Original Article

Objective Assessment of the Quality and Accuracy of Deformable Image Registration

Ines-Ana Jurkovic, Nikos Papanikolaou, Sotirios Stathakis, Neil Kirby, Panayiotis Mavroidis[1]

Department of Radiation Oncology, University of Texas Health Science Center at San Antonio, San Antonio, TX, [1]Department of Radiation Oncology, University of North Carolina, Chapel Hill, North Carolina, USA

Abstract

Background: The increased use of deformable registration algorithms in clinical practice has also increased the need for their validation. **Aims and Objectives:** The purpose of the study was to investigate the quality, accuracy, and plausibility of three commercial image registration algorithms for 4-dimensional computed tomography (4DCT) datasets using various similarity measures. **Materials and Methods:** 4DCT datasets were acquired for 10 lung cancer patients. 23 similarity measures were used to evaluate image registration quality. To ensure selected method's invertibility and assess resultant mechanical stress, the determinant of the Jacobian for the displacement field and 3-D Eulerian strain tensor were calculated. All the measures and calculations were applied on to extended deformable multi pass (EXDMP) and deformable multi pass (DMP) methods. **Results:** The results indicate the same trend for several of the studied measures. The Jacobian determinant values were always positive for the DMP method. The Eulerian strain tensor had smaller values for the DMP method than EXDMP in all of the studied cases. The negative values of the Jacobian determinant point to non-physical behavior of the EXDMP method. The Eulerian strain tensor values indicate less tissue strain for the DMP method. Large differences were also observed in the results between complete and cropped datasets (coefficient of determination: 0.55 vs. 0.93). **Conclusion:** A number of error and distance measures showed the best performance among the tested measures. The evaluated measures might detect CT dataset differences with higher precision if the analysis is restricted to a smaller volume.

Keywords: 4DCT, deformable image registration, image dissimilarity indices; Jacobian determinant, image similarity measures, strain tensor

Received on: 31-05-2019 **Review completed on: 09-07-2020** **Accepted on: 14-07-2020** **Published on: 13-10-2020**

Introduction

Deformable image registration (DIR) is extensively used in radiation therapy applications. Possible DIR uses include auto-segmentation of structures,[1] dose accumulation,[2] and treatment optimization. A multi-institution study[3] suggested caution in globally accepting the results from deformable registration. Studies that evaluated different DIR algorithms, including the most common commercial software, showed reasonable overall accuracy of the registration; however, they also observed large DIR errors in some of the studied cases.[3-5] Thus, it is necessary to assess accuracy of DIR software before it is used for any of its radiation therapy applications. Several studies suggested different ways of quantifying the registration results.[3,6-8] Many of the studies in this field use patient-specific phantoms to evaluate DIR algorithms by looking at fixed points in the images (control points) and measuring their displacement in the registration resultant image (landmark error). Landmark matching is a process that has been applied on large image datasets; however, the limited number of images that are usually analyzed may not give a complete and accurate picture of DIR performance on the whole volume. Furthermore, identifying landmarks is time consuming process as well as not feasible for every patient. Beyond this, there is need for a fast and reliable metrics that can be automatically applied on large volume datasets and can provide a dependable tool for the DIR assessment. Many publications examined the Dice, Tanimoto, or mutual information (MI) measures, but some of these measures proved to be unreliable when used as the only tool for DIR assessment. For example, Dice can

Access this article online

Quick Response Code:

Website: www.jmp.org.in

DOI: 10.4103/jmp.JMP_47_19

Address for corresponding: Dr. Panayiotis Mavroidis, Department of Radiation Oncology, University of North Carolina, Chapel Hill, North Carolina, USA.
E-mail: panayiotis_mavroidis@med.unc.edu

How to cite this article: Jurkovic IA, Papanikolaou N, Stathakis S, Kirby N, Mavroidis P. Objective assessment of the quality and accuracy of deformable image registration. J Med Phys 2020;45:156-67.

show improved similarity when compared with the original unwrapped data[9] but nonphysical behavior may exist, which may strongly influence the accumulated dose distributions and/or contour segmentation. The term "nonphysical behavior" refers to the situations where the produced registration features are physically impossible based on the geometrical and mechanical characteristics of the involved tissues. Hence, to assess for possible nonphysical behavior of DIR, it is essential to understand tissue mechanics.[10,11]

This study utilizes a combination of metrics to assess both image similarity and for nonphysical deformation scenarios to evaluate DIR accuracy. More specifically, a broad range of image similarity measures are utilized here, which include methods previously applied in radiation oncology for this purpose and also several similarity measures[3,6,12-16] that were developed and utilized in other fields.

In deformable registration, each voxel occupies its own location in space for both the initial and deformed configuration. However, no material element should be permitted to invert as it leads to nonphysical transformations.[17] The goal of the transformation is to be plausible or at least locally invertible. For any transformation to satisfy this requirement, according to the inverse function theorem, it is sufficient for it to be continuously differentiable and have a positive Jacobian determinant (J). As a measure to estimate the expansion and contraction during the deformation (i.e., volume change), the Jacobian determinant is widely used.[18] To avoid a region of positive finite volume to be deformed into a region of zero, negative (folding), or infinite volume, it is required that $0< J <\infty$. Tissues are composite materials that are continually changing and their behavior is described by continuum models, which have been developed and used in continuum mechanics and biomedical engineering. To describe the kinematics and mechanics of deformations different strain measures are employed as well as measures of volume and surface changes. The most common strain measures are the Lagrangian or material strain tensor and the Eulerian or spatial strain tensor, which are defined by means of the deformation gradient as the basic measure of local deformation and rotational motion. The Jacobian determinant and the Eulerian strain tensors are used here to evaluate for nonphysical deformation scenarios.

Materials and Methods

Patient selection

To determine which measure has the best correlation with DIR performance, the metrics were applied to 4DCT lung datasets. For each patient 10 computed tomography (CT) datasets were available (the breathing cycle was sampled at 10 different phases). The work was approved by the appropriate ethical committees related to the institution in which it was performed and that subjects gave informed consent to the work. For evaluation purposes, the phase 50 (end of exhale) was registered to the phase 0 (end of inhale), which was chosen as the primary CT set for DIR. Selected patients had varying tumor volume sizes [Table 1].

