Technical Note: Intensity-based quality assurance criteria for deformable image registration in image-guided radiotherapy

Lando Bosma, Cornel Zachiu, Baudouin Denis de Senneville, Bas Raaymakers, Mario Ries

Technical Note: Intensity-based quality assurance criteria for deformable image registration in image-guided radiotherapy

L S Bosma[1], C Zachiu[1], B Denis de Senneville[1,2], B W Raaymakers[1], M Ries[3]

[1] Department of Radiotherapy, UMC Utrecht, Heidelberglaan 100, 3508 GA Utrecht, The Netherlands

[2] Institut de Mathématiques de Bordeaux (IMB), UMR 5251 CNRS/University of Bordeaux, F-33400 Talence, France

[3] Imaging Division, UMC Utrecht, Heidelberglaan 100, 3508 GA Utrecht, The Netherlands

Version typeset July 7, 2022

Author to whom correspondence should be addressed. email: L.S.Bosma@umcutrecht.nl

Abstract

Background: Deformable image registration is increasingly used in radiotherapy to adapt the treatment plan and accumulate the delivered dose. Consequently, clinical workflows using deformable image registration require quick and reliable quality assurance to accept registrations. Additionally, for online adaptive radiotherapy, quality assurance without the need for an operator to delineate contours while the patient is on the treatment table is needed. Established quality assurance criteria such as the Dice similarity coefficient or Hausdorff distance lack these qualities and also display a limited sensitivity to registration errors beyond soft tissue boundaries.

Purpose: The purpose of this study is to investigate the existing intensity-based quality assurance criteria structural similarity and normalized mutual information for their ability to quickly and reliably identify registration errors for (online) adaptive radiotherapy and compare them to contour-based quality assurance criteria.

Methods: All criteria were tested using synthetic and simulated biomechanical deformations of 3D MR images as well as manually annotated 4D CT data. The quality assurance criteria were scored for classification performance, for their ability to predict the registration error, and for their spatial correlation with the registration error.

Results: We found that besides being fast and operator-independent, the intensity-based criteria have the highest area under the receiver operating characteristic curve and provide the best input for models to predict the registration error on all data sets. Structural similarity furthermore provides spatial information with a higher spatial correlation to the benchmark than the inverse consistency error, Jacobian determinant, and curl magnitude.

Conclusions: Intensity-based quality assurance criteria can provide the required confidence in decisions about using mono-modal registrations in clinical workflows. They

thereby enable automated quality assurance for deformable image registration in adaptive radiotherapy treatments.

I. Introduction

Radiotherapy is increasingly moving towards image-guided adaptive therapy workflows, which aim to compensate for the effect of motion both in between as well as during therapy sessions. To this end, the patients' internal anatomy can be imaged using cone-beam CT[1] or MRI[2,3] before and during treatment, which enables deformable image registration algorithms to extract anatomical motion information from these images. This information can subsequently be used to mitigate the effect of motion. To use motion information in clinical workflows, the motion estimations need to be reliable, accurate, and precise. Incorrect estimations can accumulate over time and decrease treatment quality and compromise patient safety. Additionally, the quality assurance needs to be fast, as the patients' anatomy can continue to change during assessment. Better tools for quality assurance of registration results has been identified as the main factor that may allow centres to use DIR more in clinical practise[4].

Commonly used quality assurance criteria that are advised for deformable image registration by the AAPM TG 132 Report[5] like the Dice similarity coefficient and Hausdorff distance score registrations by indicating some form of contour correspondence with a single number. While for applications like contour propagation and MLC-tracking this has been found to be sufficient, there are severe disadvantages for scoring deformable image registrations for dose accumulation and/or plan adaptation in this way. First, these criteria lack speed as they need two (sets of) delineated contours. This is labor intensive and time consuming, in particular for multi-slice or 3D data. Therefore, these criteria are not suited for online and/or real-time applications with the patient on the treatment table. Second, as these criteria only score the delineations, they lack reliability by being insensitive to registration errors in the soft tissue beyond the contoured organ boundaries. Furthermore, as they output a single number, these criteria do not provide any spatial information on the registration errors. Also the advised target registration error of anatomical landmarks annotated by experts suffers from similar shortcomings. Selecting the appropriate landmarks is a

laborious and time-consuming process and a lot of landmarks covering the region of interest are required as they provide an inherently local description of the registration performance.

