Validating Dose Uncertainty Estimates Produced by AUTODIRECT: An Automated Program to Evaluate Deformable Image Registration Accuracy

Hojin Kim, PhD[1,2], Josephine Chen, PhD[1], Justin Phillips, PhD[1], Jason Pukala, PhD[3], Sue S. Yom, MD, PhD[1], and Neil Kirby, PhD[4]

Abstract

Deformable image registration is a powerful tool for mapping information, such as radiation therapy dose calculations, from one computed tomography image to another. However, deformable image registration is susceptible to mapping errors. Recently, an automated deformable image registration evaluation of confidence tool was proposed to predict voxel-specific deformable image registration dose mapping errors on a patient-by-patient basis. The purpose of this work is to conduct an extensive analysis of automated deformable image registration evaluation of confidence tool to show its effectiveness in estimating dose mapping errors. The proposed format of automated deformable image registration evaluation of confidence tool utilizes 4 simulated patient deformations (3 B-spline-based deformations and 1 rigid transformation) to predict the uncertainty in a deformable image registration algorithm's performance. This workflow is validated for 2 DIR algorithms (B-spline multipass from Velocity and Plastimatch) with 1 physical and 11 virtual phantoms, which have known ground-truth deformations, and with 3 pairs of real patient lung images, which have several hundred identified landmarks. The true dose mapping error distributions closely followed the Student t distributions predicted by automated deformable image registration evaluation of confidence tool for the validation tests: on average, the automated deformable image registration evaluation of confidence tool–produced confidence levels of 50%, 68%, and 95% contained 48.8%, 66.3%, and 93.8% and 50.1%, 67.6%, and 93.8% of the actual errors from Velocity and Plastimatch, respectively. Despite the sparsity of landmark points, the observed error distribution from the 3 lung patient data sets also followed the expected error distribution. The dose error distributions from automated deformable image registration evaluation of confidence tool also demonstrate good resemblance to the true dose error distributions. Automated deformable image registration evaluation of confidence tool was also found to produce accurate confidence intervals for the dose–volume histograms of the deformed dose.

Keywords

deformable image registration, AUTODIRECT, DIR uncertainty, dose mapping error, Student t distribution

Abbreviations

AUTODIRECT, automated deformable image registration evaluation of confidence tool; ART, adaptive radiation therapy; CT, computed tomography; DIR, deformable image registration; DVH, dose–volume histogram; DVF, deformation vector field; HN, head and neck; ROI, region of interest

Received: October 27, 2016; Revised: February 27, 2017; Accepted: March 22, 2017.

[1] Department of Radiation Oncology, University of California, San Francisco, CA, USA
[2] Department of Radiation Oncology, Asan Medical Center, University of Uslan College of Medicine, Seoul, Korea
[3] Department of Radiation Oncology, University of Florida Health Cancer Center at Orlando Health, Orlando, FL, USA
[4] Department of Radiation Oncology, University of Texas Health Science Center San Antonio, San Antonio, TX, USA

Corresponding Author:
Neil Kirby, PhD, Department of Radiation Oncology, University of Texas Health Science Center San Antonio, 7979 Wurzbach Road, Mail Code 7889, San Antonio, TX 78229, USA.
Email: kirbyn@uthscsa.edu

Introduction

Deformable image registration (DIR) has been used for various clinical applications in radiation oncology. The resulting deformation vector field (DVF) from DIR can reduce the time needed for image segmentation[1-7] and can be used to transfer dose from one image to another[1-3,8-11] for adaptive radiation therapy or to evaluate the composite dose from multiple treatment courses. Deformable image registration, however, has inherent uncertainties, which makes the application to dose transfer much more challenging, as there is currently no clinical means to verify its accuracy.

Many studies[12-16] have highlighted the accuracy of DIR algorithms in terms of landmark tracking and image similarity. A few studies[17-21] have investigated how DIR uncertainty affects dose transfer, and some[9,22,23] have developed the automated processes for spatial uncertainty modeling. Bender et al[9] suggested finding the region of spatial DIR uncertainty by assessing the inverse inconsistency of a deformation field, but this ignores DIR errors in the inverse consistent regions. A distance-to-dose difference tool was introduced by Saleh-Sayha et al[22] that maps the effect of dose gradients on the uncertainty. It, however, does not account for other causes of DIR inaccuracy, including the increased errors frequently found in regions of homogeneous image intensity. The work proposed by Murphy et al[23] obtained a spatial DIR error map by analyzing the effect of varying the region of interest (ROI) on the resulting DVF. This, however, does not consider the variability between different DIR algorithms, as well as other causes of DIR uncertainty other than those caused by the ROI.