Table 1: Tumor volume size per patient

Patient #	1	2	3	4	5	6	7	8	9	10
Tumor volume (cm^3)	1.7	11.0	2.5	108.9	14.1	29.9	21.7	96.6	23.5	17.8

Deformable image registration

Several registration options are available in Velocity AI (Varian Medical Systems, Palo Alto, CA): DICOM, rigid, rigid + scale, deformable, deformable multi pass (DMP), rigid + DMP, extended DMP (EXDMP) and structure-guided deformable. Velocity's primary registration algorithm uses a multi-resolution approach based on Mattes MI, the transform used is a cubic B-Spline, the interpolator used is a bi-linear interpolation and the optimizer is based on the method of steepest gradient descent. As far as, degrees of freedom of this approach Velocity is using a B-Spline of order 3 (cubic) with a uniform knot vector. The number of control points (per-dimension) is configurable with a minimum of 5 control points per-axis (no other constraints are imposed onto this value). The registration methods chosen for evaluation were rigid (RIGID – translation and rotation in the x, y, and z planes), DMP (a three pass coarse to medium to fine resolution deformable that yields finer touch up) and EXDMP (a six pass deformable that goes into finer resolution than the DMP). A workflow chart for the Rigid, DMP, and EXDMP registration algorithms is shown in Figure 1. DMP performs DIR sequentially from low to high resolution, i.e., after registration has been completed in one resolution stage, results are used as initial conditions for the next stage. The resolution used in each stage is determined automatically. The multi-resolution approach increases the number of control points used by the B-Spline transform between successive resolution levels. The software manufacturer suggests the use of the DMP method for CT to CT registration and the EXDMP when DMP fails to provide satisfactory results.

Evaluated similarity/dissimilarity measures

For DIR evaluation, registration between the 0 and 50 phases was used for comparison and DIR accuracy assessment, where the 0 phase dataset represented the reference image set. In this study, 23 measures were evaluated using two groups of 3D datasets: The complete CT dataset and the cropped CT dataset (3D CT dataset cropped to the tumor volume region). These measures were: The cross correlation (CC), normalized CC (NCC), distance correlation (DC), root mean squared error (RMSE), normalized absolute error, mean norm of the difference (MND), structural similarity index (SSIM), feature similarity index (FSIM), dimensionless global error (ERGAS), gradient magnitude similarity deviation (GMSD), quality index (Q), Dice similarity coefficient (DSC), Tanimoto coefficient (TC), bias (B), Bray-Curtis dissimilarity (BCD), Pearson correlation coefficient (PCC), Spearman rank correlation coefficient (SRCC), Euclidean distance (ED), Morisita-Horn

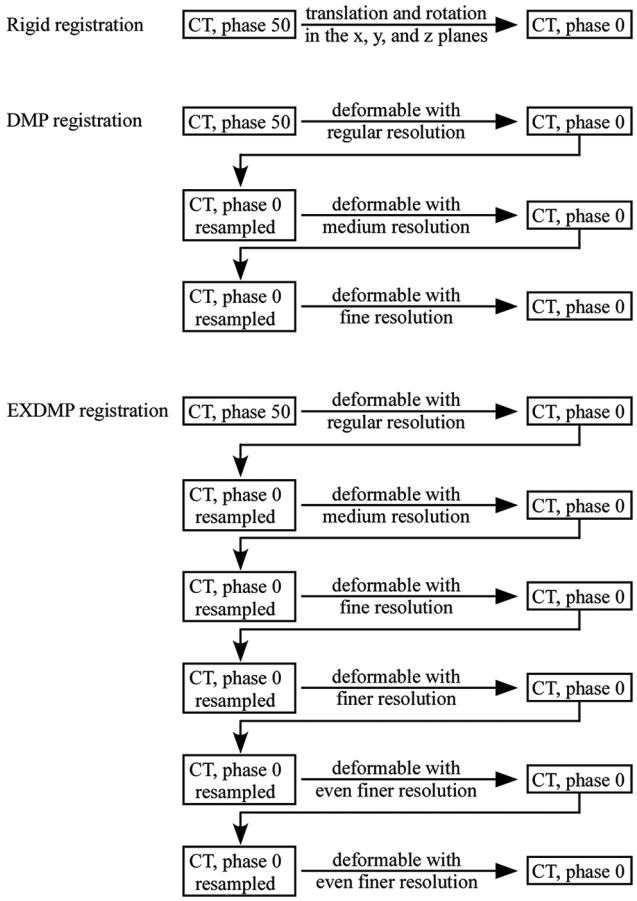

Figure 1: A workflow chart for the rigid, deformable multi pass and extended deformable multi pass registration algorithms

dissimilarity (MHD), Sorensen dissimilarity (SD), simple matching dissimilarity (SMD), structural content (SC), and the 2D voxel mapping (MI is used in the Velocity AI algorithm, so it was not used in the assessment of the algorithm performance).[3,6,12-16] A short description of those measures is provided in Appendix 1.

Nonphysical behavior evaluation (deformation and strain)

For each of the studied cases and for each of the two examined DIR methods (EXDMP and DMP), the binary deformation fields were exported from the Velocity AI (Varian Medical Systems, Palo Alto, CA) and the data were used in MATLAB for the deformation field assessment.

First, the deformation gradient, F, was calculated. The calculated deformation gradient F can be decomposed into the product of a proper orthogonal tensor®, describing the rigid body displacements, and a symmetric tensor (U), describing the stretch deformation:[19]

$$F = RU \text{ or } F = VR \quad (1)$$

where U is the right stretch tensor and V is the left stretch tensor. Based on these two stretch tensors, two commonly used deformation tensors are defined, the right Cauchy-Green tensor C (= U^2), and the left Cauchy-Green tensor B (= V^2). Both deformation tensors can be obtained from the deformation gradient:

$$F^TF = (RU)^T(RU) = U^TR^TRU = U^TU = UU = U^2 = C \quad (2)$$

Using the same approach it can be easily verified that

$$FF^T = V^2 = B \quad (3)$$

The tensor that was used in the calculations is the Eulerian strain tensor which is defined as follows:

$$e^* = \frac{1}{2}(I - B^{-1}) \quad (4)$$

where I is the identity matrix. It can be noted that if there is no deformation $B^{-1} = I$ and $e^* = 0$. A change in the volume due to deformation can be calculated using the Jacobian determinant, and it is defined as:[20]