The need for reliable quality assurance is further reinforced by the recent success of deep neural networks (DNNs) in medical image processing. In the recent past, DNN solutions have been employed for deformable image registration[6–8] as well as for quality assurance of image registration[9–11]. A limitation is that DNNs frequently lack several desirable properties of probabilistic models, such as uncertainty quantification and priors as well as a lack of transparency and that generalization of the trained models can be difficult. To facilitate the clinical translation of DNNs, these disadvantages can be largely alleviated if an independent quality assurance based on deterministic methods as an additional safeguard layer is performed.

In this paper, we evaluate therefore four deterministic contour-based criteria and two deterministic and fast operator-independent intensity-based quality assurance criteria on their ability to serve as the basis of a binary classifier to accept registrations for further clinical use, and to serve as the input for a model to predict the registration error. We also assess their potential to provide spatial information.

II. Methods

We compared four contour-based criteria and two intensity-based criteria. The contour-based criteria are: the Dice similarity coefficient[12], the Jaccard similarity index[13], the Hausdorff distance[14], and the mean Hausdorff distance[15]. The operator-independent intensity-based criteria are normalized mutual information[16] and structural similarity[17]. The contour-based criteria and normalized mutual information output a single scalar. Structural similarity provides a value for each voxel and can therefore also give the distribution of errors on a region of interest or a map of the registration error, indicating where a registration fails. As the benchmark for quality assurance we used the endpoint error[18] or -if no benchmark deformation vector field was available- the target registration error. To compare the criteria, we average the endpoint error, target registration error, and structural similarity over a contour-area and for normalized mutual information only consider the voxel intensities in this area.

All criteria are tested on three different data sets. First, on a set of synthetically deformed 3D MR images of prostate anatomies for ten patients. This allows us to use the endpoint error as a benchmark and provides a high number of deformations. Acquisition details can be found in the supplementary material. The synthetic deformations are introduced by randomly displacing every 30^{th} voxel in all three dimensions and using B-spline interpolation to determine the deformations of intermediate voxels. We generate 500 deformations for each of the 10 patients, drawing voxel displacements from a normal distribution with a standard deviation of 2 mm. To test the influence of the signal-to-noise ratio (SNR), we synthetically added increasing levels of Rician noise to the images, lowering their SNR from 12 to 9, 6, and 4, respectively.

Secondly, the criteria are tested on 3D MR datasets subjected to simulated biomechanical deformations. These simulations take into account the tissue-specific physical properties and represent an approximation to typical physiological deformations. This provides an anatomically correct benchmark. For a prostate patient, we simulated four motion patterns that are typically observed during treatments of 6 to 10 minutes using the finite element modeling software FEBio[19]. The motion patterns represent a rectal filling (maximum average displacement of the prostate of 4.3 mm), a bladder filling (3.2 mm), the average observed motion of a prostate during treatment (1.5 mm), and residual motion only (0.6 mm). These simulations were then used to create a 4D cine MR image series consisting of 11 images by deforming a 3D MR scan of a prostate cancer patient treated on the MR-Linac Unity system (Elekta AB, Stockholm, Sweden) installed at the UMC Utrecht. For more details of the motion patterns and finite element modeling, see[20]. Subsequently, the cine MR images are registered using five different variational DIR algorithms previously proposed in the context of MR guided radiotherapy[21–25]. To increase the size of the dataset, registrations were also performed on these images after 2-, 3-, and 4-fold downsampling or after adding four levels of Rician noise. In total, 1600 registration results have been investigated for this biomechanical simulations experiment. For these first two datasets the clinically delineated and the deformed prostate contours are used to compute the contour-based criteria and to average the endpoint error and intensity-based criteria over.

