Automation and individualization of radiotherapy treatment planning in head and neck cancer patients
Kierkels, Roel Godefridus Josefina

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2019

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
An automated, quantitative, and case-specific evaluation of deformable image registration in computed tomography images

Kerkels, Roel G J
den Otter, Lydia A
Korevaar, Erik W
Langendijk, Johannes A
van der Schaaf, Arjen
Knopf, Antje-Christin
Sijtsema, Nanna M

Physics in Medicine and Biology 63, 4 (2017)
Abstract

A prerequisite for adaptive dose tracking in radiotherapy is the assessment of the deformable image registration (DIR) quality. In this work, various metrics that quantify DIR uncertainties are investigated using realistic deformation fields of 26 head and neck and 12 lung cancer patients. Metrics related to the physiologically feasibility (the Jacobian determinant, harmonic energy [HE], and octahedral shear strain [OSS]) and numerically robustness of the deformation (the inverse consistency error [ICE], transitivity error [TE], and distance discordance metric [DDM]) were investigated. The deformable registrations were performed using a B-spline transformation model. The DIR error metrics were log-transformed and correlated (Pearson) against the log-transformed ground-truth error on a voxel level. Correlations of $r \geq 0.5$ were found for the DDM and HE. Given a DIR tolerance threshold of 2.0 mm and a negative predictive value of 0.90, the DDM and HE thresholds were 0.49 mm and 0.014, respectively. In conclusion, the log-transformed DDM and HE can be used to identify voxels at risk for large DIR errors with a large negative predictive value. The HE and/or DDM can therefore be used to perform automated quality assurance of each CT-based DIR for head and neck and lung cancer patients.
6.1 Introduction

Deformable image registration (DIR) algorithms are increasingly used in the field of radiation oncology, and are a key component in adaptive radiotherapy (Jaffray et al. 2010, Grégoire et al. 2012, Schwartz et al. 2013). DIR has the potential to quantify temporal or anatomical changes over the fractionated course of therapy and is mainly used for contour propagation, dose mapping and response assessment (Brock et al. 2017). Since DIR supports clinical decisions, quantification of the DIR error is required per case and on a voxel level (Jaffray et al. 2010, Bertelsen et al. 2011). The quality of the propagated contours and mapped doses is highly dependent on the combination of anatomical site, clinical application, image modalities and DIR algorithm (Nie et al. 2013, Janssens et al. 2009, Rosu et al. 2005). The propagated contours can be assessed by visual inspection and edited when needed, whereas this is not the case for the mapped dose distributions, which directly rely on the DIR quality. A quantitative (and preferably automated) method for the assessment of DIR quality is therefore needed.

A growing number of publications report on the DIR accuracy for various anatomical sites, imaging modalities, and DIR algorithms (e.g. based on intensity matching, or geometry-based registrations) (Kirby et al. 2016, Nie et al. 2013, Al-Mayah et al. 2015, Brock 2010, Zhong et al. 2010, Castadot et al. 2008, Janssens et al. 2009, Rosu et al. 2005, Kashani et al. 2008, Mohamed et al. 2015). The main components in DIR that introduce some kind of uncertainty can be found in the similarity metric, transformation model, optimization function, and the image resampler. Moreover, the biomechanical characteristics of the different patient tissues are not directly modeled by the DIR algorithm, which leads to uncertainties in the transformation. Therefore, an important step during the commissioning of a DIR system is the tuning of its parameters (Brouwer et al. 2014). Although such tuning improves DIR performance, case-specific DIR errors remain to be examined.

Recently, the AAPM Task Group 132 (TG132) published recommendations on the implementation, commissioning and evaluation of DIR algorithms (Brock et al. 2017). This Task Group stressed that it is important to understand and report the uncertainties associated with the DIR system and each registration. Moreover, recommendations are given on “validation and quality assurance of image registration and fusion at treatment planning, treatment delivery, adaptive re-planning, and response assessment” (Brock et al. 2017). A platform for the evaluation of DIR uncertainties is therefore required. Several qualitative and quantitative measures of DIR error assessment are described in the TG132 report.
The most commonly used quantitative methods for DIR error assessment define the target registration error based on landmark comparisons and contour evaluation metrics such as the Dice similarity coefficient and mean distance to agreement (Castillo et al 2009, