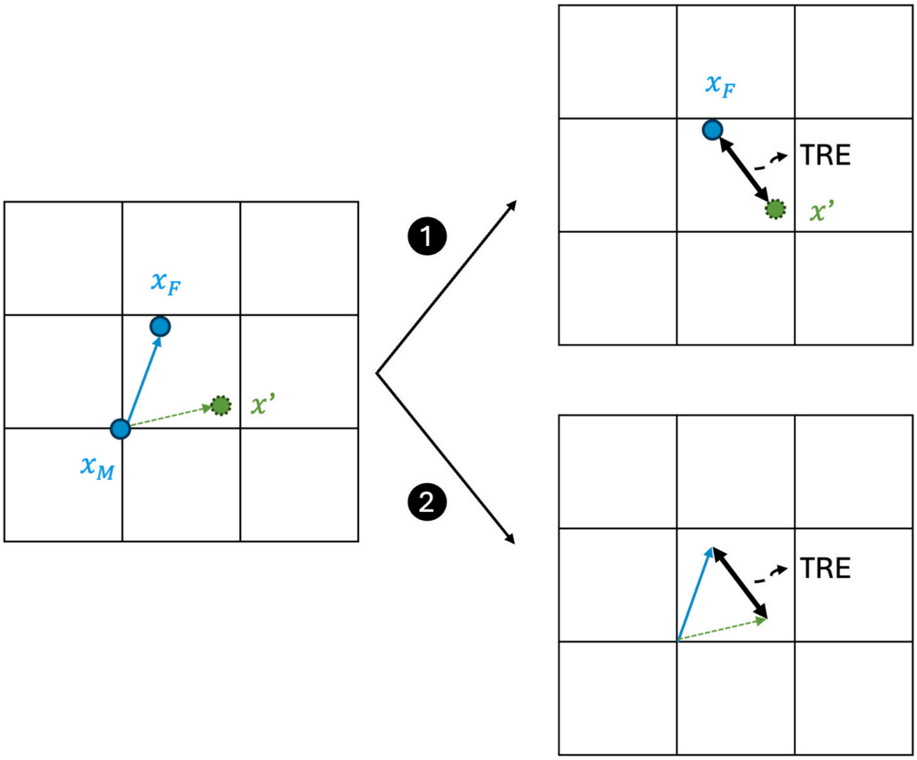


Fig. 3. Schematic showing two ways of obtaining target registration error (TRE) for a marker: 1) by calculating the Euclidean distance between the mapped marker position x′ and the marker position on the fixed image x_F, 2) by finding the difference between the deformation vector given by the DIR algorithm at the marker coordinate (green dashed arrow) and the true deformation vector (blue solid arrow). x_M represents the original marker position on the moving image.

Parameter Optimization

SlicerElastix

The upper panel in Table S1 in the supplementary material shows the parameters used during each DIR run on SlicerElastix, for the first deformation scenario. The results for other deformation scenarios were similar, but with a minor change: the second and fourth scenarios gave an optimal number of resolutions of 4 instead of 5. The difference in mean TRE was nonetheless subtle, within 0.03 mm (comparing Run 4 and Run 5). This implies that 4 or 5 resolutions is adequate for a B-spline transformation of images with voxel size of about 1 mm^3 and deformations of approximately 2.7 mm to 6.4 mm.

The final grid spacing determines the control point spacing of the B-spline transformation at the finest resolution level. A lower value allows a more flexible deformation, which may improve accuracy but can also result in an unrealistic deformation. Therefore, careful tuning is recommended and the optimal value of 16 mm obtained in this study may not be applicable to other cases (e.g. images with different voxel size, deformation extent, etc).

As expected, NCC was selected as the optimal similarity metric, since the DIR input images were of a single modality, i.e. CT scans. Also, a higher number of iterations leads to a lower mean TRE, indicating a higher voxel mapping accuracy. It is noteworthy that the use of mask greatly increases the DIR accuracy, depicted by the considerable reduction in mean TRE from Run 11 to Run 13 (Fig. 4**A**). Decreasing the coverage of the mask (from around the water region to around prostate, bladder and rectum) limits the registration focus to the region of interest, thereby improving the DIR performance.