Finally, the quality assurance criteria are tested on ten thoracic 4D CT datasets from the DIR-lab database [1]. This publicly available dataset provides a spatially sparse anatomi-

[1]See https://med.emory.edu/departments/radiation-oncology/research-laboratories/deformable-image-

cally plausible benchmark. For images of full inhale and full exhale, 300 manually annotated anatomical landmarks are available to quantify the displacement[26,27]. As for the biomechanical dataset, we increase the size of this dataset eightfold by downsampling and adding noise. In addition, we use the five registration algorithms twice with different parameters. In total, 800 registrations have been investigated for this data set. Expert delineated lungs in full inhale and full exhale state are used to compute the contour-based criteria and to average the endpoint error and intensity-based criteria over.

We first evaluated the quality assurance criteria as the basis of a binary classifier for accepting deformable image registrations for clinical use. To this end, the mean endpoint error is used to divide the data into acceptable and unacceptable cases. We then trained a logistic regression model on the different quality assurance criteria. For the synthetic prostate data 10-fold cross-validation is used with one unseen patient in each test set. For the biomechanically simulated data 10-fold cross-validation with a random proportion of the data in the test set is used. And for the manually annotated data, 5-fold cross-validation is used with two previously unseen patients in each test, averaging over all possible combinations. The models are then tested and the area under the receiver operating characteristic (AUROC) curve is determined. The AUROC is the probability that for a randomly chosen acceptable and unacceptable case the classifier identifies them correctly.

Secondly, we compare the prediction performance for the investigated criteria. For this, we train a linear regression model to predict a registration error in mm based on the output of the different criteria. Then we evaluate the Pearson correlation between the predicted registration error and true registration error, and the absolute difference between the two (which we call prediction error). The same training and test sets as listed above are used for the synthetic and simulated data. For the manually annotated data we used 10-fold cross-validation with one unseen patient in the test set.

Finally, we compared the spatial information in the applicable quality assurance criteria. To this end, we train linear regression models using the voxel-by-voxel output from structural similarity, inverse consistency, the absolute deviation of the Jacobian determinant from unity $|1 - J(\mathbf{u} + 1)|$, and the curl magnitude $\|\nabla \times \mathbf{u}\|_2$. The benchmark is the voxel-by-voxel endpoint error. Using 10-fold cross-validation, we evaluate the models by computing the

registration/index.html

II. METHODS

gamma criterion[28]. We test the criteria on the biomechanical simulation of the prostate anatomy as it has known and realistic deformations and the prostate is modeled to have a Jacobian close to unity and close to vanishing curl magnitude. We use all voxels from a cube of 75x75x75 mm surrounding the prostate ($1.6{\cdot}10^5$ voxels) for all datapoints where the average endpoint error is in the top 10%, for memory purposes. For a set of gamma tolerances, we score the average gamma criterion over the cube as well as the percentage of voxels passing the gamma criterion ($\gamma \leq 1$).

III. Results

Table 1 shows the results for the synthetic deformations. The intensity-based criteria have the highest areas under the receiver operating characteristic curve (AUROC), and their prediction models show the highest correlation with the endpoint error and the lowest prediction error. This deviation from the true endpoint error is at least 1.5 times lower for both intensity-based criteria than for any contour-based criterion. For all criteria, the mean slope of their linear regression is lower than 1. For NMI (0.75) and SSIM (0.76) the slope is much closer to one than for any contour-based criterion (0.32 at most). This indicates a better sensitivity and smaller underestimation of the registration error. The full receiver operating characteristic curve can be found in Figure S1 in the supplementary material. Figure 1 shows a linear regression analysis for the prediction performance on a single unseen test patient. The patient with results closest to the mean of all ten patients as reported in Table 1 is shown. We can observe the higher correlations, smaller errors, and better slope alignments for the intensity-based criteria. The inter-patient performances for Dice and Jaccard shown here are considerably worse than their intra-patient performances (not shown). For the intensity-based criteria this difference is relatively small. For all criteria, the AUROC decreases with decreasing signal-to-noise-ratio (SNR), see Table S3 in the supplementary material. However, even on images with an SNR of 4, the intensity-based criteria perform better than all contour-based criteria do on the original images with an SNR of 12. The results are qualitatively the same for different choices of the cutoff to separate acceptable and unacceptable registrations, see Table S4 in the supplementary material.