Our general goal for radiotherapy is to deliver doses to patients that are within 5% of that prescribed, when accounting for all sources of uncertainty (International Commission on Radiation Units, ICRU Report 24). For many patients, DIR becomes a contributing component to this uncertainty, but there is not currently a clinically established method to evaluate this uncertainty for a patient and how it contributes to the overall treatment uncertainty. Recently, Kirby et al[24] proposed a new software tool, the automated DIR evaluation of confidence tool (AUTODIRECT) that evaluates the spatial dose mapping accuracy of a DIR algorithm on a patient-by-patient basis. The automated confidence methodology utilizes a small number of test deformations (currently 4 for computational efficiency) to estimate the uncertainty in the DIR algorithm's performance. It analyzes the spatial dose mapping errors from the 4 test cases by modeling them using a Student t distribution.[25]

This validation is differentiated from the previous proof-of-principle study[24] in 2 ways. First, the implementation of AUTODIRECT in the previous study overpredicted DIR uncertainty. This issue has now been corrected in the current implementation, and it yields results close to the expected statistical behavior. Second, to demonstrate the performance and clinical relevance of the proposed framework, the validation is comprehensively performed with a total of 15 data sets. Twelve of these are phantoms (11 virtual and 1 physical) with known ground-truth deformations. The remaining 3 are actual lung
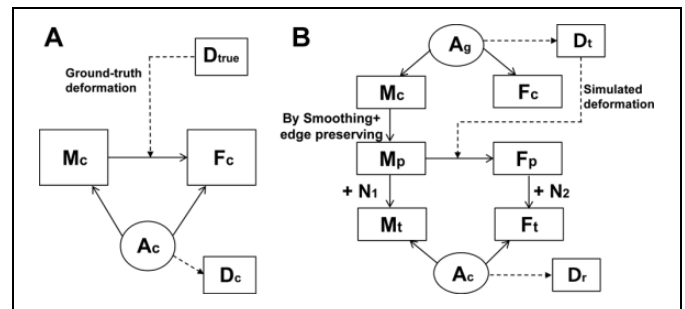
Figure 1. A, Diagram depicting the clinical use of a deformable image registration (DIR) algorithm. B, Diagram depicting the procedure used by automated deformable image registration evaluation of confidence tool (AUTODIRECT) to create a pair of patient-specific images with known deformation to test the performance of a DIR algorithm.

patient data sets containing hundreds of pairs of manually defined landmarks. This extends the testing of AUTODIRECT to several anatomical sites (head-and-neck [HN], pelvis, and lung). Also, as virtual phantoms are simulated deformations, the addition of the physical phantom and real patient data sets benchmarks AUTODIRECT with real deformations.

Methods

Review of AUTODIRECT

For the phantom data sets in this study, the true deformation is known (D_{true}) and can be compared to the DVF (D_c) predicted by the clinical DIR algorithm (A_c) when applied to the phantom images, as seen in Figure 1A.

Figure 1B shows the workflow of AUTODIRECT.[24] It requires fixed and moving CT images (F_c and M_c) and 2 noise scans (N_1 and N_2) describing the noise characteristics of the imaging system to produce DIR error estimates. More specifically, Figure 1B illustrates how AUTODIRECT applies a generator DIR algorithm (A_g) to M_c and F_c to produce a test DVF (D_t). A filter with edge preserving and smoothing functions processes M_c to remove noise that could potentially skew DIR accuracy determination, yielding a processed moving image (M_p). The test deformation field (D_t) is applied to M_p to create an artificial-processed fixed image (F_p). This test deformation is the ground-truth deformation for the pair M_p and F_p. Two unique noise scans are then added to M_p and F_p to create the final test moving and fixed images (M_t and F_t). The pair of M_t and F_t is passed to the clinical DIR algorithm (A_c), and the resulting deformation field (D_r) is compared to D_t. In AUTODIRECT, this procedure is performed 4 times using different generator algorithms (3 B-spline DIRs and 1 rigid transform). The comparison of the 4 test and resulting deformations are used to create a statistical model of the dose errors for each voxel. When small sample sizes (<20) are used to estimate Gaussian distributions, a Student t distribution is better suited to represent the probability distribution of variables. As the current version of AUTODIRECT has small number of samples (n = 4), the distribution of DIR dose mapping error is

assumed to conform to the Student t distribution with 3 (N−1) degrees of freedom. Thus for each voxel, a Student t variable can be defined as expressed in Equation 1.

$$t = \frac{E - \bar{\varepsilon}}{\sqrt{\left(\frac{N+1}{N-1}\right) \times \sigma}} \text{ where } E = d_{\text{actual}} - d_{\text{pred}}, \ \bar{\varepsilon} = \frac{1}{N}\sum\nolimits_{k=1}^{N} \varepsilon_i \quad (1)$$
$$\varepsilon_i = \frac{1}{N}\sum\nolimits_{k=1}^{N}(d_i - p_i), \ \sigma = \frac{1}{N}\sum\nolimits_{k=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2$$

where N is the number of the test deformations ($N = 4$). E is the actual dose mapping error, given by the difference between the true dose at that voxel (d_{actual}) and the dose predicted (d_{pred}) by the clinical (D_c) deformation. Both d_i and p_i are the doses warped by the test ($D_{t,\ i}$) and resulting ($D_{r,\ i}$) deformations, respectively. d_i and p_i represent the known errors produced by A_c when tested using the simulated data. As seen in Equation 1, the mean and standard deviation of the known errors are used to predict the distribution of errors in D_c, which are not known.