$$dV = JdV_o \quad (5)$$

For incompressible material $dV = dV_0$ and $J = 1$. In the above expression the Jacobian determinant of the deformation is defined as the determinant of the deformation gradient:[21]

$$J = detF \quad (8)$$

Results

Deformable image registration

There is a considerable computational time difference between the DMP and the EXDMP methods. For all the studied cases, it took 3–4 times longer per image set to complete the EXDMP registration compared to the DMP registration for the same datasets. Not all the measures were evaluated using the complete CT datasets mainly due to the computational memory limitations and the large number of data being evaluated. For example, voxel mapping was applied only on the 3D datasets which were cropped to the wider volume of the tumor location. The same goes for several other coefficients among which were the Dice and Tanimoto. Since this study does not perform an inter-comparison or evaluation of the different measures, this issue has no impact on the analysis. What is important in this analysis is the use of the same CT image volume (complete or cropped) for all three image registration algorithms per measure.

Rigid registration visibly produced the worst registration results, which were validated with the overall outcome of the implemented measures. For the complete CT datasets and taking into the account all the evaluated measures, it was found that the RIGID registration was the worst in 75% of the cases, and for the cropped volume data in 96% of the cases.

Evaluated measures

The measures that consistently outlined the RIGID registration as the least accurate (in both datasets) were the Q and DC (CC, RMSE, and GMSD produced same result in 90% of the cases). From the measures that were taken using the limited volume datasets, the RIGID registration was outlined as the worst registration in 100% of the cases for these measures: CC,

Table 2: Results of the distance correlation and universal quality index metrics

Measure	Complete dataset				Cropped dataset			
	Measure outlined best registration method	In percentage of cases	Measure outlined worst registration method	In percentage of cases	Measure outlined best registration method	In percentage of cases	Measure outlined worst registration method	In percentage of cases
DC	EXDMP	50	EXDMP	0	EXDMP	70	EXDMP	0
	DMP	50	DMP	0	DMP	30	DMP	0
	RIGID	0	RIGID	100	RIGID	0	RIGID	100
Q	EXDMP	80	EXDMP	0	EXDMP	80	EXDMP	0
	DMP	20	DMP	0	DMP	20	DMP	0
	RIGID	0	RIGID	100	RIGID	0	RIGID	100

EXDMP: Extended deformable multi pass, DMP: Deformable multi pass, DC: Distance correlation, Q: Quality index

RMSE, MND, GMSD, DC, PCC, SRCD, BCD, ED, ERGAS, Q, MHD, and 2D voxel mapping. The breakdown of some of the results is shown in Table 2, where the DC and Q measures are used as an example.

According to our results, MI did not perform as well on the cropped dataset as it did on the complete dataset, which to some extent is contradictory to the results of some other measures (such as the MND, FSIM, ERGAS, and B), which would give better results on the cropped dataset. Furthermore, the results acquired based on the cropped dataset indicate the EXDMP registration as the favorable one in more cases compared to the results of the complete datasets for the same measures. Unfortunately, due to the computational memory limitations, some of the measures that performed well in the cropped datasets evaluation were not assessed for the complete datasets. These measures are: The BCD, PCC, SRCD, ED, MHD, SD, SMD, and the 2D voxel mapping.

The results of the 2D voxel mapping with the corresponding coefficients of determination could be independently validated using the Velocity AI 2D voxel map response option. The cropped comparison volumes were larger in the calculation that was done in MATLAB, using rectangular regions of interest (more voxels included in comparison), than the volumes used for calculation in Velocity AI. As shown in Table 3, the data indicate that the value of R2 (coefficient of determination) increases with the volume involved in mapping (larger number of voxel points).

No correlation was found between the tumor volume size and any of the measures regardless of the size of the 3D volume being evaluated. The differences between the measures' values are shown in Figure 2 for the three most common measures, namely the CC, MI and TC. Patients #2 and #6 showed the largest difference in measure values, when compared to the rest of the patients, regardless of the registration method assessed. These two patients have tumor volumes located in the lower lung and posteriorly.[5,22,23] Figure 2 illustrates the variability of behavior of the similarity metrics per patient.

When the results are broken down per patient [Table 4], it is seen that when all the calculated measures were taken into account for cropped volumes, the DMP and EXDMP methods

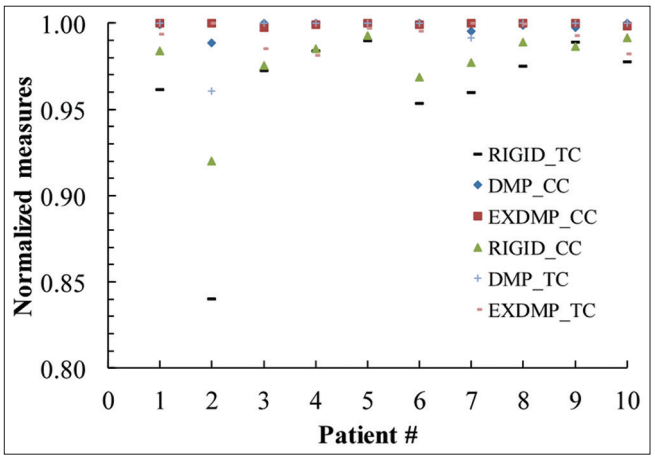

Figure 2: Comparison of the resulting values of the three most commonly used measures in the deformable image registration accuracy assessment (the data were obtained using the complete three-dimensional dataset)

share the occurrences in the best value column equally (50:50). Even when we select only the measures, which indicate the RIGID transformation is the worst (least accurate) method, the ranking stays the same [Table 5]. Overall, across all the measures and evaluated volume datasets, DMP was ranked as best in 61% of the cases, EXDMP in 34% of the cases, and RIGID in 5% of the cases for the complete dataset.