For the biomechanically simulated deformations of the prostate (Table A1 in the supplementary material), we find qualitatively similar results. The intensity-based criteria outper-

Table 1: Classification and prediction results for the quality assurance criteria evaluated on the prostate for synthetic deformations. The results are averaged over the ten test patients (and all data points). Shown are the area under the receiver operating characteristic curve (AUROC), the Pearson correlation between the predicted and true endpoint errors, and their absolute difference as the prediction error.

QA criterion	AUROC	Correlation	Prediction error (mm)
Dice similarity coefficient	0.84	0.75	0.37
Jaccard index	0.84	0.75	0.37
Hausdorff distance	0.71	0.49	0.38
Mean Hausdorff distance	0.78	0.62	0.36
Mutual information	**0.92**	**0.91**	0.24
Structural similarity	**0.92**	**0.91**	**0.22**

form all contour-based criteria on all evaluations. The mean prediction errors for the NMI (0.04 mm) and the SSIM (0.07 mm) are at least halve as low as those for the contour-based criteria.

The gamma criterion evaluation results for the spatial correspondence are shown in Table 2 and Table S5 in the supplementary material. For any choice of tolerances, structural similarity has a mean gamma value at least a factor 1.4 lower than any other criterion. On average, the percentage of voxels passing the criterion is at least a factor of 1.2 higher than for any other criterion. For a 10%/2mm tolerance, (where 10% represents an error of 0.23 mm on average), the gamma pass rate for structural similarity is 95%. In Figure 2, a typical example of a transversal slice of the true and predicted endpoint errors from structural similarity, the inverse consistency error, Jacobian determinant, and curl magnitude is shown. We can observe the ability of the model based on structural similarity to localize the largest registration error, resulting in a higher gamma pass rate.

For the manually annotated 4D CT thoracic data sets (Table A2 in the supplementary material) the results are qualitatively similar to those above. The intensity-based criteria score best and at least as good as the contour-based on all evaluations. The mean prediction error for normalized mutual information is at least 1.3 times lower than those for the contour-based criteria.

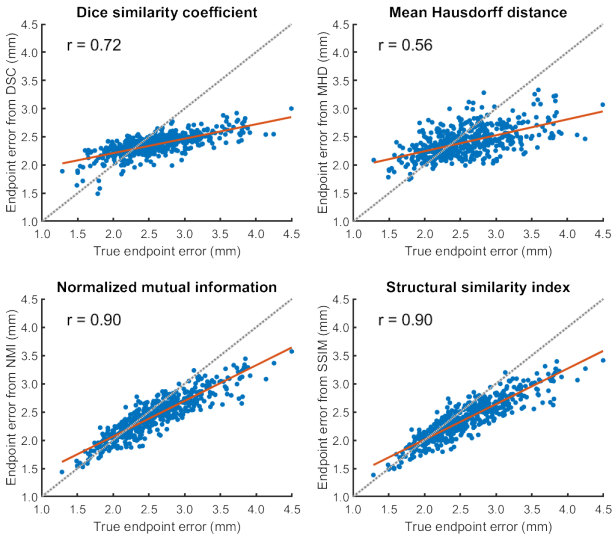

Figure 1: Prediction performance for a single test patient of the synthetic deformations for the Dice similarity coefficient (DSC), mean Hausdorff distance (MHD), normalized mutual information (NMI), and structural similarity index (SSIM). Plotted are the predicted endpoint errors and the true endpoint errors. A linear regression analysis is shown, and the Pearson correlation coefficient r is indicated. We can see the higher correlation that is also more aligned with the line with slope 1 for the intensity-based criteria. They also show a smaller spread around this line.

Table 2: Gamma criterion pass rate percentage evaluated on a box surrounding the prostate for different tolerances for the criteria holding spatial information.

QA criterion	5%/1mm	5%/2mm	10%/1mm	10%/2mm	10%/3mm	20%/2mm
Structural similarity	**75**	**81**	**92**	**95**	**97**	**98**
Inverse consistency	50	56	79	82	85	96
Jacobian determinant	56	62	83	86	89	97
Curl magnitude	57	63	83	85	88	96

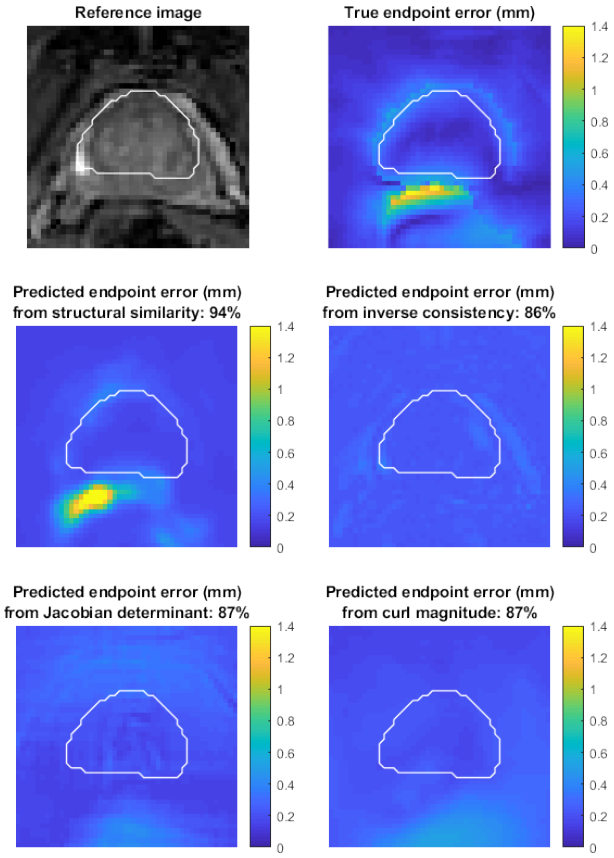

Figure 2: Transversal slice of the cube used for evaluation of the spatial correspondence. Shown are the reference image, the true endpoint error and the predicted endpoint errors using structural similarity, inverse consistency error, Jacobian determinant, and curl magnitude. The prostate contour is shown in white. The gamma pass rate percentage for 10%/2mm over the cube is indicated in the title. The datapoint with the results closest to the mean over the cross-validation is shown.

IV. Discussion

In this work, we evaluated multiple existing criteria on their capabilities for quality assurance of mono-modal image registration for MRI and CT. We have compared the operator-independent intensity-based normalized mutual information and structural similarity to the more established contour-based Dice similarity coefficient, Jaccard index, Hausdorff distance, and mean Hausdorff distance, and to the DVF-based spatial criteria inverse consistency error, Jacobian determinant, and curl magnitude. Both intensity-based criteria outperform all contour-based criteria on almost all datasets and evaluations. Across the three datasets, the prediction error is at least a factor of 1.6, 2.7 and 1.1 lower for the intensity-based criteria compared to the best performing contour-based criterion. This confirms the hypothesis that using the additional information in image intensities has benefits for quality assurance. Importantly, this comparatively high performance is maintained even for low SNR.