Once the t variable is defined for each voxel, Equation 2 shows how to estimate the range of possible dose errors from our proposed framework at a certain confidence interval:

$$-k \leq t = \frac{E - \bar{\varepsilon}}{\sqrt{\left(\frac{N+1}{N-1}\right) \cdot \sigma}} \leq k$$
$$\Rightarrow \left(\begin{array}{c}\text{Confidence Interval} \\ \text{from AUTODIRECT}\end{array}\right)$$
$$= \left(d_{\text{pred}} + \bar{\varepsilon} - k\sqrt{\left(\frac{N+1}{N-1}\right)} \cdot \sigma, \ d_{\text{pred}} + \bar{\varepsilon} - k\sqrt{\left(\frac{N+1}{N-1}\right)} \cdot \sigma\right), \quad (2)$$

where k represents the confidence interval defined by the t distribution, that is, $k = 3.182$ at 95% confidence interval for 3 $(N-1)$ degrees of freedom. Notably, once Equation 2 is derived under assumption of the t statistic, the minimum and maximum doses for the confidence interval are completely defined by the mean and variance of the dose mapping errors from the test cases without the true deformation. Hence, if the t statistic assumption is validated, AUTODIRECT will demonstrate the ability to provide valuable dose uncertainty information for the clinical case without prior knowledge of the true deformation. For the clinical usage of this method, when the ground-truth deformation is not known, the test deformation fields would be utilized to calculate the statistical parameters in Equation 2 to derive the dose confidence interval.

Evaluation of AUTODIRECT

For validation testing of AUTODIRECT, we employed 11 pairs of virtual phantom images based on patient CT images for various body sites (1 prostate, 1 craniospinal, and 9 HN cases). The prostate data set is from a treatment of the prostate and lymph nodes to a dose of 45 Gy in 25 fractions. The craniospinal data set is from a treatment with the prescription

of 39.6 Gy in 22 fractions. The HN cases had several different disease sites: base of tongue, tonsil, nasopharynx, and supraglottic larynx. These cases had prescriptions that were either 70 Gy in 35 fractions or 69.96 Gy in 33 fractions. For each set, a patient image was digitally deformed with ImSimQA, software available from Oncology Systems Limited (OSL), Shrewsbury, Shropshire, UK, to create a target image. It allows the user to manually define the deformation for a small set of control points and then utilizes a thin-plate spline algorithm to define the DVF for the entire volume. The 3 virtual phantoms on 3 different body sites created by Nie et al[21] utilized 30 pairs of anatomical landmarks defined by a physician to simulate anatomical changes in the prostate region due to bladder filling, changes along the cranio-spinal axis due to switching between a prone and supine position, and changes in the HN region due to patient weight loss. The remaining 8 virtual HN cases, created by Pukala et al,[19] simulated anatomy-driven deformations due to mandible translation/rotation as well as tumor and parotid shrinkage. ImSimQA can export the applied deformation, which serves here as the ground-truth deformation for each data set. An additional HN physical phantom set was also used, which was developed by Singhrao et al[17] with a measured ground-truth deformation determined with 891 optical markers. This physical phantom represented deformation from HN flexion. The physical phantom was modeled after an actual patient. The patient's nasopharynx dose distribution (70 Gy in 35 fractions) was fused to the phantom for the dose analysis performed here.

To demonstrate its effectiveness in actual clinical cases, the workflow was also tested with 3 pairs of lung patient images with 300 identified pairs of landmarks, provided by DIR-LAB[14] (http://www.dir-lab.com). These 3 data sets were chosen at random from the 5, high-resolution (512×512 in the axial plane) four-dimensional computed tomography (DCT) data sets available on this site. For each case, the image in inhale breathing cycle was set to be the moving image, while the exhale was set to be the fixed image. Thus, the ground-truth deformations for these cases were the movement of the lungs from exhalation to inhalation. The true mapping between landmarks in the 2 images was provided by the DIR-LAB. As the landmarks were widely distributed throughout the entire lung, we simulated a whole lung treatment for each case. The prescriptions for these cases were 12 Gy in 6 fractions to the patient midplanes, delivered with anterior-posterior (AP)/ posterior-anterior (PA) fields. Because these landmarks are manually identified, their positions are not exact and have an inherent error associated with them. To account for this inherent error, we computed a uniform random variable ranging from −0.5 to 0.5 mm (half of the image resolution in x-, y-directions) for each landmark, multiplied this spatial error by the local dose gradient (Gy/mm), and added this additional dose error to the dose error calculated assuming perfect landmark correspondence.

For this version of AUTODIRECT, the 4 generator algorithms (Ag in Figure 1B) for the simulated deformations consisted of 1 rigid and 3 DIR algorithms. The DIR algorithms
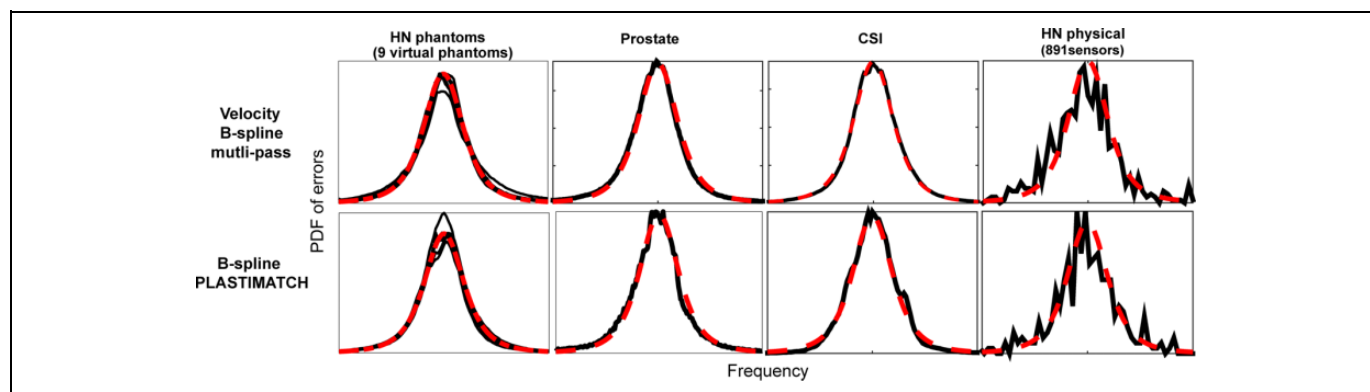
Figure 2. Probability density functions of errors. Comparison of the theoretical *t* distribution (red, dashed) and the distribution of dose mapping errors estimated from automated deformable image registration evaluation of confidence tool (AUTODIRECT) (black, solid) for 12 test data sets when Velocity B-spline multipass and B-spline Plastimatch clinical deformable image registrations (DIRs) are evaluated.