Since the results of the registration method accuracy varied so widely across the studied measures, the sensitivity analysis was performed for the most prominent measures as suggested in the study by Yaegashi et al.[7] The similarity measures of the 4DCT images were evaluated with respect to the 50 phase CT dataset. The image similarity with respect to this phase decreases as the respiratory phase increases. To find which measure is the most sensitive we looked at the rate of change of each measure. The dissimilarity measures used were converted to similarity measures and the error and distance measures were normalized to produce compatible comparisons. Yaegashi *et al*. looked at the image per image correspondence (calculating the degree of similarity between two images), while in our study the complete 3D volume was used for assessment, which may explain some of the differences found between the two studies. The

Table 3: Two dimensional voxel mapping R^2 values comparison per patient (best value in bold)

Patient #	Method	Cropped dataset	Velocity AI tumor volume + 0.5 cm	Velocity AI tumor volume + 5.0 cm
1	DMP	**0.96**	0.50	0.91
	EXDMP	**0.96**	**0.74**	**0.92**
	RIGID	0.91	0.10	0.75
2	DMP	0.90	0.45	0.82
	EXDMP	**0.93**	**0.61**	**0.88**
	RIGID	0.71	0.16	0.60
3	DMP	**0.95**	0.71	0.91
	EXDMP	**0.95**	**0.75**	**0.92**
	RIGID	0.89	0.42	0.81
4	DMP	**0.96**	**0.86**	**0.93**
	EXDMP	**0.96**	**0.86**	**0.93**
	RIGID	0.93	0.68	0.87
5	DMP	**0.93**	0.64	0.90
	EXDMP	**0.93**	**0.80**	**0.91**
	RIGID	0.91	0.38	0.85
6	DMP	**0.95**	0.53	**0.79**
	EXDMP	**0.95**	**0.60**	0.78
	RIGID	0.88	0.23	0.63
7	DMP	0.94	0.75	0.88
	EXDMP	**0.95**	**0.81**	**0.92**
	RIGID	0.87	0.15	0.76
8	DMP	**0.95**	**0.78**	**0.91**
	EXDMP	**0.95**	0.77	**0.91**
	RIGID	0.92	0.62	0.83
9	DMP	0.93	0.16	0.84
	EXDMP	**0.94**	**0.65**	**0.90**
	RIGID	0.90	0.10	0.79
10	DMP	**0.98**	**0.80**	**0.93**
	EXDMP	0.97	0.76	0.91
	RIGID	0.93	0.29	0.80

EXDMP: Extended deformable multi pass, DMP: Deformable multi pass, AI: Artificial intelligence

Table 4: List of the method preferences using all the studied measures for the cropped three-dimensional computed tomography dataset

	Percentage occurrence, all measures	
Patient #	DMP	EXDMP
1	42	**58**
2	0	**100**
3	**95**	5
4	**95**	5
5	**63**	37
6	**89**	11
7	0	**100**
8	37	**63**
9	26	**63**
10	**95**	5

The largest values per patient are shown in bold. EXDMP: Extended deformable multi pass, DMP: Deformable multi pass

aforementioned study suggested the MI measure as the most sensitive one, whereas according to our study other measures appear to be more sensitive, such as the SSIM, ED, MND, RMSE, and TC [Figure 3]. Among the similarity measures, RMSE, MND, and GMSD indicated the RIGID method as the least accurate one in 100% of the cases, while the TC and FSIM gave the same result in 90% of the cases.

Figure 3 also illustrates the measures that have comparable image similarity sensitivities, namely DC, CC, and MHD with their values almost constant at about 1.0. The 2D voxel mapping method accuracy results, which were obtained through the corresponding coefficient of determination, matched exactly the CC, Pearson correlation dissimilarity, and the Morisita-Horn dissimilarity results from the cropped CT dataset. The Tanimoto coefficient method accuracy results matched the results obtained by simple matching dissimilarity, Sorensen dissimilarity, and gradient magnitude similarity deviation. Patient #1 is a patient with the smallest tumor volume and is also the one that consistently showed rigid transformation as the best one for the number of used measures when the complete dataset was evaluated.

Quantifying deformation and strain

For all the cases, and the two evaluated methods, Eulerian strain tensors were calculated and their maximum values were compared with published data.[24,25] Since the volume data that were used for calculation consisted of various tissues with

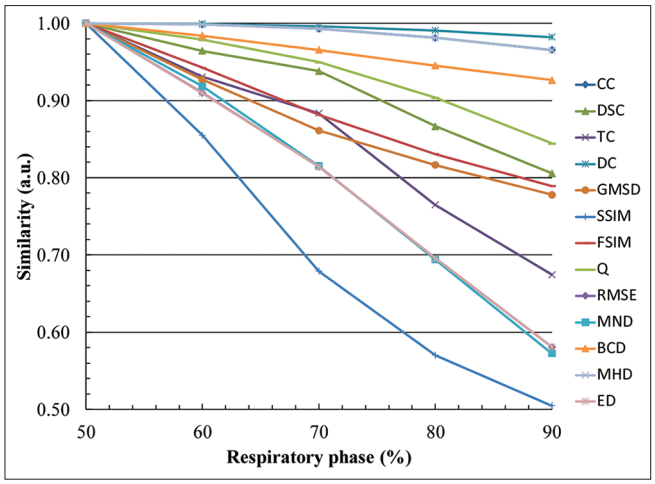

Figure 3: Comparison of the measures for each respiratory phase. The measures were applied on the cropped dataset

different mechanical properties and biochemical data, the results covered a wide range of values [Table 6].

The strain tensor comparison shows consistently larger values for the EXDMP method implying larger mechanical deformations as indicated by the Eulerian strain tensor, which is also confirmed by the minimum Jacobian determinant values [Table 7]. The only exception in the above pattern is patient #4, who is the only patient that exhibits plausible physical behavior when EXDMP method is used ($J > 0$).

Local tissue expansion corresponds to a Jacobian determinant >1 and local tissue contraction corresponds to a Jacobian <1. The results for the DMP method based on the J minimum values in Table 7 indicate that in all the studied cases a certain amount of tissue contraction is observed. The difference in behavior between the two studied methods is even more visible in the Jacobian determinant color map for one of the central transverse slices of the cropped CT dataset for the two different patient cases [Figure 4].

Discussion

The results of DMP showing the best performance in many cases were unexpected due to the fact that the EXDMP method has a longer computational time. By further analyzing the obtained results from the cropped datasets, it could be seen that in the cases where for evaluated measures EXDMP is predominantly best (and RIGID constantly worst) DMP was always better for the patients 3, 4, or 6. Interestingly, only these three patients from all the evaluated cases had tumors located in the upper lung and posteriorly, and this was independent of the tumor volume size as these three cases have widely different tumor sizes [Table 1].