Plastimatch

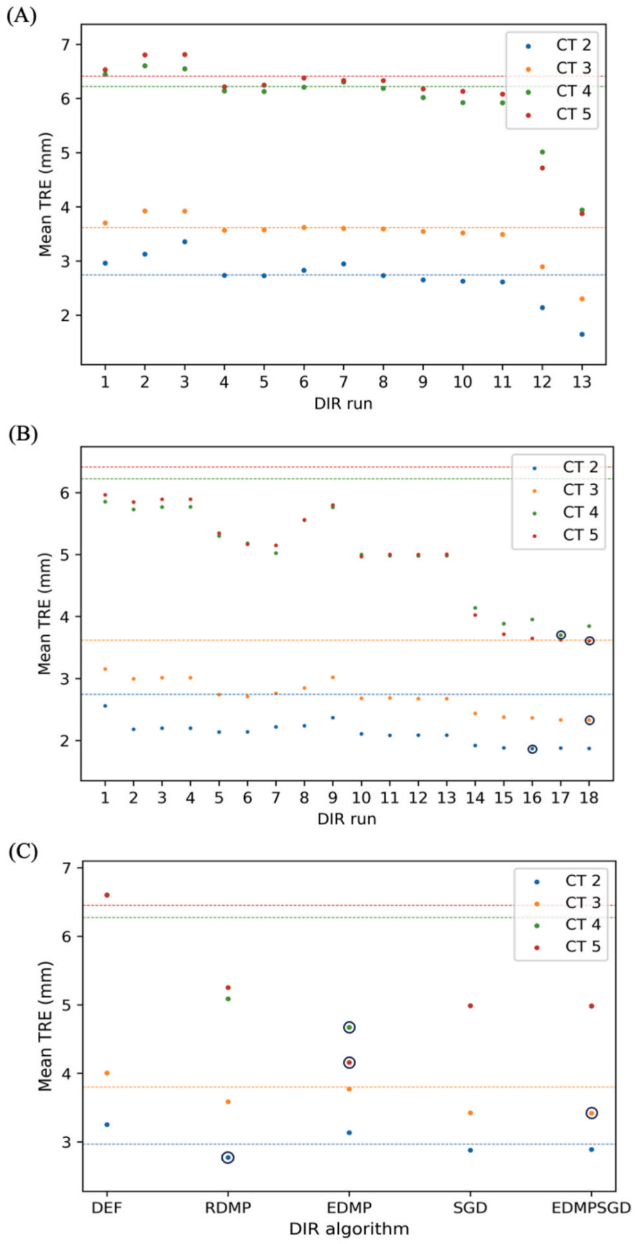
The lower panel in Table S1 shows the parameters used during each DIR run on Plastimatch, for the first deformation scenario. As Gaussian filtering is used for the Demons algorithm, the standard deviation of the filter kernel together with the filter width determines the smoothing extent on the deformation field. The optimisation returned an optimal standard deviation of 2 mm but different optimal filter width for each CT. Indeed, the necessary amount of smoothing depends on the image intensity distribution and how noisy the deformation field is. It is therefore not surprising that the optimal filter width is dependent on the nature and source of the data.

Acceleration dictates how fast the algorithm converges, and at the same time may affect the robustness of the results. In our study, a larger deformation entailed a larger optimal acceleration, which could indicate a possible causal relationship. However, as limited information was provided by Plastimatch on this parameter, our deduction is inconclusive and the effect of other factors could not be ruled out.

As stated in the image registration guidebook by Plastimatch [25], the homogenization value should increase with voxel sizes, going down to about 1 for 1 mm voxels. This agrees with our results, where 1 is the optimal homogenization for our CT images with voxel size of $0.885 \times 0.885 \times 1.0 mm$. The mean TRE generally decreased with more iterations (with minimal fluctuations), in line with our expectation. When comparing Runs 14 to 18 in Fig. 4**B**, it is observed that the mean TRE plateaued more quickly for smaller deformation scenarios (i.e. CT 2 and CT 3).

Mean marker movement and lowest mean TRE achieved

Table 2 shows the mean marker movements and the lowest mean TRE given by each software (including RayStation) in different deformation scenarios. Comparing across the software, it is clear that RayStation achieved the best performance, with the mean TREs reduced to about 50% of the corresponding mean marker movements. Conversely, Velocity had the largest values for all scenarios, and the algorithm option giving the lowest mean TRE varied according to the deformation scenario (Fig. 4**C**).



Acronyms: DEF - Deformable; RDMP – Rigid Registration + Deformable Multi Pass; EDMP – Extended Deformable Multi Pass; SGD – Structure Guided Deformable; EDMPSGD – Extended Deformable Multi Pass.

Fig. 4. Mean target registration error (TRE) obtained from each DIR run/algorithm option on (A) SlicerElastix, (B) Plastimatch and (C) Velocity, for all four deformation scenarios. The dashed lines represent the mean marker movement while the black circles mark out the lowest mean TRE achieved in each scenario (corresponding to data tabulated in Table S1).