were B-spline-based algorithms[26,27] implemented in a research build of Velocity, with a wide range of control point spacings and passes. Thus, the 4 different generator algorithms produced simulated deformations with varying amounts of smoothness and pliability. For the clinical DIR algorithms (A_c in Figure 1), we applied 2 B-spline DIR algorithms: the B-spline multipass from the commercial version of Velocity (version 2.7) and the B-spline method from an open-source software, Plastimatch[27] (http://www.plastimatch.org) (Version 1.6.1). Velocity utilizes mutual information as its similarity metric, whereas mean-square difference was applied here for the Plastimatch DIR algorithm. Thus, 30 comparisons were made between AUTODIRECT-computed uncertainty maps and the true distribution of DIR errors: one comparison for each combination of 2 different clinical DIR algorithms and 15 different test image pairs (11 virtual phantom, 1 physical phantom, and 3 landmark image data set).

The noise scans (N_1 and N_2 in Figure 1B) were acquired by scanning a water phantom to capture noise and image artifacts, such as ring and nonuniformity artifacts. Two types of noise scans were acquired for this study. The first was from a large pelvic-shaped water phantom for the prostate phantom, and the other was a small cylindrical water phantom for the HN and craniospinal irradiation (CSI) phantom studies.

To assess how well AUTODIRECT modeled the distribution of DIR-related dose errors, the true error and the mean and variance of the test mapping errors for every voxel were calculated by AUTODIRECT for the 30 validation cases. If the AUTODIRECT method is accurate, then the computed *t* statistic should follow the theoretical *t* distribution. In addition, AUTODIRECT was used to estimate 50%, 68%, and 95% confidence intervals for the DIR dose errors using Equation 2. The percentage of voxels within the 3 designated confidence intervals was calculated. The spatial distribution of dose mapping errors was also visualized on the axial CT image and in the form of dose–volume histograms (DVHs) of specific structures.

Results

Phantom Studies

Figure 2 illustrates the distribution of dose mapping errors compared to the theoretical *t* distribution estimated by AUTODIRECT for the B-spline multipass DIR algorithms in Velocity and in Plastimatch. The HN physical phantom yielded a somewhat noisy distribution mainly due to the small number of samples (891 landmarks). The AUTODIRECT-predicted error distributions underestimated the errors for HN phantom 9 with the Velocity B-spline multipass, and slightly overestimated the errors for HN phantom 4 with the B-spline Plastimatch DIR algorithm. For the remaining cases, however, the *t* statistic estimated from AUTODIRECT closely conformed to the theoretical error distribution closely. The AUTODIRECT-derived confidence values were calculated at 3 different designated levels: 50%, 68%, and 95%, corresponding to $k = 0.765$, 1.189, and 3.182 in Equation 2, respectively. Figure 3 plots the percentage of voxels of the deformed dose lying within the 3 confidence intervals for the 12 data sets, which resulted in the average values of 48.8%, 66.3%, and 93.8% and of 50.1%, 67.6%, and 93.8%, for the Velocity and Plastimatch algorithms, respectively.

Figure 4 displays the examples of the spatial distribution of DIR errors for the prostate, CSI, HN phantom1, and HN phantom 9 phantoms. There is a high degree of correlation between the gradient of the dose distributions and DIR uncertainty. Figure 4 also shows the similarity between the actual dose errors (difference between true and clinical deformations) and the predicted uncertainty (mean of the dose mapping errors as defined in Equation 2) by the AUTODIRECT workflow. These examples demonstrate the ability of AUTODIRECT to capture the dose transfer uncertainty without knowing the actual errors.

The resulting uncertainty is employed to estimate the error bars in the form of the DVHs for critical structures
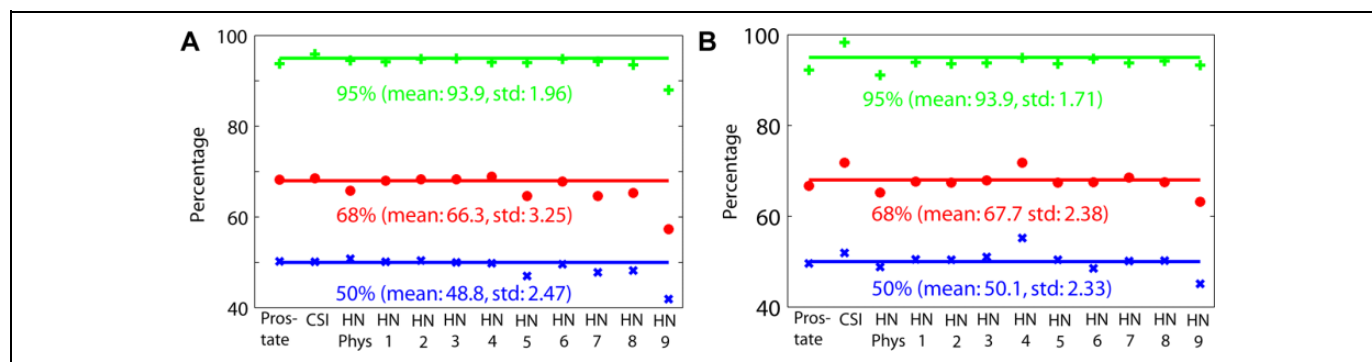
Figure 3. Percentage of voxels within the corresponding automated deformable image registration evaluation of confidence tool (AUTO-DIRECT)-predicted confidence interval for 12 phantom cases when the confidence intervals are defined to be 50%, 68%, and 95% with (A) Velocity B-spline multipass clinical deformable image registration (DIR), (B) B-spline clinical DIR from Plastimatch.
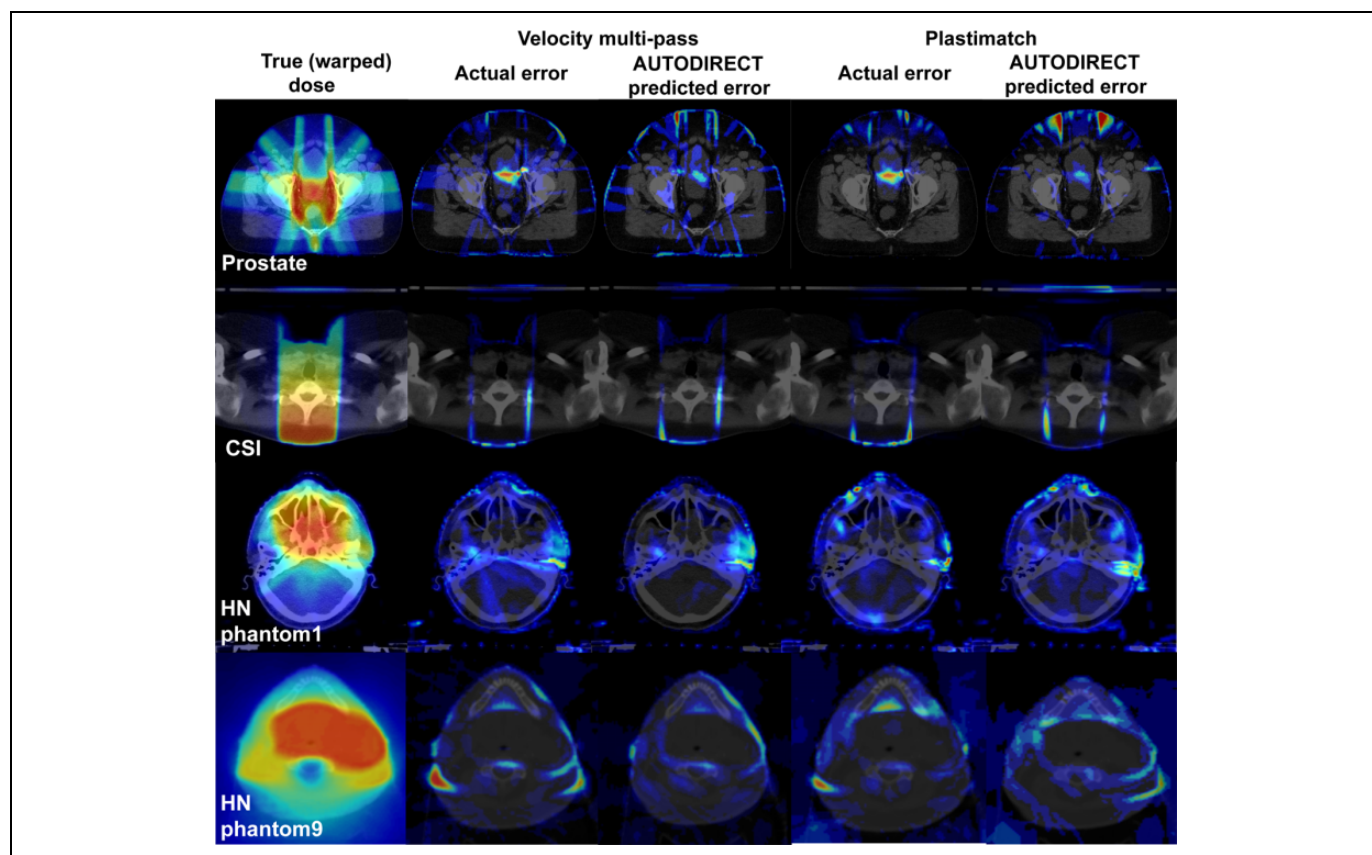

Figure 4. Spatial dose mapping uncertainty: True warped dose (first column), actual error between true warped dose and dose warped by 2 clinical deformable image registrations (DIRs; second and fourth columns), and automated deformable image registration evaluation of confidence tool (AUTODIRECT)-predicted dose (third and fifth columns) for prostate, CSI, HN phantom 1, and HN phantom 9 cases, respectively.

and the target volumes, as seen in Figure 5. The representative DVHs in Figure 5 represent organs at risk that receive relatively high doses in regions of deformations, as well as targets. The predicted error bars (dashed-dotted, red) mostly encompass the true deformed dose (solid, black) for the given structures.