The measures that always scored RIGID registration as the worst one were CC, RMSE, MND, GMSD, 2D voxel map, DC, PCC, SRCC, BCD, MHD, ERGAS, and Q for the cropped CT dataset and DC and Q for the complete CT dataset. The sensitivity study showed that RMSE, MND, ED, GMSD, TC, and FSIM measures have the highest image similarity sensitivity and at the same time found RIGID registration to be the least accurate one

Table 5: List of the method preferences using only the measures where the RIGID method was found to be the least accurate one, using the cropped three-dimensional computed tomography dataset

	Percentage occurrence, all measures	
Patient #	**DMP**	**EXDMP**
1	35	**65**
2	0	**100**
3	**94**	6
4	**94**	6
5	**59**	41
6	**88**	12
7	0	**100**
8	29	**71**
9	29	**71**
10	**100**	0

The largest values per patient are shown in bold. EXDMP: Extended deformable multi pass, DMP: Deformable multi pass

Table 6: Mechanical properties of different tissues as assessed from the deformation data

Patient #	**Eulerian strain tensor**	
	EXDMP	**DMP**
1	0.70	0.39
2	0.88	0.31
3	0.53	0.11
4	0.55	0.10
5	0.50	0.05
6	1.33	0.30
7	1.05	0.29
8	0.87	0.17
9	0.82	0.16
10	0.78	0.12

EXDMP: Extended deformable multi pass, DMP: Deformable multi pass

Table 7: Jacobian determinant scalar values used for the evaluation of the nonphysical deformable image registration behavior

Patient #	**Minimum Jacobian determinant**	
	EXDMP	**DMP**
1	−0.41	0.21
2	−0.56	0.56
3	−0.18	0.61
4	0.00	0.77
5	−0.25	0.74
6	−0.37	0.42
7	−1.19	0.54
8	−0.32	0.36
9	−0.81	0.57
10	−0.63	0.63

EXDMP: Extended deformable multi pass, DMP: Deformable multi pass

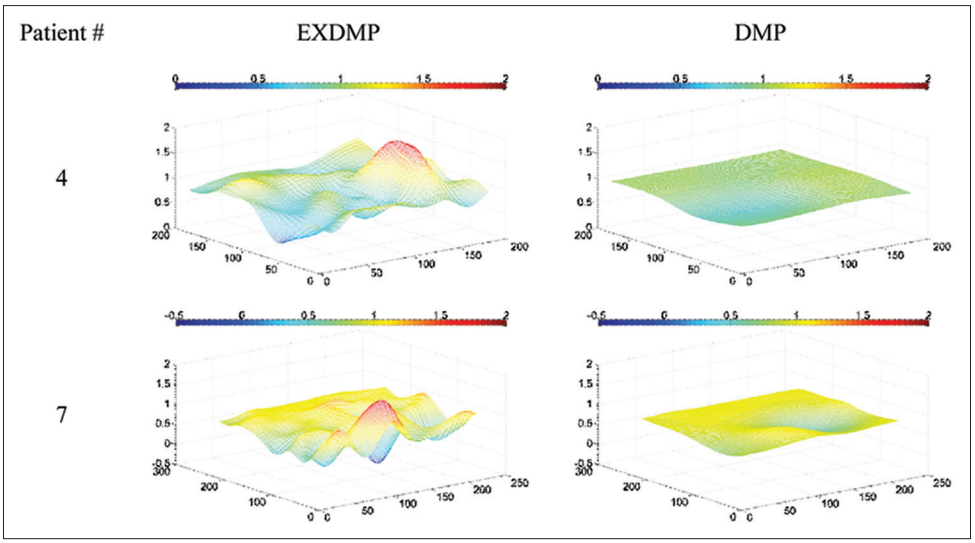

Figure 4: Jacobian determinant map emphasizing the transformation difference between the extended deformable multi pass and deformable multi pass method

in more than 90% of the studied cases. This suggests that these measures can be used for DIR accuracy evaluation.

Based on both the Jacobian and strain tensor calculations, it can be noted that while one DIR method may be helpful for the task of contour propagation it can be at the same time problematic when used for the task of dose accumulation and/or for the task of measuring a local volume change. The strain tensor values for the DMP method are well associated with published data, which report that the ultimate tensile strain for different tissues, such as tendon, ligament, skin, and aorta varies from 0.1 to 1.2 in one of the studies[24] and from 0.14 to 0.18 for ligament and tendons in another one.[25] The Jacobian determinant also indicated nonphysical behavior from EXDMP. Together, the Jacobian determinant and strain measures can give valuable information of the DIR's physical behavior. The development of all the inclusive measures for the evaluation of a DIR algorithm, which will take into account all the aforementioned issues (accuracy, quality, similarity, sensitivity, and plausibility) will be addressed in future work.

At present, based on the results of the nonphysical behavior analysis and the results of similarity measure analysis, only the GMSD, SD, SMD, and TC indicated the DMP method as the best one in 80% of studied cases for the cropped dataset, and only one measure-B, indicated the same for the complete dataset, without picking RIGID as one of the best methods. Finally, the DMP method was shown to be better than the EXDMP when it comes to the physicality of the deformable registration and to the correct assessment of the volume change and mechanical stress in the deformation process.

Although the present study presents some interesting findings, it is also subject to a number of limitations. Task Group 132 (TG-132)[26] presents techniques and workflows for image registration as well as a few common evaluation measures. In the present study, the large majority of registration evaluation measures reported in the literature have been on the same clinical dataset to evaluate the performance of three different registration algorithms given the fact that there has not been yet any measure established as reference or 'golden' standard. However, although the analysis provides a quantitative mean of evaluating the quality of registration and the measures used have been validated by other studies, none of the registrations were assessed by a radiation oncologist or a radiologist. Most studies that evaluate image registration algorithms employ only few evaluation measures and their conclusions are subject to their results. However, as it is shown here, there is a considerable variability in the results of the different evaluation measures even when evaluating exactly the same dataset and registration algorithms. On the other hand, intra- and inter-observer variability (in manual image registration or contour delineation) has been shown to be larger than that of the evaluation measures. Hence, the assessment of the performed registrations by a single physician could not be adequate for our purpose. Finally, 10 patients is a small cohort. Hence, the conclusions derived by the presented findings should be considered with caution in the light of the reduced statistical power of the analysis.