Additionally, structural similarity provides a spatial map of registration errors. This allows to observe distributions of the SSIM over a volume or identify local failures of image registration. We found its spatial correspondence to the benchmark to be considerably higher than conventional DVF-based spatial criteria. The second-best is the Jacobian determinant, but it misses registration errors not arising from the estimation of physiologically implausible deformations. Structural similarity does require image contrast to identify local misregistrations. This spatial map gives rise to possibilities such as to: only flag registration errors in regions where the planned dose (gradient) is above a particular threshold, find a map of the registration error multiplied by the planned dose (gradient), or spatially vary the cutoff value for the SSIM when using it to classify registrations. Additionally, when a registration is correct in the majority of the evaluated volume but fails locally, an aggregated single number lacks sensitivity. An error map or distribution might be able to reveal local misregistrations in this case. The spatial distribution thereby enables semi-automatic quality assurance by indicating problematic regions for an operator to investigate.

The advantage of using synthetic and simulated deformations is that the endpoint error can be used as the benchmark quality assurance criterion. The disadvantage is that for these deformed images the noise in the original image is deformed in the same way as the signal and (transient) image artifacts will appear in both images. Therefore these images are expected to be more similar than separately acquired independent images. For this reason, and to test

against a lower soft-tissue contrast, we also included 4D CT images with manually annotated landmarks. Their disadvantage is that the target registration error is only locally defined and prone to inter-observer differences. The intensity-based criteria showed consistently high performances also for this different contrast with separately acquired images. We should note that the results for the intensity similarity measures may depend on the presence of artefacts and other inconsistencies. All experiments in this paper were done on mono-modal images. Mono-modal image registration is an important aspect of real-time/online adaptive radiotherapy where fast and (semi-)automated quality assurance is required. Intensity-based quality assurance criteria are not suitable to validate cross-contrast image registrations. In these cases, criteria based on the expertise of the operator or potentially DNN solutions provide better options.

V. Conclusion

The presented study analyzed different contour-based and intensity-based quality assurance criteria for deformable image registration on a range of mono-modal data sets. Intensity-based criteria outperform contour-based criteria on almost all evaluations in terms of classification of unacceptable registrations and prediction of registration errors on both MRI and CT data. Both normalized mutual information and structural similarity are operator-independent, fast, robust, and show the highest specificity and sensitivity to detect misregistrations.

Between the two, structural similarity has the advantage of providing spatial information or a distribution of registration errors. Overall, structural similarity presents itself as a sound choice for fast (semi-)automated quality assurance to decide on accepting mono-modal registrations in clinical workflows. It is especially suitable for workflows under time-pressure or aiming to reduce operator burden.

VI. Acknowledgement

The collaboration project is co-funded by the PPP Allowance made available by Health~Holland, Top Sector Life Sciences & Health, to stimulate public-private partner-

ships.

References

[1] M. Guckenberger, Image-guided Radiotherapy Based on Kilovoltage Cone-beam Computed Tomography–A Review of Technology and Clinical Outcome, Eur Oncol Haematol **7**, 121–124 (2011).

[2] B. Raaymakers et al., Integrating a 1.5 T MRI scanner with a 6 MV accelerator: proof of concept, Physics in Medicine & Biology **54**, N229 (2009).

[3] S. Mutic and J. F. Dempsey, The ViewRay system: magnetic resonance–guided and controlled radiotherapy, in *Seminars in radiation oncology*, volume 24, pages 196–199, Elsevier, 2014.

[4] M. Hussein, A. Akintonde, J. McClelland, R. Speight, and C. H. Clark, Clinical use, challenges, and barriers to implementation of deformable image registration in radiotherapy–the need for guidance and QA tools, The British Journal of Radiology **94**, 20210001 (2021).

[5] K. K. Brock, S. Mutic, T. R. McNutt, H. Li, and M. L. Kessler, Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132, Medical physics **44**, e43–e76 (2017).

[6] X. Yang, R. Kwitt, M. Styner, and M. Niethammer, Quicksilver: Fast predictive image registration–a deep learning approach, NeuroImage **158**, 378–396 (2017).

[7] H. Sokooti, B. De Vos, F. Berendsen, B. P. Lelieveldt, I. Išgum, and M. Staring, Nonrigid image registration using multi-scale 3D convolutional neural networks, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 232–239, Springer, 2017.