Correlation analysis between DSC and mean TRE

The relationship between DSC and mean TRE of each organ, resulting from different deformation scenarios and DIRs, were plotted for each software (including RayStation) as shown in Fig. 5. Generally, it can be seen that the mean TRE displayed a decreasing trend with increasing DSC. For Velocity and Plastimatch, however, the data points were more dispersed above or below the regression lines. This suggests weaker correlations overall, which was indeed confirmed by the Pearson correlation tests (Table 3). The negative correlation strength ranged from moderate ($-0.60 < r \leq -0.40$) to very strong ($r \leq -0.80$) for Velocity, and strong ($-0.80 < r \leq -0.60$) to very strong for Plastimatch. In contrast, both RayStation and SlicerElastix achieved very strong negative correlation between the two metrics for all the organs/combination considered.

Table 2
Mean marker movement and lowest mean target registration error (TRE) achieved by each DIR software for all four deformation scenarios. The mean marker movements for Velocity are reported behind the slash symbol(/).

CT	Mean marker movement (mm)	Lowest mean TRE achieved (mm)			
		RayStation	Velocity	SlicerElastix	Plastimatch
CT 2	2.74 ± 0.24/ 2.97 ± 0.23	1.48 ± 0.16	2.77 ± 0.23	1.65 ± 0.15	1.87 ± 0.14
CT 3	3.62 ± 0.28/ 3.80 ± 0.27	2.03 ± 0.21	3.42 ± 0.25	2.30 ± 0.18	2.33 ± 0.20
CT 4	6.22 ± 0.53/ 6.27 ± 0.51	2.61 ± 0.32	4.67 ± 0.31	3.94 ± 0.32	3.70 ± 0.34
CT 5	6.41 ± 0.47/ 6.45 ± 0.47	3.06 ± 0.31	4.16 ± 0.33	3.88 ± 0.36	3.61 ± 0.36

Prostate Analysis

Similar to what has been discussed in **Section 3.2**, RayStation performed the best in terms of voxel mapping accuracy (Table 4). For the entire range of prostate marker movements (up to approximately 8 mm), RayStation gave mean TREs that were close to or lower than 2 mm.

Interestingly, the homogeneous medium within the prostate did not give rise to a terribly erroneous voxel mapping. In fact, a quick evaluation of the error percentages relative to the marker movements would reveal that most of the software mapped the markers within the prostate even more accurately than those at the prostate wall.

Pearson correlation tests done between DSC and mean TRE, separately for markers within the prostate and on the prostate wall, showed very strong negative correlation for all the software tested, with the exception of one case; for markers within the prostate, DIR using Velocity yielded a comparatively lower correlation strength between DSC and mean TRE (Table 5). This indicates that among all, Velocity is most prone to mapping voxels within a homogeneous medium independently of the contour region with higher contrast.

Classification Performance

The accuracy, precision, recall, F1-score, specificity and AUC were reported in Table 6. The results marked with '-' represent cases with no true positive, false positive or false negative. Generally, a higher DSC threshold resulted in a better classification performance. It can also be seen that SlicerElastix achieved the best performance overall. For all three organs, the scores at their corresponding best DSC threshold were at least 0.70, with most above 0.80, except for the precision and F1-score of the bladder. Meanwhile, the AUCs were all close to 1. This agrees with the earlier finding where SlicerElastix displayed the highest correlation between DSC and mean TRE, thus allowing a more accurate classification based on the DSC value.

Discussion

Through this study, we assessed and compared four DIR algorithms by quantifying the mean TREs and their correlation with DSC. These analyses were done for bladder and rectum (important organs at risk in prostate RT), and one specifically focused on prostate, to investigate the DIR performance in both high-contrast and low-contrast regions.

We would like to highlight a few findings that are congruent with those in the previous work [19]: 1) for all software, the lowest mean TREs achieved generally had a larger magnitude with larger mean marker movements (Table 2), implying the challenges associated with

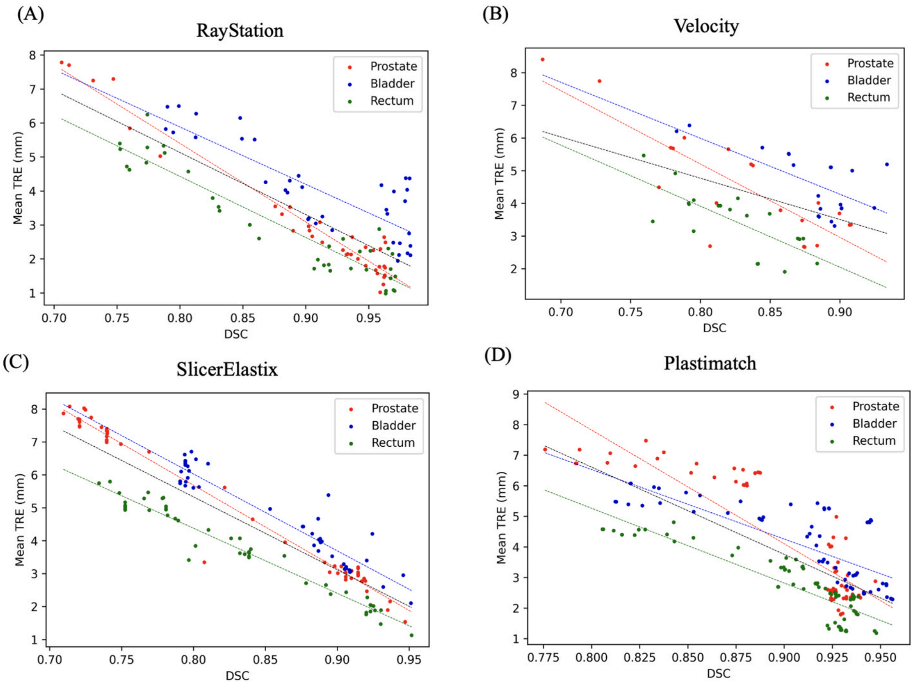


Fig. 5. Plots of mean target registration error (TRE) vs Dice similarity coefficient (DSC) for prostate, bladder and rectum, using DIR on (A) RayStation, (B) Velocity, (C) SlicerElastix, and (D) Plastimatch. The red, blue, and green dashed lines are the regression lines for prostate, bladder, and rectum, respectively, while the black dashed line represents the regression line for all the data points.

Table 3
Pearson correlation coefficient (r) between DSC and mean TRE for prostate, bladder, rectum, and all three organs together, for each DIR software. The *P*-values for all the reported r's were <0.01.

	RayStation	Velocity	SlicerElastix	Plastimatch
Prostate	-0.977	-0.836	-0.987	-0.874
Bladder	-0.829	-0.668	-0.948	-0.786
Rectum	-0.936	-0.740	-0.978	-0.880
All	-0.835	-0.503	-0.901	-0.761

Table 4
Mean marker movement and lowest mean target registration error (TRE) achieved by each DIR software for all four deformation scenarios, quantified using markers within the prostate (upper panel) and on the prostate wall (lower panel). The mean marker movements for Velocity are reported behind the slash symbol(/).

CT	Mean marker movement (mm)	Lowest mean TRE achieved (mm)			
		RS	VL	SE	PT
CT 2	3.00 ± 0.12/2.94 ± 0.08	0.85	2.20	1.24	1.64
CT 3	2.82 ± 0.24/3.19 ± 0.32	1.41	3.09	1.50	2.02
CT 4	7.46 ± 0.48/7.58 ± 0.50	1.43	3.26	2.72	3.38
CT 5	6.92 ± 0.25/7.17 ± 0.25	1.51	2.89	2.37	2.73

CT	Mean marker movement (mm)	Lowest mean TRE achieved (mm)			
		RS	VL	SC	PT
CT 2	3.00 ± 0.10/3.06 ± 0.11	1.17	2.83	1.80	1.92
CT 3	3.36 ± 0.34/3.65 ± 0.33	1.51	3.55	2.24	2.49
CT 4	7.98 ± 0.64/8.01 ± 0.58	2.01	5.57	5.04	4.55
CT 5	7.55 ± 0.46/7.46 ± 0.56	2.03	4.99	4.21	3.58

Abbreviations: RS – RayStation; VL – Velocity; SE – SlicerElastix; PT – Plastimatch.

Table 5
Pearson correlation coefficient (r) between DSC and mean TRE of markers within the prostate (labelled as IN) and on the prostate wall (labelled as OUT), for each DIR software. The *P*-values for all the reported r's were <0.01.

	RayStation	Velocity	SlicerElastix	Plastimatch
IN	-0.983	-0.774	-0.977	-0.893
OUT	-0.952	-0.858	-0.987	-0.854

larger deformation scenarios, 2) the correlation between DSC and mean TRE for bladder was relatively weaker compared to prostate and rectum, regardless of the DIR software used (Table 3).