Conclusion

In this study, we have demonstrated the performance of various measures that may be used for the evaluation of rigid and deformable registration accuracy of a 4DCT dataset. The EXDMP method showed an overwhelming nonphysical and unrealistic behavior as well as poor image similarity in a number of studied cases, making DMP the method of choice. However, care must be taken when deciding which method should be used, because this also depends on the task for which it is applied, i.e., dose accumulation, contour propagation, or measuring local volume (or surface area) change. For example, better voxel mapping (EXDMP) may lead to more accurate

contour propagation, etc. It was also shown that the evaluated measures might detect CT dataset differences with higher precision if the analysis is restricted to a smaller volume (i.e., differences were observed in the results of the measures depending on the size of the CT dataset being evaluated).

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Foskey M, Davis B, Goyal L, Chang S, Chaney E, Strehl N, *et al*. Large deformation three-dimensional image registration in image-guided radiation therapy. Phys Med Biol 2005;50:5869-92.
2. Schaly B, Kempe JA, Bauman GS, Battista JJ, Van Dyk J. Tracking the dose distribution in radiation therapy by accounting for variable anatomy. Phys Med Biol 2004;49:791-805.
3. Kashani R, Hub M, Balter JM, Kessler ML, Dong L, Zhang L, *et al*. Objective assessment of deformable image registration in radiotherapy: A multi-institution study. Med Phys 2008;35:5944-53.
4. Brock KK, Deformable Registration Accuracy Consortium. Results of a multi-institution deformable registration accuracy study (MIDRAS). Int J Radiat Oncol Biol Phys 2010;76:583-96.
5. Kadoya N, Fujita Y, Katsuta Y, Dobashi S, Takeda K, Kishi K, *et al*. Evaluation of various deformable image registration algorithms for thoracic images. J Radiat Res 2014;55:175-82.
6. Lu Y, Sun Y, Liao R, Ong SH. A New Similarity Measure for Deformable Image Registration Based on Intensity Matching. 2013 IEEE 10th International Symposium on Biomedical Imaging (ISBI); 2013. p. 234-7.
7. Yaegashi Y, Tateoka K, Fujimoto K, Nakazawa T, Nakata A, Saito Y, *et al*. Assessment of Similarity Measures for Accurate Deformable Image Registration. J Nucl Med Rad Ther 2012;3:137.
8. Heinrich MP, Jenkinson M, Bhushan M, Matin T, Gleeson FV, Brady SM, *et al*. Non-local shape descriptor: A new similarity metric for deformable multi-modal registration. Med Image Comput Comput Assist Interv 2011;14:541-8.
9. Fatyga M, Dogan N, Weiss E, Sleeman, IV WC, Zhang B, Lehman WJ, *et al*. A voxel-by-voxel comparison of deformable vector fields obtained by three deformable image registration algorithms applied to 4DCT lung studies. Front Oncol 2015;5:17.
10. Maurel W, Thalmann D, Wu Y, Thalmann NM. Biomechanical Models for Soft Tissue Simulation. Berlin, Heidelberg: Springer Berlin Heidelberg; 1998.
11. Cowin SC, Doty SB. Tissue Mechanics. New York: Springer Science & Business Media; 2007.
12. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: From error visibility to structural similarity. IEEE Trans Image Proc 2004;13:600-12.
13. Zhang L, Zhang D, Mou X, Zhang D. FSIM: A feature similarity index for image quality assessment. IEEE Trans Image Proc 2011;20:2378-86.
14. Renza D, Martinez E, Arquero A. A new approach to change detection in multispectral images by means of ERGAS index. IEEE Geosci Remote Sensing Letters 2013;10:76-80.
15. Xue W, Zhang L, Mou X, Bovik AC. Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index. IEEE Transactions on Image Processing 2014;23:684-95.
16. Wang Z, Bovik AC. A universal image quality index. IEEE Signal Proc Letters 2002;9:81-4.
17. Veress AI, Phatak N, Weiss JA. Deformable Image Registration with Hyperelastic Warping. In: Suri JS, Wilson DL, Laxminarayan S, editors. Handbook of Biomedical Image Analysis, New York: Springer US; 2005. p. 487-533.
18. Yanovsky I, Thompson PM, Klunder AD, Toga AW, Leow AD. Local volume change maps in nonrigid registration: When are computed changes real. International Conference on Medical Image Computing and Computer Assisted Intervention, Workshop on Statistical Registration; 2007. p. 1-8.
19. Rubin D, Krempl E, Lai WM. Introduction to Continuum Mechanics. Newnes; 2012.
20. Weiss JA, Veress AI, Gullberg GT, Phatak NS, Sun Q, Parker D, *et al*. Strain Measurement Using Deformable Image Registration. In: Holzapfel PGA, Ogden PRW, editors. Mechanics of Biological Tissue, Berlin Heidelberg: Springer -Verlag; 2006, p. 489–501.
21. Bower AF. Applied Mechanics of Solids. Boca Raton: CRC Press; 2009.
22. Paganetti H, Jiang H, Adams JA, Chen GT, Rietzel E. Monte Carlo simulations with time-dependent geometries to investigate effects of organ motion with high temporal resolution. Int J Radiat Oncol Biol Phys 2004;60:942-50.
23. Chan MK, Kwong DL, Ng SC, Tam EK, Tong AS. Investigation of four-dimensional (4D) Monte Carlo dose calculation in real-time tumor tracking stereotatic body radiotherapy for lung cancers. Med Phys 2012;39:5479-87.
24. Holzapfel GA. Biomechanics of soft tissue. Handbook Materials Behav Models 2001;3:1049-63.
25. Weiss JA, Gardiner JC. Computational modeling of ligament mechanics. Crit Rev Biomed Eng 2001;29:303-71.
26. 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:e43-76.

Appendix

Appendix 1

Cross Correlation and Normalized Cross Correlation

Cross correlation (CC) is a standard approach in feature detection. As such it can be used as a similarity measure to calculate the degree of similarity between two images. This metric computes pixel-wise CC. This measure is good for evaluation of alignment in within a single imaging modality. Then it can detect subtle changes in image intensity and/or shape of a structure. This is why it is a good choice for comparison of the computed tomography (CT) datasets of the same subject. In image-processing applications in which the brightness of the image and template can vary due to lighting and exposure conditions, the images can be first normalized. This is typically done at every step by subtracting the mean and dividing by the standard deviation, which leads to CC normalization. Normalized CC (NCC) is often used as a similarity measure for the comparison of the accuracy of different DIR algorithms.

Distance Correlation

Distance correlation (DC) is used in statistics and probability theory and it is a measure of the statistical dependence between two random variables. This measure gets a value of zero only if the two compared variables are statistically independent. The measure is derived from other quantities: Distance variance, distance covariance and distance standard deviation. It was introduced to address a deficiency of Pearson's correlation, which prevented it from becoming zero in the cases of dependent variables. The distance correlation is expressed as follows:

$$dCor(P,R)=\frac{dCov(P,R)}{\sqrt{dVar(P)dVar(R)}} \quad \text{(A1)}$$

where $dCov$ is the distance covariance and $dVar$ is the distance variance.

Root Mean Squared Error

Error metrics are usually used to measure the quality of reconstructed images compared to the original ones. As the value that is produced by this metric decreases, the image resemblance improves (i.e., higher similarity). If a pixel in the original image is denoted as P_i and in the reconstructed image as R_i, the mean square error (MSE) between the two images is defined as:

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(P_i-R_i)^2 \quad \text{(A2)}$$

The root mean squared error is defined as the square root of the MSE.

Normalized Absolute Error

The normalized absolute error is one of the objective image quality measures, which is in line with the MSE, root mean squared error (RMSE), MAE, LMSE, SC measures and it is expressed as follows:

$$NAE=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}|O\{P(i,j)\}-O\{R(i,j)\}|}{\sum_{i=1}^{m}\sum_{j=1}^{n}|O\{P(i,j)\}|} \quad \text{(A3)}$$

where $O\{\}$ is an operator and P and R are the original and transformed (degraded) images, respectively. The large value of this measure means that image quality is poor. This measure is mainly used in the quality assessment of compressed images.

Mean Norm of the Difference

The mean norm of the difference (MND) is a distance metric, which assumes that intensities are only subject to zero mean Gaussian noise and it is calculated as the sum of root squared differences. It is one of the simpler measures as it looks at the intensity relationship in the context of intra-modality registration. Intensity values are assumed to be identical across images or related by an affine transformation (linear relationship). Mean norm of the difference is then the mean of the calculated value.

Structural Similarity Index

The structural similarity index (SSIM) measures image quality based on an initial uncompressed and/or distortion free image as reference.[12] SSIM was designed to improve the traditional image quality measures like MSE and peak signal to noise ratio (PSNR). SSIM is calculated on various windows of an image. The measure between windows of two images (*P, R*) of common size is:

$$SSIM(P,R)=\frac{(2\mu_P\mu_R+c_1)(2\sigma_{PR}+c_2)}{(\mu_P^2+\mu_R^2+c_2)(\sigma_P^2+\sigma_R^2+c_2)} \quad \text{(A4)}$$

where μ_P is the average of P, μ_R is the average of R, σ_P^2 is the variance of P, σ_R^2 is the variance of R, σ_{PR} is the covariance of P and R, c_1 and c_2 are the two variables used to stabilize the division and they are defined by the dynamic range of the pixel values. The resultant index is a decimal value between -1 and 1, where value 1 is only reachable when the two datasets are identical.

Feature Similarity Index

The conventional measures such as PSNR and MSE operate directly on the intensity of the image; feature similarity index (SSIM) is motivated by the need to capture the loss of structure in the image. The feature similarity index is based on the fact that human visual system understands an image mainly according to its low-level features.[13] The phase congruency, which is a dimensionless measure of the significance of a local structure, is used as the primary feature in FSIM; the image gradient magnitude is employed as the secondary feature. FSIM computation consists of two stages. In the first stage, the local similarity map is computed. In the second stage, the similarity map is pooled into a single similarity score. Phase congruency and gradient magnitude play complementary roles in characterizing the image local quality. After obtaining the local similarity map, phase congruency is used again as a weighting function to derive a single quality score. FSIM was designed for gray-scale images making it a good similarity measure for CT images comparison.

Dimensionless Global Relative Error of Synthesis

The erreur relative globaleadimensionnelle de synthese (ERGAS), which stands for dimensionless global relative error of synthesis, is mathematically expressed as:

$$ERGAS=\sqrt{\frac{100}{SR}\frac{1}{N}\sum_{k=1}^{N}\left(\frac{RMSE_k}{\mu k}\right)^2} \quad \text{(A5)}$$

where N is the total number of bands, SR is the scale ratio of the spatial resolutions of the *MS* (multi-spectral) and *PAN* (panchromatic) images, and μ_k is the average of the kth band. This index is capable of measuring the global distortion of an image. An ERGAS value equal to zero indicates absence of radiometric distortion, but there is still possibility of spectral distortion. It is mainly used in image fusion applications (fusing multiple input images in multiple output images).

Gradient Magnitude Similarity Deviation

The gradient magnitude similarity deviation is mainly used in the evaluation of the perceptual quality of output images in applications such as image restoration, image compression and multimedia streaming.[15] Image gradients are sensitive to image distortions, while different local structures in a distorted image suffer different degrees of degradations. The pixel-wise gradient magnitude similarity (GMS) between the reference and distorted images combined with the standard deviation of the GMS map can predict accurately perceptual image quality. The resulting Gradient Magnitude Similarity Deviation (GMSD) algorithm is much faster than most of the state-of-the-art image quality assessment methods, and it has competitive prediction accuracy.

Universal Image Quality Index (Q)

Universal image quality index is a global measure defined as:[16]

$$Q=\frac{4\sigma_{PR}\bar{P}\bar{R}}{(\sigma_P^2+\sigma_R^2)((\bar{P})2+(\bar{R})2)} \quad \text{(A6)}$$

where and are the mean values of the original and distorted images respectively, σ_P^2 and σ_R^2 are the variances, and σ_{PR} is the covariance. Q has a range from −1 to 1. If the two images are identical Q equals 1. This index was developed to replace commonly used measures such as MSE, mean absolute error (MAE) and RMSE and it takes into account three different components: Degree of image correlation, luminance distortion and contrast distortion. It is mainly used in assessing image quality in image compression, blurring, locally adaptive resolution coding, etc.

Mutual Information

Mutual information (MI) is among the most popular image similarity measures. MI is an information theory measure of the statistical dependence between the amount of information that one variable contains about another variable (i.e., a measure of how well one image explains the other). In image processing, the most common measure of information is entropy. Entropy is calculated from an image intensity histogram, *H (P)*. In image registration joint entropy is also considered and is calculated using the joint histogram of two images, *H (P, R)*. If two images are totally unrelated their joint entropy would be the sum of the images' individual entropies. As the similarity of the images increases, the joint entropy decreases compared with the sum of the individual entropies. The optimal registration can be gained by maximizing mutual information, *MI (P, R)*.

$$MI(P, R) = H(P) + H® – H(P, R) \quad \text{(A7)}$$

Normalized MI is defined as follows:

$$NMI(P, R) = (H(P) + H®)/H(P, R) \quad \text{(A8)}$$

Normalized MI is more robust for the inter-modality image registration than MI. MI is one of the similarity measures used mostly for the comparison of the DIR accuracy of different algorithms, and is shown to have an advantage over other similarity measures for the evaluation of the accuracy of deformable image registration.[4]

Dice Similarity Coefficient

The dice similarity coefficient (DSC) is often used as a similarity measure in the assessment of the deformable image registration accuracy together with the MI, TC and NCC.[7] DSC is a similarity measure between two images *P* and *R*, and is defined as:

$$DSC = \frac{I|P \cap R|}{|P| + |R|} \quad \text{(A9)}$$

DSC ranges from 0 to 1, where 1 indicates complete correspondence.

Tanimoto Coefficient

The Tanimoto coefficient (TC) is also known as extended Jaccard coefficient, and is commonly used as similarity measure in DIR algorithms accuracy evaluation. A large value of the TC indicates better correspondence between the images. The TC between two images, *P* and *R*, is defined as

$$TC = \left|\frac{P \cap R}{P \cup R}\right| = \frac{P \cdot R}{|P| + |R| - P \cdot R} \quad \text{(A10)}$$

Bias (B)

The bias is often used in image fusion applications (satellite imagery),[9] and it is a statistical analysis measure derived from fused images by comparing the original image with the resultant fused image to see the difference in their spectral quality. It is a measure that is frequently used with the CC, ERGAS and Q measures. This measure calculates the added bias to resultant image by looking at each pixel of the image.

Bray-Curtis Dissimilarity

The Bray–Curtis dissimilarity (BCD) is mainly used in ecology and biology. BCD is a statistical measure that is used to quantify the compositional dissimilarity between two different sites, based on counts at each site. The Bray–Curtis dissimilarity is bound between 0 and 1, where 0 means the two sites have the same composition. BCD is often cited as a distance based measure and it is also used in image quality assessment.

Pearson Correlation Coefficient

The Pearson correlation coefficient measures the strength and the direction of the linear relationship between two variables. The value is bound between -1 and 1 where 1 is strong positive relation, 0 is no relation and -1 is a strong negative correlation. It is the most commonly used correlation coefficient and works very well when the deviations are linear but not when they are curvilinear.

Spearman Rank Correlation Coefficient

The Spearman's rank-order correlation is the nonparametric version of the Pearson product-moment correlation. Spearman's correlation coefficient measures the strength of association between two ranked variables.

Euclidean Distance

In mathematics, the Euclidean distance (ED) or Euclidean metric is the straight-line distance between two points in Euclidean space. The Euclidean distance between points *P* and R is the length of the line segment connecting them. In image analysis, ED represents the distance between certain pixels within an image. In image processing, it is often used as a qualifying metric in a distance transform.

Morisita-Horn Dissimilarity

Morisita's index of similarity was first proposed to measure similarity between two communities. The Morisita index is most easily interpreted as probability. The Morisita index varies from 0 (no similarity) to about 1.0 (complete similarity). Morisita's index has been recommended as the best overall measure of similarity for ecological use.

Sorensen Dissimilarity

The Sorensen similarity index is a very simple index, similar to the Jaccard's index. It may be represented in terms of dissimilarity (1-index). This coefficient weights matches in species composition between the two samples more heavily than mismatches. If many species are present in a community but not present in a sample, it can be useful to use Sorensen's coefficient rather than Jaccard's. The Sorensen and Jaccard coefficients are very closely correlated. The Morisita-Horn index and the adjusted Jaccard and adjusted Sorensen indices of similarity are recommended for quantitative data because they are not greatly affected by the sample size.

Simple Matching Dissimilarity

This is the simplest coefficient for binary data. It is a statistical index that is used for comparing the similarity and diversity of sample sets. It may be represented in terms of dissimilarity (1-coefficient). This coefficient makes use of negative matches as well as positive matches.

Structural Content

Structural Content is a correlation based measure and measures the similarity between two images. Its structural content is given by the equation:

$$SC = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} (R(i,j))^2}{\sum_{i=1}^{m} \sum_{j=1}^{n} (P(i,j))^2} \quad \text{(A11)}$$

Where *P (i, j)* represents the reference image and *R (i, j)* represents the distorted image.

Two-Dimensional Voxel Mapping

Performance of DIR algorithms can also be assessed by recording the CT values of each voxel in the two 3D dataset (original and registered) and subsequently comparing the values of the voxels that correspond to the same location. A voxel map is then created by comparing (mapping) the CT values in the first 3D dataset to the corresponding CT values in the second 3D dataset and finding the line of best linear fit and calculating the corresponding R^2 value (coefficient of determination). If all the values completely match, they should be lying on a straight line with the R^2 value being equal one. In this study, for each dataset a 3D region of interest that encompasses only the region of the tumor volume was chosen and processed using MATLAB. The outcomes were then compared with the voxel mapping results from the Velocity AI (Varian Medical Systems, Palo Alto, CA) (in Velocity AI, voxel mapping was applied on two different volume sizes, tumor volume plus 0.5 cm margin and the tumor volume plus 5 cm margin).