



**HAL**  
open science

## 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

### ► To cite this version:

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. *Medical Physics*, 2023, 50 (9), pp.5715-5722. 10.1002/mp.16367 . hal-04268726

**HAL Id: hal-04268726**

**<https://hal.science/hal-04268726>**

Submitted on 2 Nov 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

L S Bosma<sup>1</sup>, C Zachiu<sup>1</sup>, B Denis de Senneville<sup>1,2</sup>, B W Raaymakers<sup>1</sup>,  
M Ries<sup>3</sup>

<sup>1</sup> Department of Radiotherapy, UMC Utrecht, Heidelberglaan 100, 3508 GA Utrecht, The Netherlands

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

<sup>3</sup> 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

39 thereby enable automated quality assurance for deformable image registration in adap-  
40 tive radiotherapy treatments.  
41

## 42 **1. Introduction**

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

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

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

71 The need for reliable quality assurance is further reinforced by the recent success of deep  
72 neural networks (DNNs) in medical image processing. In the recent past, DNN solutions  
73 have been employed for deformable image registration<sup>6-8</sup> as well as for quality assurance of  
74 image registration<sup>9-11</sup>. A limitation is that DNNs frequently lack several desirable proper-  
75 ties of probabilistic models, such as uncertainty quantification and priors as well as a lack  
76 of transparency and that generalization of the trained models can be difficult. To facilitate  
77 the clinical translation of DNNs, these disadvantages can be largely alleviated if an indepen-  
78 dent quality assurance based on deterministic methods as an additional safeguard layer is  
79 performed.

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

## 85 II. Methods

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

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

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

127 Finally, the quality assurance criteria are tested on ten thoracic 4D CT datasets from  
128 the DIR-lab database<sup>1</sup>. This publicly available dataset provides a spatially sparse anatomi-

---

<sup>1</sup>See [https://med.emory.edu/departments/radiation-oncology/research-laboratories/deformable-image-](https://med.emory.edu/departments/radiation-oncology/research-laboratories/deformable-image)

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

136 We first evaluated the quality assurance criteria as the basis of a binary classifier for ac-  
137 cepting deformable image registrations for clinical use. To this end, the mean endpoint error  
138 is used to divide the data into acceptable and unacceptable cases. We then trained a logistic  
139 regression model on the different quality assurance criteria. For the synthetic prostate data  
140 10-fold cross-validation is used with one unseen patient in each test set. For the biomechan-  
141 ically simulated data 10-fold cross-validation with a random proportion of the data in the  
142 test set is used. And for the manually annotated data, 5-fold cross-validation is used with  
143 two previously unseen patients in each test, averaging over all possible combinations. The  
144 models are then tested and the area under the receiver operating characteristic (AUROC)  
145 curve is determined. The AUROC is the probability that for a randomly chosen acceptable  
146 and unacceptable case the classifier identifies them correctly.

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

154 Finally, we compared the spatial information in the applicable quality assurance criteria.  
155 To this end, we train linear regression models using the voxel-by-voxel output from structural  
156 similarity, inverse consistency, the absolute deviation of the Jacobian determinant from unity  
157  $|1 - J(\mathbf{u} + 1)|$ , and the curl magnitude  $\|\nabla \times \mathbf{u}\|_2$ . The benchmark is the voxel-by-voxel  
158 endpoint error. Using 10-fold cross-validation, we evaluate the models by computing the

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

### 166 III. Results

167 Table 1 shows the results for the synthetic deformations. The intensity-based criteria have  
168 the highest areas under the receiver operating characteristic curve (AUROC), and their  
169 prediction models show the highest correlation with the endpoint error and the lowest pre-  
170 diction error. This deviation from the true endpoint error is at least 1.5 times lower for both  
171 intensity-based criteria than for any contour-based criterion. For all criteria, the mean slope  
172 of their linear regression is lower than 1. For NMI (0.75) and SSIM (0.76) the slope is much  
173 closer to one than for any contour-based criterion (0.32 at most). This indicates a better  
174 sensitivity and smaller underestimation of the registration error. **The full receiver operat-**  
175 **ing characteristic curve can be found in Figure S1 in the supplementary material.** Figure  
176 1 shows a linear regression analysis for the prediction performance on a single unseen test  
177 patient. The patient with results closest to the mean of all ten patients as reported in Table  
178 1 is shown. We can observe the higher correlations, smaller errors, and better slope align-  
179 ments for the intensity-based criteria. **The inter-patient performances for Dice and Jaccard**  
180 **shown here are considerably worse than their intra-patient performances (not shown).** For  
181 the intensity-based criteria this difference is relatively small. For all criteria, the AUROC  
182 decreases with decreasing signal-to-noise-ratio (SNR), **see Table S3 in the supplementary**  
183 **material.** However, even on images with an SNR of 4, the intensity-based criteria perform  
184 better than all contour-based criteria do on the original images with an SNR of 12. **The**  
185 **results are qualitatively the same for different choices of the cutoff to separate acceptable**  
186 **and unacceptable registrations, see Table S4 in the supplementary material.**

187 For the biomechanically simulated deformations of the prostate (Table A1 in the supple-  
188 mentary 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	<b>0.92</b>	<b>0.91</b>	0.24
Structural similarity	<b>0.92</b>	<b>0.91</b>	<b>0.22</b>

189 form all contour-based criteria on all evaluations. **The mean prediction errors for the NMI**  
 190 **(0.04 mm) and the SSIM (0.07 mm) are at least halve as low as those for the contour-based**  
 191 **criteria.**

192 **The gamma criterion evaluation results for the spatial correspondence are shown in**  
 193 **Table 2 and Table S5 in the supplementary material. For any choice of tolerances, structural**  
 194 **similarity has a mean gamma value at least a factor 1.4 lower than any other criterion.**  
 195 **On average, the percentage of voxels passing the criterion is at least a factor of 1.2 higher**  
 196 **than for any other criterion. For a 10%/2mm tolerance, (where 10% represents an error of**  
 197 **0.23 mm on average), the gamma pass rate for structural similarity is 95%. In Figure 2, a**  
 198 **typical example of a transversal slice of the true and predicted endpoint errors from structural**  
 199 **similarity, the inverse consistency error, Jacobian determinant, and curl magnitude is shown.**  
 200 **We can observe the ability of the model based on structural similarity to localize the largest**  
 201 **registration error, resulting in a higher gamma pass rate.**

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



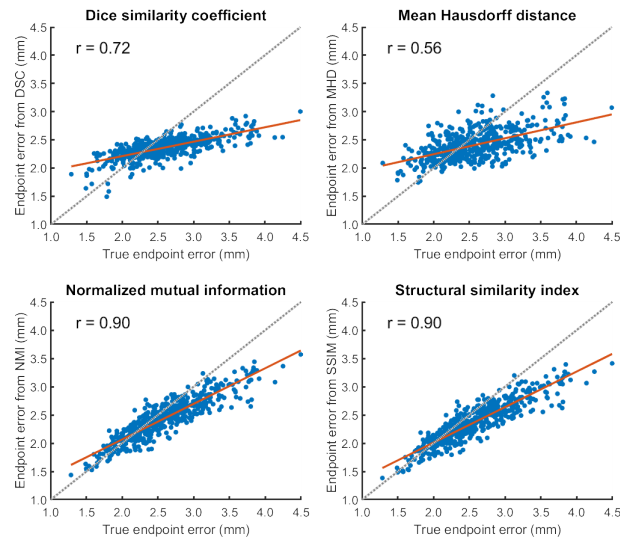


Figure 1: Prediction performance for a single test patient of the synthetic deformations for the Dice similarity coefficient (DSC), mean Hausdorff distance (MHD), **normalized** mutual information (NMI), and structural similarity index (SSIM). Plotted are the predicted endpoint errors and the true endpoint 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	<b>75</b>	<b>81</b>	<b>92</b>	<b>95</b>	<b>97</b>	<b>98</b>
Inverse consistency	50	56	79	82	85	96
Jacobian determinant	56	62	83	86	89	97
Curl magnitude	57	63	83	85	88	96

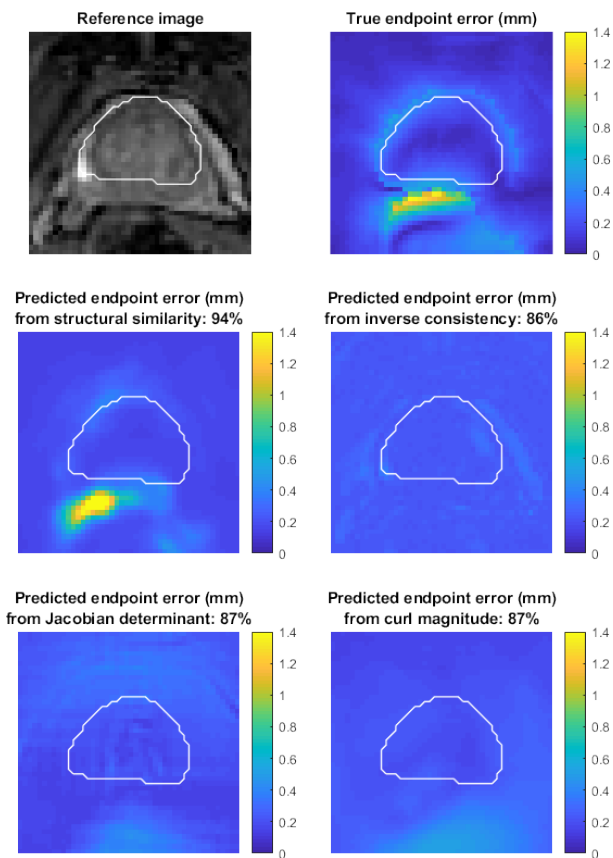


Figure 2: Transversal slice of the cube used for evaluation of the spatial correspondence. Shown are the reference image, the true endpoint error and the predicted endpoint errors using structural similarity, inverse consistency error, Jacobian determinant, and curl magnitude. The prostate contour is 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

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

## 249 V. Conclusion

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

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

## 262 VI. Acknowledgement

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

265 ships.

## References

- <sup>1</sup> 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).
- <sup>2</sup> 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).
- <sup>3</sup> 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.
- <sup>4</sup> 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).
- <sup>5</sup> 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).
- <sup>6</sup> X. Yang, R. Kwitt, M. Styner, and M. Niethammer, Quicksilver: Fast predictive image registration—a deep learning approach, *NeuroImage* **158**, 378–396 (2017).
- <sup>7</sup> 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.
- <sup>8</sup> 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.
- <sup>9</sup> 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).
-

- 296 <sup>10</sup> B. D. de Senneville, J. V. Manjón, and P. Coupé, RegQCNET: Deep quality control  
297 for image-to-template brain MRI affine registration, *Physics in Medicine & Biology* **65**,  
298 225022 (2020).
- 299 <sup>11</sup> H. Sokooti, S. Yousefi, M. S. Elmahdy, B. P. Lelieveldt, and M. Staring, Hierarchical  
300 Prediction of Registration Misalignment using a Convolutional LSTM: Application to  
301 Chest CT Scans, *IEEE Access* (2021).
- 302 <sup>12</sup> L. R. Dice, Measures of the amount of ecologic association between species, *Ecology* **26**,  
303 297–302 (1945).
- 304 <sup>13</sup> P. Jaccard, Nouvelles recherches sur la distribution florale, *Bull. Soc. Vaud. Sci. Nat.*  
305 **44**, 223–270 (1908).
- 306 <sup>14</sup> F. Hausdorff, *Grundzüge der mengenlehre*, volume 7, von Veit, 1914.
- 307 <sup>15</sup> D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, Comparing images using  
308 the Hausdorff distance, *IEEE Transactions on pattern analysis and machine intelligence*  
309 **15**, 850–863 (1993).
- 310 <sup>16</sup> M. Hossny, S. Nahavandi, and D. Creighton, Comments on 'Information measure for  
311 performance of image fusion', *Electronics letters* **44**, 1066–1067 (2008).
- 312 <sup>17</sup> Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, Image quality assessment:  
313 from error visibility to structural similarity, *IEEE transactions on image processing* **13**,  
314 600–612 (2004).
- 315 <sup>18</sup> S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski, A database and  
316 evaluation methodology for optical flow, *International journal of computer vision* **92**,  
317 1–31 (2011).
- 318 <sup>19</sup> S. A. Maas, B. J. Ellis, G. A. Ateshian, and J. A. Weiss, FEBio: finite elements for  
319 biomechanics, *J Biomech Eng* **134**, 011005 (2012).
- 320 <sup>20</sup> L. Bosma, C. Zachiu, M. G. Ries, B. D. de Senneville, and B. W. Raaymakers, Quan-  
321 titative investigation of dose accumulation errors from intra-fraction motion in MRgRT  
322 for prostate cancer, *Physics in Medicine & Biology* (2021).

- 323 <sup>21</sup> B. K. Horn and B. G. Schunck, Determining optical flow, in *Techniques and Applications*  
324 *of Image Understanding*, volume 281, pages 319–331, International Society for Optics  
325 and Photonics, 1981.
- 326 <sup>22</sup> C. Zachiu, N. Papadakis, M. Ries, C. Moonen, and B. Denis de Senneville, An improved  
327 optical flow tracking technique for real-time MR-guided beam therapies in moving or-  
328 gans, *Physics in Medicine & Biology* **60**, 9003 (2015).
- 329 <sup>23</sup> B. Denis de Senneville, C. Zachiu, M. Ries, and C. Moonen, EVolution: an edge-based  
330 variational method for non-rigid multi-modal image registration, *Physics in Medicine*  
331 *and Biology* **61**, 7377–7396 (2016).
- 332 <sup>24</sup> C. Zachiu, B. Denis de Senneville, C. T. Moonen, B. W. Raaymakers, and M. Ries,  
333 Anatomically plausible models and quality assurance criteria for online mono-and multi-  
334 modal medical image registration, *Physics in Medicine & Biology* **63**, 155016 (2018).
- 335 <sup>25</sup> C. Zachiu et al., Anatomically-adaptive multi-modal image registration for image-guided  
336 external-beam radiotherapy, *Physics in Medicine & Biology* (2020).
- 337 <sup>26</sup> E. Castillo, R. Castillo, J. Martinez, M. Shenoy, and T. Guerrero, Four-dimensional  
338 deformable image registration using trajectory modeling, *Physics in Medicine & Biology*  
339 **55**, 305 (2009).
- 340 <sup>27</sup> R. Castillo, E. Castillo, R. Guerra, V. E. Johnson, T. McPhail, A. K. Garg, and T. Guer-  
341 rero, A framework for evaluation of deformable image registration spatial accuracy using  
342 large landmark point sets, *Physics in Medicine & Biology* **54**, 1849 (2009).
- 343 <sup>28</sup> D. A. Low, W. B. Harms, S. Mutic, and J. A. Purdy, A technique for the quantitative  
344 evaluation of dose distributions, *Medical physics* **25**, 656–661 (1998).
-