Lung Landmarked Image Study

The AUTODIRECT workflow with Velocity and Plastimatch clinical DIRs produced the resulting error distributions in Figure 6, respectively, for the 3 lung patient data set with 300 landmarks. As stated earlier, the errors include an estimate of the inherent placement error by simulating uniform random
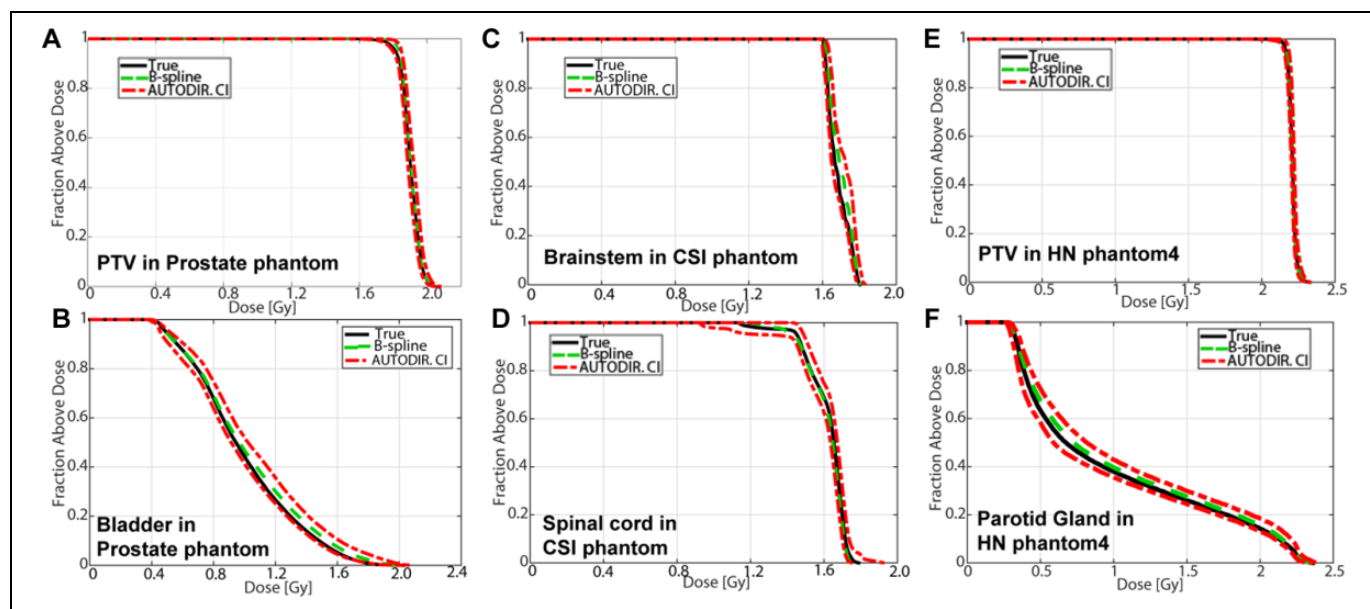
Figure 5. True warped dose (black, solid), dose warped by the clinical deformation (green, dashed) and range of warped dose estimated by automated deformable image registration evaluation of confidence tool (AUTODIRECT) at 95% confidence interval (CI; red, dashed-dotted) for (A, B) prostate and bladder contours in prostate phantom, (C, D) brainstem and spinal cord contours in CSI phantom, and (E, F) planning target volume, (PTV) and parotid gland contours in HN phantom4 data sets.
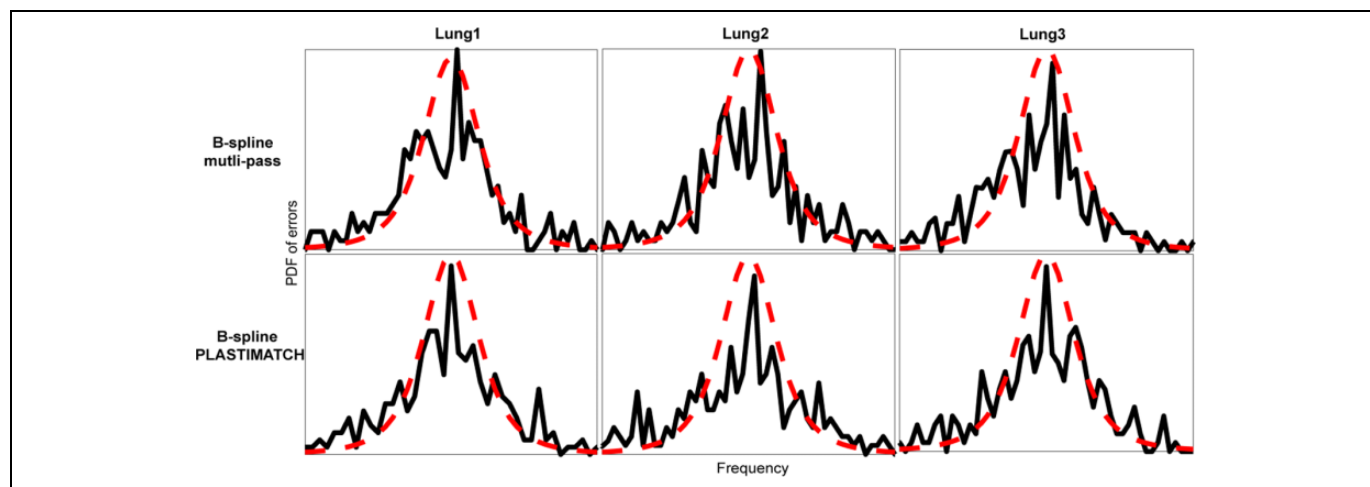
Figure 6. Probability density functions of errors in 3 pairs of real lung images with 300 landmarks from automated deformable image registration evaluation of confidence tool (AUTODIRECT; first row) Velocity B-spline multipass, and (second row) B-spline Plastimatch deformable image registrations (DIRs) were used (dotted in red: *t* distribution, solid in black: predicted by AUTODIRECT).

