


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

Lando S. Bosma<sup>1</sup>  | Cornel Zachiu<sup>1</sup>  | Baudouin Denis de Senneville<sup>1,2</sup> | Bas W. Raaymakers<sup>1</sup> | Mario Ries<sup>3</sup>

<sup>1</sup>Department of Radiotherapy, UMC Utrecht, Utrecht, The Netherlands

<sup>2</sup>Institut de Mathématiques de Bordeaux (IMB), University of Bordeaux, Talence, France

<sup>3</sup>Imaging Division, UMC Utrecht, Utrecht, The Netherlands

## Correspondence

L. S. Bosma, Department of Radiotherapy, UMC Utrecht, Heidelberglaan 100, 3508 GA Utrecht, The Netherlands.  
Email: [L.S.Bosma@umcutrecht.nl](mailto:L.S.Bosma@umcutrecht.nl)

## Funding information

Health~Holland, Grant/Award Number: LSHM19031

## 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 information.

**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 gamma pass rate of the predicted registration error than commonly used spatial quality assurance criteria.

**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.

## KEYWORDS

adaptive radiotherapy quality assurance, deformable image registration, quality assurance

## 1 | 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 patient's internal anatomy can be imaged using cone-beam CT<sup>1</sup> or MRI<sup>2,3</sup> before and during treatment, which enables

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Medical Physics* published by Wiley Periodicals LLC on behalf of American Association of Physicists in Medicine.

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, quality assurance needs to be fast, as the patient's anatomy can continue to change during the assessment. Better tools for quality assurance of registration results has been identified as the main factor that may allow centers to use DIR more in clinical practice.<sup>4</sup>

Commonly used quality assurance criteria that are advised for deformable image registration by the AAPM TG 132 Report<sup>5</sup> 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 to 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<sup>6–8</sup> as well as for quality assurance of image registration.<sup>9–11</sup> 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.

## 2 | METHODS

We compared four contour-based criteria and two intensity-based criteria. The contour-based criteria are: the Dice similarity coefficient,<sup>12</sup> the Jaccard similarity index,<sup>13</sup> the Hausdorff distance,<sup>14</sup> and the mean Hausdorff distance.<sup>15</sup> The operator-independent intensity-based criteria are normalized mutual information<sup>16</sup> and structural similarity.<sup>17</sup> 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<sup>18</sup> 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 consider the normalized mutual information for voxel intensities in this area only.

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<sup>th</sup> voxel in all three dimensions, drawing voxel displacements from a normal distribution with a standard deviation of 2 mm. B-spline interpolation is used to determine the deformations of intermediate voxels, ensuring spatial continuity. We generate 500 deformations for each of the 10 patients. To test the influence of the signal-to-noise ratio (SNR), we synthetically added noise patterns of Rician shape to the images, lowering their SNR from 12 to 9, 6, and 4, respectively. For low SNR this approximates the noise in MR images.<sup>19,20</sup>

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 of 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 min using the finite element modeling software FEBio.<sup>21</sup> 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.<sup>22</sup> Subsequently, the cine MR images are registered using five different variational DIR algorithms previously proposed in the context of MR-guided radiotherapy.<sup>23–27</sup> To increase the size of the dataset, registrations were also performed after 2-, 3-, and 4-fold downsampling the image resolution and 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.<sup>1</sup> This publicly available dataset provides a spatially sparse anatomically plausible benchmark. For images of full inhale and full exhale, 300 manually annotated anatomical landmarks are available to quantify the displacement.<sup>28,29</sup> 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 ranks the acceptable case higher than the unacceptable case.

Secondly, we compare the prediction performance for the investigated criteria. For this, we train a linear regression model to predict a registration error in millimeters 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 predict the endpoint error using a fixed effect linear regression model fit on the spatial quality assurance metric. These predictors are the voxel-by-voxel output from structural similarity and other commonly used spatial quality assurance criteria: 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 for this model is the known voxel-by-voxel endpoint error. We evaluate the models by computing the gamma criterion with the known endpoint error as the distribution<sup>30,31</sup> using 10-fold cross-validation. 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 75 x 75 x 75 mm surrounding the prostate ( $1.6 \cdot 10^5$  voxels) for all data points where the average endpoint error on the prostate 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 ( $\gamma \leq 1$ ).