Elaborating on the second point, it is noticed that blue points (representing bladder) in Fig. 5 were scattered above the regression lines at DSC>0.90. Despite the high DSCs, the mean TREs ranged up to 5 or 6 mm. Bearing in mind the volume dependence of DSC as reported in past studies [26,27], this observation once again demonstrated the caveat with respect to larger organs, where a high DSC does not necessarily guarantee a high voxel mapping accuracy. This could also explain the low precision for bladder (Table 6), observed across all the software.

Among all the tested software, SlicerElastix manifested a remarkably strong correlation between DSC and mean TRE for all organs (Table 3). All the correlation coefficients were lower than -0.90, indicating its potential for easy evaluation of voxel mapping accuracy through segmentation metrics. Although the correlations were weaker for the remaining software, all but one of the recorded values lay in the strong to very strong correlation range. This finding may disagree with a couple of past studies [28,29], but could be attributed to the regularization factor which has been accounted for in all the tested algorithms. RayStation uses an objective function consisting of an image similarity term, grid regularization terms, and anatomical penalty terms. Velocity imposes constrained regularization to prevent abnormal voxel behaviours such as jumping and folding. SlicerElastix allows the tuning of the B-spline model final grid spacing, to attain a balance between the flexibility and physical plausibility of the deformation. Lastly, the Demons algorithm implemented on Plastimatch includes a Gaussian filter to smooth the resulting DVF.

Despite having considered a regularization factor, Velocity displayed the lowest agreement between DSC and mean TRE (Table 3). This, we speculate, could be due to the inherent regularization not being equally

Table 6
Evaluation metric scores, stratified by software and organ, for classification task using DSC to determine whether the mean TRE exceeded 3 mm (positive) or not (negative). The 95% confidence interval for each AUC is shown in the bracket.

		RayStation				**Velocity**				**SlicerElastix**				**Plastimatch**			
DSC Threshold		**0.75**	**0.8**	**0.85**	**0.9**	**0.75**	**0.8**	**0.85**	**0.9**	**0.75**	**0.8**	**0.85**	**0.9**	**0.75**	**0.8**	**0.85**	**0.9**
P	**Acc**	0.83	0.89	0.89	0.94	0.45	0.65	0.80	0.75	0.71	0.73	0.79	0.87	0.49	0.56	0.64	0.85
	Pre	0.81	0.87	0.87	0.96	0.39	0.50	0.67	1.00	0.52	0.53	0.59	0.70	0.49	0.52	0.57	0.76
	Rec	1.00	1.00	1.00	0.96	1.00	1.00	0.86	0.29	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	F1	0.90	0.93	0.93	0.96	0.56	0.67	0.75	0.44	0.68	0.70	0.74	0.82	0.65	0.69	0.73	0.86
	Spe	0.40	0.60	0.60	0.90	0.15	0.46	0.77	1.00	0.58	0.61	0.69	0.81	0	0.14	0.30	0.70
	AUC	0.98 (0.94 – 1.00)				0.84 (0.60 – 1.00)				0.96 (0.91 – 0.99)				0.90 (0.82 – 0.96)			
B	**Acc**	0.25	0.36	0.47	0.67	0	0.10	0.15	0.70	0.06	0.42	0.48	0.81	0.32	0.32	0.47	0.60
	Pre	0.25	0.28	0.32	0.43	0	0	0	0	0.06	0.09	0.10	0.23	0.32	0.32	0.38	0.44
	Rec	1.00	1.00	1.00	1.00	-	-	-	-	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	F1	0.40	0.44	0.49	0.60	-	-	-	-	0.11	0.17	0.18	0.38	0.48	0.48	0.55	0.61
	Spe	0	0.15	0.30	0.56	0	0.10	0.15	0.70	0	0.39	0.45	0.80	0	0	0.22	0.41
	AUC	0.89 (0.77 – 0.98)				-				1.00 (1.00 – 1.00)				0.91 (0.84 – 0.97)			
R	**Acc**	0.61	0.86	0.97	0.97	0.20	0.55	0.75	0.80	0.35	0.73	0.98	0.94	0.64	0.64	0.79	0.85
	Pre	0.61	0.81	0.96	1.00	0.20	0.31	0.40	-	0.31	0.52	0.94	1.00	0.64	0.64	0.75	0.83
	Rec	1.00	1.00	1.00	0.95	1.00	1.00	0.50	0	1.00	1.00	1.00	0.80	1.00	1.00	1.00	0.96
	F1	0.76	0.90	0.98	0.98	0.33	0.47	0.44	-	0.47	0.68	0.97	0.89	0.78	0.78	0.86	0.89
	Spe	0	0.64	0.93	1.00	0	0.44	0.81	1.00	0.08	0.62	0.97	1.00	0	0	0.42	0.65
	AUC	1.00 (1.00 – 1.00)				0.83 (0.61 – 1.00)				1.00 (1.00 – 1.00)				0.98 (0.94 – 1.00)			