noise multiplied by the dose gradient. Due to the small number of samples, the error distributions were noisy relative to the results of the virtual phantom study. Importantly, however, the outlines tend to be close to the theoretical *t* distribution for the 3 lung cases with both B-spline clinical DIRs.

Discussion

An AUTODIRECT is an automated software tool that was proposed to predict patient-specific dose mapping accuracy for a clinical DIR. This study verified the accuracy of the dose

mapping uncertainty estimates provided by the workflow using 12 virtual/physical phantoms. The workflow was also tested on 3 clinical lung data sets with 300 landmarks. For the virtual/physical phantoms with known deformation, the true dose error distributions closely followed the *t* distribution predicted by AUTODIRECT. The predictions of the uncertainty for the lung patient data set with 300 landmarks were qualitatively close to the true error distribution. The validation in terms of the proximity of the dose mapping error to the theoretical one is significant, as stated, since the true deformation is not known for prospective patient analysis. The results above demonstrate

that our proposed framework may provide clinically important information that is currently missing from existing platforms. In particular, we believe this information could be valuable to physicians evaluating composite dose distributions for patients undergoing retreatment with radiation.

The 4 test deformations ($D_{t,i}$) play a crucial role in predicting the spatial uncertainty of the DIRs. The initial proof-of-principal work[24] on AUTODIRECT also used 4 generator algorithms: 2 B-spline-based, 1 demons-based,[28] and 1 rigid registration. The specific demons DIR algorithm used in that study was shown to skew DIR uncertainty predictions, resulting in overestimates. Also, the rigid registration was shown to be a reasonable test deformation, yielding predicted errors that were similar to the actual ones. For this reason, the current version adopted 3 B-spline and 1 rigid registration as the test deformations. The B-spline test deformations captured the essence of the ground-truth warping, with deformations in similar locations and of similar magnitude. This test DIR set is shown here to create accurate dose warping uncertainty estimation, as identified in the 12 phantom studies and the 3 patient data sets. It is important to note that this set of test deformations is not universally optimal for all clinical DIR algorithms. Some additional analyses (results not shown) were also performed using a demons-based DIR as the clinical DIR (A_c). In these cases, the predicted error distributions were also overestimated. These dose uncertainty overestimates were as much as 60% for some of the benchmarks in the initial proof-of-principle work. Thus, it may be important to customize the test DIR algorithms employed by AUTODIRECT for the clinical DIR algorithm being used.

It is important to also address the computation burden of this technique. This burden can be divided into 4 different parts: creating the test deformations, creating the test image sets from these deformations, applying the clinical DIR algorithm to the test image sets, and a final analysis to compute the statistical parameters. For the current implementation of AUTODIRECT, these steps take 140, 180, 140, and 80 seconds, respectively (9 minutes total). This computation was timed for a Windows desktop computer with an Intel Xeon E5-1620 v3 processor and 32 GB of random-access memory, (RAM).

Based on the validation performed in this work, the AUTODIRECT workflow shows promise for estimating DIR dose warping uncertainty when transferring dose from one planning CT to another. There are, however, some limitations to the current validation study. The virtual phantoms were created by manually applying reasonable estimates for typical anatomical deformations that occur in patients. In cases, such as weight loss, there is some uncertainty in the true nature of the underlying anatomical deformation. In cases, such as bladder and rectal filling, the DVF is ill-defined in some regions. In these cases, the validation testing results really are only an indication that AUTODIRECT calculations match well with the assumptions made when creating the virtual phantoms. An additional limitation for the virtual phantom validations is that they were produced

with synthetic splines and both the test and clinical DIR algorithms utilized splines for warping. This has the potential to skew the produced DIR error predictions. For this reason, the physical phantom and the lung data sets are crucial to this study. Another limitation of the current AUTODIRECT platform is that the image processing is most appropriate for fan-beam, kilo-voltage CT imaging. Work is ongoing to extend the AUTODIRECT process to imaging typically used during image-guided radiotherapy to enable its use for adaptive therapy applications.

Conclusion

Tested on 12 virtual and physical phantoms and 3 pair of lung images from a real patient, the AUTODIRECT framework was found to reliably estimate the DIR-driven dose mapping errors for 2 widely available B-spline algorithms. Thus, the AUTODIRECT workflow shows promise for providing clinically important information on the uncertainty of a deformed dose distribution.