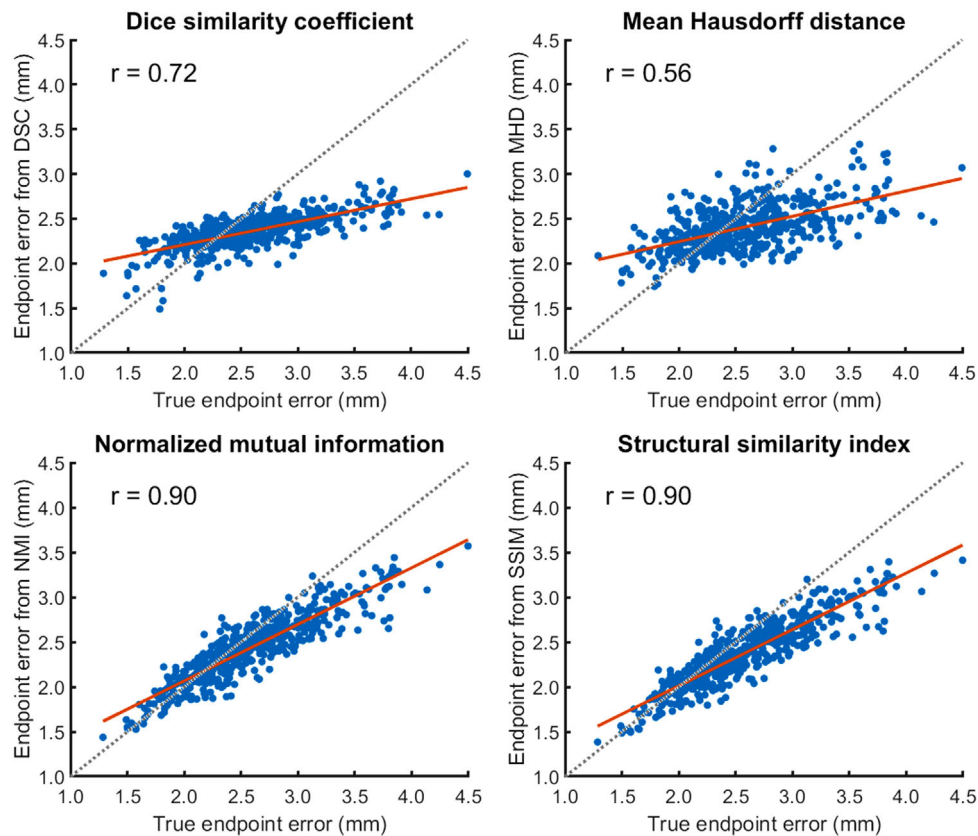
### 3 | RESULTS

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 normalized mutual information (0.75) and structural similarity (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 a smaller underestimation of the registration error. The full receiver operating characteristic curve can be found in Figure S1 in the Supporting Information. Figure 1 shows a linear regression analysis for the prediction performance on a single

<sup>1</sup> See <https://med.emory.edu/departments/radiation-oncology/research-laboratories/deformable-image-registration/index.html>

**TABLE 1** Classification and prediction results for the quality assurance criteria evaluated on the prostate for synthetic deformations. We show the mean (standard deviation) over the ten test patients and their 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	0.86 (0.04)	0.75 (0.20)	0.37 (0.09)
Jaccard index	0.86 (0.04)	0.75 (0.11)	0.37 (0.09)
Hausdorff distance	0.72 (0.05)	0.49 (0.18)	0.38 (0.09)
Mean Hausdorff distance	0.80 (0.04)	0.62 (0.24)	0.36 (0.07)
Mutual information	<b>0.95 (0.01)</b>	<b>0.91 (0.19)</b>	<b>0.24 (0.09)</b>
Structural similarity	<b>0.95 (0.01)</b>	<b>0.91 (0.19)</b>	<b>0.22 (0.10)</b>



**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 (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.

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 higher correlations, smaller errors, and better slope alignments for the intensity-based criteria. The interpatient performances for Dice and Jaccard shown here are considerably worse than their inpatient 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 S1 in the Supporting Information. 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 S2 in the Supporting Information.