Abbreviations: P – Prostate; B – Bladder; R – Rectum; Acc – Accuracy; Pre – Precision; Rec – Recall; F1 – F1-score; Spe – Specificity; AUC – Area under the Receiver Operating Characteristic curve.

effective in all clinical scenarios. Introducing a regularization function which allows user tuning may prove useful in this case. Besides, the external processing and interpolation of DVF might have induced uncertainties on the mapped marker positions, and hence the TREs for Velocity.

As Velocity does not allow import of external marker positions, the marker positions were redefined on the Velocity interface, causing the mean marker movements recorded to be slightly different (Table 2, Table 4). This variability in defining marker positions represents a source of uncertainty arising from random human errors. Nonetheless, this uncertainty should have been taken into account by the standard error of the mean values, as the error bars for the two sets of values are seen to be overlapping.

Benchmarking the lowest mean TREs achieved against the 3 mm upper limit recommended by TG132 [20], it is observed that within the mean marker movement range of 2.74 – 6.41 mm, only RayStation managed to satisfy this criterion (Table 2). This uncovered the potential of RayStation being a more reliable tool for performing DDA, as compared to other tested software. Within the homogeneous medium of the inner prostate, RayStation and SlicerElastix were able to map the markers with an accuracy of less than 3 mm, while the largest mean TREs for Velocity and Plastimatch were slightly above 3 mm (Table 4).

A past study by Kirby *et al.* [29] using a two-dimensional deformable pelvic phantom reported a decent Velocity performance, where it achieved the lowest percentage of registration error above 3 mm and the lowest mean error. In contrast, our current study found that Velocity had the lowest voxel mapping accuracy (highest mean TRE for all deformation scenarios) among the tested software. This could presumably be caused by the way of error quantification, as the past study took into account the whole pelvic region, while our study localized the evaluation to the three organs of interest, i.e. prostate, bladder and rectum. In fact, taking a close look at the results of another related study by Nie *et al.* [30] would reveal that the percentage of larger errors (4 mm and above) were higher in regional analysis (which focused on localized evaluation) compared to global analysis.

Our findings on the performance of various commercial and open-source DIR algorithms, as quantified by TRE and DSC, contribute directly to addressing the critical issue of DIR uncertainties in radiotherapy. As highlighted by Nenoff *et al.* [31], quantifying these uncertainties is paramount, yet there is currently no universal consensus within the radiotherapy community on how to effectively do so or to establish thresholds for acceptable DIR results. Our phantom-based approach offers a standardized method for evaluating algorithm performance under controlled conditions, thereby providing valuable insights into the inherent variability and potential inaccuracies of different DIR solutions. The observed discrepancies in TRE and DSC among the algorithms underscore the need for careful commissioning and QA of DIR systems, echoing the recommendations made by Nenoff *et al.* [31] for handling uncertainties.

Due to the overriding of CT numbers, it should be acknowledged that the image contrast could be either higher (e.g. at the organ boundary region) or lower (e.g. within the prostate) than actual patient data. That being said, our study had allowed investigation into these two extreme conditions, thus we believe that these results would still offer useful insights into the DIR performance of various algorithms given different image quality and contrast.

It is important to note that the results presented here are solely applicable to the pelvic site; for other cancer sites with different image heterogeneity and deformation nature, the DIR software performance may vary, hence warranting further study. In addition, our current study only covered two commercial and two open-source software. Further research could be done on other commonly used commercial software, including Eclipse (Varian Medical Systems, Palo Alto, USA) and MIM (MIM Software Inc., OH, USA), to facilitate the clinical uptake of DDA. It may also be worthwhile to consider these findings within the rapidly evolving landscape of medical image registration. Recent breakthroughs in deep learning have introduced novel solutions for both affine [32] and deformable registration [33] tasks, sometimes surpassing conventional methods, particularly in terms of efficiency and generalizability across different anatomical sites and imaging modalities. Future research could therefore benefit from directly comparing the performance of the traditional DIR algorithms evaluated in this study against these state-of-the-art deep learning architectures using similar phantom-based validation methodologies. Such comparative studies would be crucial in guiding the selection of optimal registration strategies for various clinical applications and in harnessing the full potential of AI in radiation oncology.

Conclusion

In this work, we quantified and compared the voxel mapping accuracy and its correlation with DSC for four DIR software. Among all, RayStation achieved the highest voxel mapping accuracy, indicating the great promise it holds for an accurate dose mapping. For SlicerElastix, DSC was found to be an excellent indicator of voxel mapping accuracy. This means that a segmentation metric method for DDA QA is sufficient for this software.

Funding support

Hong Qi Tan is supported by the Duke-NUS Oncology Academic Program Goh Foundation Proton Research Programme (08/FY2023/EX (SL)/163-A218(b)), Clinical & Systems Innovation Support – Innovation Seed Grant (08/FY2022/P2/02-A68)

CRediT authorship contribution statement

Yun Ming Wong: Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Wen Siang Lew:** Writing – review & editing, Supervision, Resources, Project administration. **James Cheow Lei Lee:** Writing – review & editing, Supervision, Project administration. **Hong Qi Tan:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.zemedi.2025.09.001.

References

[1] Rigaud B, Simon A, Castelli J, et al. Deformable image registration for radiation therapy: principle, methods, applications and evaluation. Acta Oncol 2019;58(9): 1225–37. https://doi.org/10.1080/0284186X.2019.1620331.

[2] Sotiras A, Davatzikos C, Paragios N. Deformable medical image registration: a survey. IEEE Trans Med Imaging 2013;32(7):1153–90. https://doi.org/10.1109/TMI.2013.2265603.

[3] Wu Z, Lan T, Wang J, Ding Y, Qin Z. Medical image registration using B-spline transform. Int J Simul Syst Sci Technol 2016;17:1.1–6. https://doi.org/10.5013/IJSSST.a.17.48.01.

[4] Rueckert D, Sonoda LI, Hayes C, Hill DL, Leach MO, Hawkes DJ. Nonrigid registration using free-form deformations: application to breast MR images. IEEE Trans Med Imaging 1999;18(8):712–21. https://doi.org/10.1109/42.796284.

[5] Pennec X, Cachier P, Ayache N. Understanding the "Demon's Algorithm": 3D Non-rigid Registration by Gradient Descent. In: Taylor C, Colchester A, eds. *Medical Image Computing and Computer-Assisted Intervention – MICCAI'99*. Vol 1679. Lecture Notes in Computer Science. Springer Berlin Heidelberg; 1999:597-60doi:10.1007/10704282_64.

[6] Thirion JP. Image matching as a diffusion process: an analogy with Maxwell's demons. Med Image Anal 1998;2(3):243–60. https://doi.org/10.1016/s1361-8415(98)80022-4.

[7] Wang H, Dong L, O'Daniel J, et al. Validation of an accelerated "demons" algorithm for deformable image registration in radiation therapy. Phys Med Biol 2005;50(12):2887–905. https://doi.org/10.1088/0031-9155/50/12/011.

[8] Rigaud B, Simon A, Castelli J, et al. Evaluation of deformable image registration methods for dose monitoring in head and neck radiotherapy. Biomed Res Int 2015; 2015(1):726268. https://doi.org/10.1155/2015/726268.

[9] Pukala J, Johnson PB, Shah AP, et al. Benchmarking of five commercial deformable image registration algorithms for head and neck patients. J Appl Clin Med Phys 2016;17(3):25–40. https://doi.org/10.1120/jacmp.v17i3.5735.

[10] Kubli A, Pukala J, Shah AP, et al. Variability in commercially available deformable image registration: A multi-institution analysis using virtual head and neck phantoms. J Appl Clin Med Phys 2021;22(5):89–96. https://doi.org/10.1002/acm2.13242.

[11] Singhrao K, Kirby N, Pouliot J. A three-dimensional head-and-neck phantom for validation of multimodality deformable image registration for adaptive radiotherapy. Med Phys 2014;41(12):121709. https://doi.org/10.1118/1.4901523.

[12] Fukumitsu N, Nitta K, Terunuma T, et al. Registration error of the liver CT using deformable image registration of MIM Maestro and Velocity AI. BMC Med Imaging 2017;17(1):30. https://doi.org/10.1186/s12880-017-0202-z.

[13] Sen A, Anderson BM, Cazoulat G, et al. Accuracy of deformable image registration techniques for alignment of longitudinal cholangiocarcinoma CT images. Med Phys 2020;47(4):1670–9. https://doi.org/10.1002/mp.14029.

[14] Brock KK. Results of a multi-institution deformable registration accuracy study (MIDRAS). Int J Radiat Oncol Biol Phys 2010;76(2):583–96. https://doi.org/10.1016/j.ijrobp.2009.06.031.

[15] Kadoya N, Fujita Y, Katsuta Y, et al. Evaluation of various deformable image registration algorithms for thoracic images. J Radiat Res (Tokyo) 2014;55(1): 175–82. https://doi.org/10.1093/jrr/rrt093.

[16] Kadoya N, Nakajima Y, Saito M, et al. Multi-institutional validation study of commercially available deformable image registration software for thoracic images. Int J Radiat Oncol Biol Phys 2016;96(2):422–31. https://doi.org/10.1016/j.ijrobp.2016.05.012.

[17] Han MC, Kim J, Hong CS, et al. Performance evaluation of deformable image registration algorithms using computed tomography of multiple lung metastases. Technol Cancer Res Treat 2022;21:15330338221078464. https://doi.org/10.1177/15330338221078464.

[18] Motegi K, Tachibana H, Motegi A, Hotta K, Baba H, Akimoto T. Usefulness of hybrid deformable image registration algorithms in prostate radiation therapy. J Appl Clin Med Phys 2019;20(1):229–36. https://doi.org/10.1002/acm2.12515.

[19] Wong YM, Koh CWY, Lew KS, et al. Deformable anthropomorphic pelvis phantom for dose accumulation verification. Phys Med Biol 2024;69(12):12NT01. https://doi.org/10.1088/1361-6560/ad52e4.

[20] Brock KK, Mutic S, McNutt TR, Li H, Kessler ML. Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132. *Med Phys*. 2017;44(7):e43-e76. doi: 10.1002/mp.12256.

[21] 3D Slicer image computing platform. 3D Slicer. Accessed January 11, 2024. https://slicer.org/.

[22] Sharp G, LI R, Wolfgang J, et al. *PLASTIMATCH– AN OPEN SOURCE SOFTWARE SUITE FOR RADIOTHERAPY IMAGE PROCESSING.*; 2010.

[23] Johnson H, Harris G, Williams K. BRAINSFit: Mutual Information Registrations of Whole-Brain 3D Images, Using the Insight Toolkit. *Insight J*. Published online October 5, 2007. doi:10.54294/hmb052.

[24] Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage 2011;54(3):2033–44. https://doi.org/10.1016/j.neuroimage.2010.09.025.

[25] Image registration guidebook — Plastimatch 1.10.0 documentation. Accessed October 2, 2024. https://plastimatch.org/image_registration_guidebook.html.

[26] Kumarasiri A, Siddiqui F, Liu C, et al. Deformable image registration based automatic CT-to-CT contour propagation for head and neck adaptive radiotherapy in the routine clinical setting. Med Phys 2014;41(12):121712. https://doi.org/10.1118/1.4901409.

[27] Deeley MA, Chen A, Datteri R, et al. Comparison of manual and automatic segmentation methods for brain structures in the presence of space-occupying lesions: a multi-expert study. Phys Med Biol 2011;56(14):4557–77. https://doi.org/10.1088/0031-9155/56/14/021.

[28] Rohlfing T. Image Similarity and Tissue Overlaps as Surrogates for Image Registration Accuracy: Widely Used but Unreliable. IEEE Trans Med Imaging 2012; 31(2):153–63. https://doi.org/10.1109/TMI.2011.2163944.

[29] Kirby N, Chuang C, Ueda U, Pouliot J. The need for application-based adaptation of deformable image registration. Med Phys 2013;40(1):011702. https://doi.org/10.1118/1.4769114.

[30] Nie K, Chuang C, Kirby N, Braunstein S, Pouliot J. Site-specific deformable imaging registration algorithm selection using patient-based simulated deformations. Med Phys 2013;40(4):041911. https://doi.org/10.1118/1.4793723.

[31] Nenoff L, Amstutz F, Murr M, et al. Review and recommendations on deformable image registration uncertainties for radiotherapy applications. Phys Med Biol 2023;68(24):24TR01. https://doi.org/10.1088/1361-6560/ad0d8a.

[32] Strittmatter A, Schad LR, Zöllner FG. Deep learning-based affine medical image registration for multimodal minimal-invasive image-guided interventions – A comparative study on generalizability. Z Für Med Phys 2024;34(2):291–317. https://doi.org/10.1016/j.zemedi.2023.05.003.

[33] Strittmatter A, Zöllner FG. Multistep networks for deformable multimodal medical image registration. IEEE Access 2024;12:82676–92. https://doi.org/10.1109/ACCESS.2024.3412216.