References

1. Lu W, Olivera GH, Chen Q, et al. A Deformable registration of the planning image (kVCT) and the daily images (MVCT) for adaptive radiation therapy. *Phys Med Biol*. 2006;51(17): 4357-4374.

2. Lu W, Olivera G H, Chen Q, Chen ML, Ruchala KJ. Automatic re-contouring in 4D radiotherapy. *Phys Med Biol*. 2006;51(5): 1077-1099.

3. Zhang T, Chi Y, Meldolesi E, Yan D. Automatic delineation of on-line head-and-neck computed tomography images: toward on-line adaptive radiotherapy. *Int J Radiat Oncol Biol Phys*. 2007;68(2):522-530.

4. Orban de Xivry J, Janssens G, Bosmans G, et al. Tumour delineation and cumulative dose computation in radiotherapy based on deformable registration of respiratory correlated CT images of lung cancer patients. *Radiother Oncol*. 2007;85(2):232-238.

5. Han X, Hoogeman MS, Levendag PC, et al. Atlas-based auto-segmentation of head and neck CT Images: *Proceedings of the 11th International Conference on Medical Image Computing and Computer-Assisted Intervention, Part II*. New York, NY: Springer-Verlag, 6–10 September 2008.

6. Hardcastle N, Tomé WA, Cannon DM, et al. A multi-institution evaluation of deformable image registration algorithms for automatic organ delineation in adaptive head and neck radiotherapy. *Radiat Oncol*. 2012;7(1):90.

7. Hoffmann C, Krause S, Stoiber EM, et al. Accuracy quantification of a deformable image registration tool applied in a clinical setting. *J Appl Clin Med Phys*. 2014;15(1):4564.

8. Davis BC, Foskey M, Rosenman J, Goyal L, Chang S, Joshi S. Automatic segmentation of intra-treatment CT images for adaptive radiation therapy of the prostate. *Med Image Comput Comput Assist Interv*. 2005;8(pt 1):442-450.

9. Bender ET, Hardcastle N, Tomé WA. On the dosimetric effect and reduction of inverse consistency and transitivity errors in deformable image registration for dose accumulation. *Med Phys*. 2012;39(1):272.

10. Akino Y, Yoshioka Y, Fukuda S, et al. Estimation of rectal dose using daily megavoltage cone-beam computed tomography and deformable image registration. *Int J Radiat Oncol Biol Phys*. 2013;87(3):602-608.

11. Veiga C, McClelland J, Moinuddin S, et al. Toward adaptive radiotherapy for head and neck patients: feasibility study on using CT-to-CBCT deformable registration for dose of the day calculations. *Med Phys*. 2014;41(3):031703.

12. Castadot P, Lee JA, Parraga A, Geets X, Macq B, Grégoire V. Comparison of 12 deformable registration strategies in adaptive radiation therapy for the treatment of head and neck tumors. *Radiother Oncol*. 2008;89(1):1-12.

13. Kashani R, Hub M, Balter JM, et al. 2008 Objective assessment of deformable image registration in radiotherapy: a multi-institution study. *Med Phys*. 35(12):5944-5953.

14. Castillo R, Castillo E, Guerra R, et al. A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets. *Phys Med Biol*. 2009;54(7):1849-1870.

15. Brock KK. Results of a multi-institution deformable registration accuracy study (MIDRAS). *Int J Radiat Oncol Biol Phys*. 2010;76(2):583-596.

16. Pukala J, Meeks SL, Staton RJ, Bova FJ, Langen KM. A virtual phantom library for the quantification of deformable image registration uncertainties in patients with cancers of the head and neck. *Med Phys*. 2013;40(11):111703.

17. Singhrao K, Kirby N, Pouliot J. A three-dimensional head-and-neck phantom for validation of multi-modality deformable image registration for adaptive radiotherapy. *Med Phys*. 2014;41(12):121709.

18. Johnson PB, Padgett KR, Chen KL, Dogan N. Evaluation of the tool "Reg Refine" for user-guided deformable image registration. *J Appl Clin Med Phys*. 2016;17(3):158-170.

19. 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.

20. Kirby N, Chuang C, Ueda U, Pouliot J. The need for application-based adaptation of deformable image registration. *Med Phys*. 2013;40(1):011702.

21. 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.

22. Saleh-Sayha NK, Weiss E, Salguero FJ, Siebers JV. A distance to dose difference tool for estimating the required spatial accuracy of a displacement vector field. *Med Phys*. 2011;38(5):2318-2323.

23. Murphy MJ, Salguero FJ, Siebers JV, Staub D, Vaman C. A method to estimate the effect of deformable image registration uncertainties on daily dose mapping. *Med Phys*. 2012;39(2):573-580.

24. Kirby N, Chen J, Kim H, Morin O, Nie K, Pouliot J. An automated deformable image registration evaluation of confidence tool. *Phys Med Biol*. 2016;61(8):N203-N214.

25. Student. The probable error of a mean. *Biometrika*. 1908;6(1):1-25.

26. 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):712-721.

27. Shackleford JA, Kandasamy N, Sharp GC. On developing B-spline registration algorithms for multi-core processors. *Phys Med Biol*. 2010;55(21):6329-6351.

28. Thirion JP.Image matching as a diffusion process: an analogy with Maxwell's demons. *Med Image Anal*. 1998;2(3):243-260.