For the biomechanically simulated deformations of the prostate (Table S3 in the Supporting Information), we find qualitatively similar results. The intensity-based criteria outperform all contour-based criteria on all evaluations. The mean prediction errors for normalized

**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	2%/2mm	10%/1mm	10%/2mm	10%/2mm	20%/2mm
Structural similarity	76 (17)	82 (14)	92 (6)	95 (4)	97 (3)	99 (2)
Inverse consistency	50 (19)	56 (18)	82 (14)	85 (12)	88 (11)	96 (2)
Jacobian determinant	57 (16)	63 (14)	84 (12)	87 (10)	90 (8)	97 (2)
Curl magnitude	58 (18)	64 (17)	84 (13)	87 (11)	90 (9)	96 (2)

mutual information (0.04 mm) and structural similarity (0.07 mm) are at least half as low as those for the contour-based criteria.

The gamma pass rate evaluation for the spatial correspondence is shown in Table 2 and the average criterion results are shown in Table S4 in the Supporting Information. 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 0.24 mm on average), the gamma pass rate for structural similarity is 95%. In Figure 2, a representative 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 S5 in the Supporting Information) the results are qualitatively similar to those above. The intensity-based criteria score best and at least as well 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.

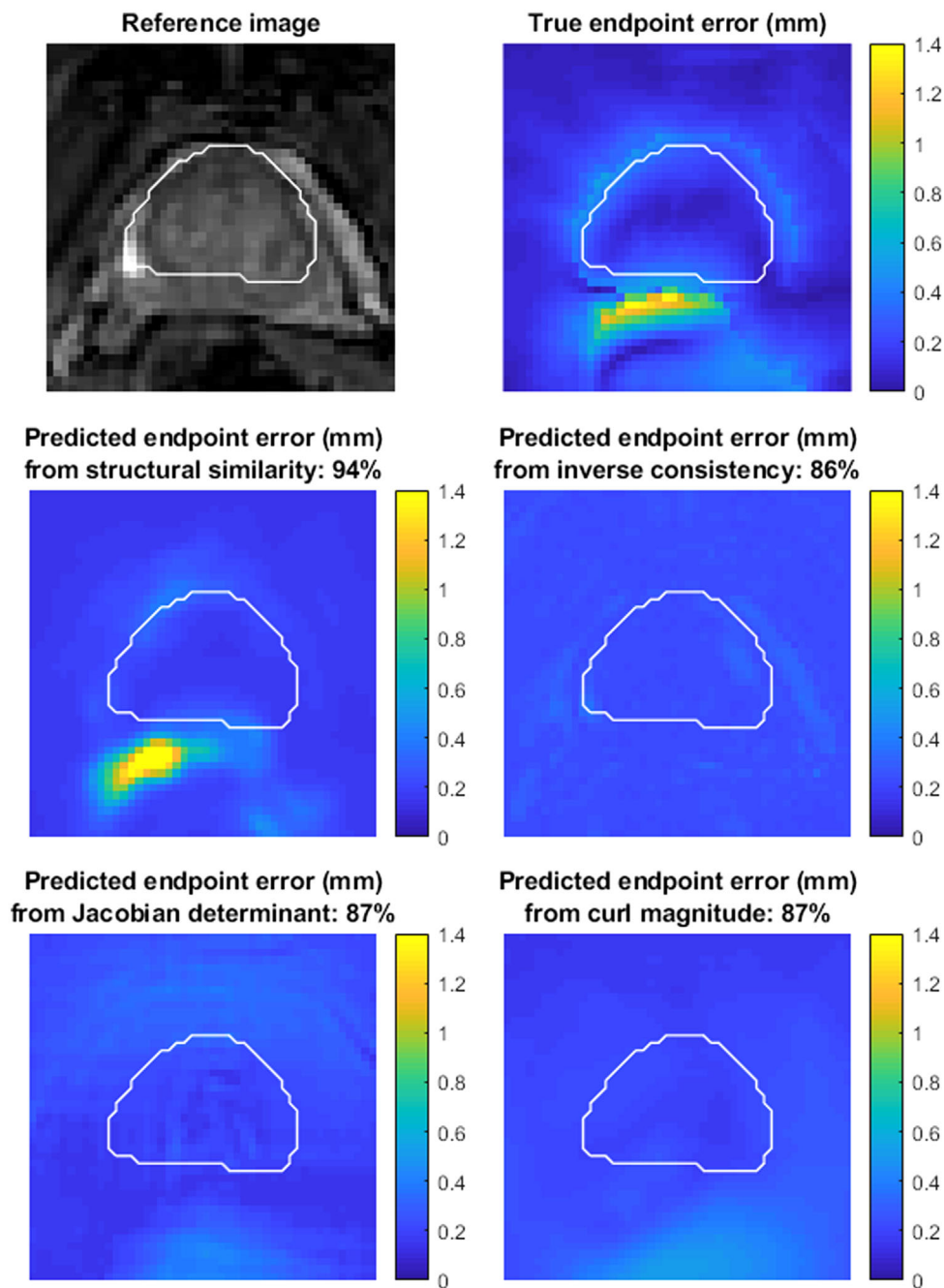
## 4 | DISCUSSION

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. The relative improvement is lower for the data set evaluated on the lungs as there are limited contrast details in the lungs as well as large dark signal voids. The benefits of intensity-based criteria are most pronounced for data sets and anatomical regions with sufficient image contrast. Importantly, the comparatively high performance is maintained even for low SNR. Since the noise on the images was simulated, we only evaluate the contour-based criteria on the clinical contours delineated on the original images in this analysis. It is expected that the reliability of contour-based criteria decreases for lower SNR as it will be harder to accurately delineate structures.

Additionally, structural similarity provides a spatial map of registration errors. This allows to observe distributions of structural similarity 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 only flagging registration errors in regions where the planned dose (gradient) is above a particular threshold, finding a map of the registration error multiplied by the planned dose (gradient), or spatially varying the cutoff value for structural similarity 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



**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 data point with the results closest to the mean over the cross-validation is shown.

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 artifacts 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 on neural networks may provide better options.

## 5 | 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.


## ACKNOWLEDGMENTS

The collaboration project is co-funded by the PPP Allowance made available by Health~Holland, Top Sector Life Sciences & Health, to stimulate public-private partnerships. The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.


## CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

## ORCID

Lando S. Bosma 

<https://orcid.org/0000-0002-8206-271X>

Cornel Zachiu 

<https://orcid.org/0000-0001-9755-4584>

## REFERENCES

- Guckenberger M. Image-guided radiotherapy based on kilovoltage cone-beam computed tomography - a review of technology and clinical outcome. *Eur Oncol Haematol*. 2011;7:121-124.
- Raaymakers B, Lagendijk J, Overweg J, et al. Integrating a 1.5 T MRI scanner with a 6 MV accelerator: proof of concept. *Phys Med Biol*. 2009;54:N229.
- Mutic S, Dempsey JF. The ViewRay system: magnetic resonance-guided and controlled radiotherapy. In: *Seminars in Radiation Oncology*. vol. 24. Elsevier; 2014. 196-199.
- Hussein M, Akintonde A, McClelland J, Speight R, Clark CH. Clinical use, challenges, and barriers to implementation of deformable image registration in radiotherapy – the need for guidance and QA tools. *Br J Radiol*. 2021;94:20210001.
- 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–e76.
- Yang X, Kwitt R, Styner M, Niethammer M. Quicksilver: Fast predictive image registration – a deep learning approach. *NeuroImage*. 2017;158:378-396.
- Sokooti H, De Vos B, Berendsen F, Lelieveldt BP, Išgum I, Staring M. Nonrigid image registration using multi-scale 3D convolutional neural networks. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2017:232-239.
- Rohé MM, Datar M, Heimann T, Sermesant M, Pennec X. SVF-Net: Learning deformable image registration using shape matching. In: *International conference on medical image computing and computer-assisted intervention*. Springer; 2017:266-274.
- Eppenhof KA, Pluim JP. Error estimation of deformable image registration of pulmonary CT scans using convolutional neural networks. *J Med Imaging*. 2018;5:024003.
- Denis de Senneville B, Manjón JV, Coupé P. RegQCNET: Deep quality control for image-to-template brain MRI affine registration. *Phys Med Bio*. 2020;65:225022.
- Sokooti H, Yousefi S, Elmahdy MS, Lelieveldt BP, Staring M. Hierarchical Prediction of Registration Misalignment using a Convolutional LSTM: Application to Chest CT Scans. *IEEE Access*; 2021.
- Dice LR. Measures of the amount of ecologic association between species. *Ecology*. 1945;26:297-302.
- Jaccard P. Nouvelles recherches sur la distribution florale. *Bull Soc Vaud Sci Nat*. 1908;44:223-270.
- Hausdorff F. *Grundzüge der mengenlehre*. Vol 7. von Veit; 1914.
- Huttenlocher DP, Klanderman GA, Rucklidge WJ. Comparing images using the Hausdorff distance. *IEEE Trans Pattern Anal Mach Intell*. 1993;15:850-863.
- Hossny M, Nahavandi S, Creighton D. Comments on 'Information measure for performance of image fusion'. *Electron Lett*. 2008;44:1066-1067.
- Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process*. 2004;13:600-612.
- Baker S, Scharstein D, Lewis J, Roth S, Black MJ, Szeliski R. A database and evaluation methodology for optical flow. *Int J Comput Vis*. 2011;92:1-31.
- Gudbjartsson H, Patz S. The Rician distribution of noisy MRI data. *Magn Reson Med*. 1995;34:910-914.
- Cárdenas-Blanco A, Tejos C, Irarrazaval P, Cameron I. Noise in magnitude magnetic resonance images. *Concepts Magnetic Resonance Part A: An Educational Journal*. 2008;32:409-416.
- Maas SA, Ellis BJ, Ateshian GA, Weiss JA. FEBio: finite elements for biomechanics. *J Biomech Eng*. 2012;134:011005.
- Bosma LS, Zachiu C, Ries MG, Denis de Senneville B, Raaymakers BW. Quantitative investigation of dose accumulation errors from intra-fraction motion in MRgRT for prostate cancer. *Phys Med Biol*. 2021;66(6):065002.
- Horn BK, Schunck BG. Determining optical flow. In: *Techniques and Applications of Image Understanding*, Vol 281. International Society for Optics and Photonics; 1981:319-331.
- Zachiu C, Papadakis N, Ries M, Moonen C, Denis de Senneville B. An improved optical flow tracking technique for real-time MR-guided beam therapies in moving organs. *Phys Med Biol*. 2015;60:9003.
- Denis de Senneville B, Zachiu C, Ries M, Moonen C. EVolution: an edge-based variational method for non-rigid multi-modal image registration. *Phys Med Biol*. 2016;61:7377-7396.
- Zachiu C, Denis de Senneville B, Moonen CT, Raaymakers BW, Ries M. Anatomically plausible models and quality assurance

- criteria for online mono-and multi-modal medical image registration. *Phys Med Biol*. 2018;63:155016.
27. Zachiu C, Denis de Senneville B, Willigenburg T, et al. Anatomically-adaptive multi-modal image registration for image-guided external-beam radiotherapy. *Phys Med Biol*. 2020;65(21):215028.
  28. Castillo E, Castillo R, Martinez J, Shenoy M, Guerrero T. Four-dimensional deformable image registration using trajectory modeling. *Phys Med Bio*. 2009;55:305.
  29. 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 Bio*. 2009;54:1849.
  30. Low DA, Harms WB, Mutic S, Purdy JA. A technique for the quantitative evaluation of dose distributions. *Med Phys*. 1998;25:656-661.
  31. Bernatowicz K, Grussu F, Ligerio M, Garcia A, Delgado E, Perez-Lopez R. Robust imaging habitat computation using voxel-wise radiomics features. *Sci Rep*. 2021;11:1-8.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Bosma LS, Zachiu C, Denis de Senneville B, Raaymakers BW, Ries M. Technical note: Intensity-based quality assurance criteria for deformable image registration in image-guided radiotherapy. *Med Phys*. 2023;50:5715–5722.  
<https://doi.org/10.1002/mp.16367>