[8] M.-M. Rohé, M. Datar, T. Heimann, M. Sermesant, and X. Pennec, SVF-Net: Learning deformable image registration using shape matching, in *International conference on medical image computing and computer-assisted intervention*, pages 266–274, Springer, 2017.

[9] K. A. Eppenhof and J. P. Pluim, Error estimation of deformable image registration of pulmonary CT scans using convolutional neural networks, Journal of medical imaging **5**, 024003 (2018).

[10] B. D. de Senneville, J. V. Manjón, and P. Coupé, RegQCNET: Deep quality control for image-to-template brain MRI affine registration, Physics in Medicine & Biology **65**, 225022 (2020).

[11] H. Sokooti, S. Yousefi, M. S. Elmahdy, B. P. Lelieveldt, and M. Staring, Hierarchical Prediction of Registration Misalignment using a Convolutional LSTM: Application to Chest CT Scans, IEEE Access (2021).

[12] L. R. Dice, Measures of the amount of ecologic association between species, Ecology **26**, 297–302 (1945).

[13] P. Jaccard, Nouvelles recherches sur la distribution florale, Bull. Soc. Vaud. Sci. Nat. **44**, 223–270 (1908).

[14] F. Hausdorff, *Grundzüge der mengenlehre*, volume 7, von Veit, 1914.

[15] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, Comparing images using the Hausdorff distance, IEEE Transactions on pattern analysis and machine intelligence **15**, 850–863 (1993).

[16] M. Hossny, S. Nahavandi, and D. Creighton, Comments on'Information measure for performance of image fusion', Electronics letters **44**, 1066–1067 (2008).

[17] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE transactions on image processing **13**, 600–612 (2004).

[18] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski, A database and evaluation methodology for optical flow, International journal of computer vision **92**, 1–31 (2011).

[19] S. A. Maas, B. J. Ellis, G. A. Ateshian, and J. A. Weiss, FEBio: finite elements for biomechanics, J Biomech Eng **134**, 011005 (2012).

[20] L. Bosma, C. Zachiu, M. G. Ries, B. D. de Senneville, and B. W. Raaymakers, Quantitative investigation of dose accumulation errors from intra-fraction motion in MRgRT for prostate cancer, Physics in Medicine & Biology (2021).

[21] B. K. Horn and B. G. Schunck, Determining optical flow, in *Techniques and Applications of Image Understanding*, volume 281, pages 319–331, International Society for Optics and Photonics, 1981.

[22] C. Zachiu, N. Papadakis, M. Ries, C. Moonen, and B. Denis de Senneville, An improved optical flow tracking technique for real-time MR-guided beam therapies in moving organs, Physics in Medicine & Biology **60**, 9003 (2015).

[23] B. Denis de Senneville, C. Zachiu, M. Ries, and C. Moonen, EVolution: an edge-based variational method for non-rigid multi-modal image registration, Physics in Medicine and Biology **61**, 7377–7396 (2016).

[24] C. Zachiu, B. Denis de Senneville, C. T. Moonen, B. W. Raaymakers, and M. Ries, Anatomically plausible models and quality assurance criteria for online mono-and multi-modal medical image registration, Physics in Medicine & Biology **63**, 155016 (2018).

[25] C. Zachiu et al., Anatomically-adaptive multi-modal image registration for image-guided external-beam radiotherapy, Physics in Medicine & Biology (2020).

[26] E. Castillo, R. Castillo, J. Martinez, M. Shenoy, and T. Guerrero, Four-dimensional deformable image registration using trajectory modeling, Physics in Medicine & Biology **55**, 305 (2009).

[27] R. Castillo, E. Castillo, R. Guerra, V. E. Johnson, T. McPhail, A. K. Garg, and T. Guerrero, A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets, Physics in Medicine & Biology **54**, 1849 (2009).

[28] D. A. Low, W. B. Harms, S. Mutic, and J. A. Purdy, A technique for the quantitative evaluation of dose distributions, Medical physics **25**, 656–661 (1998).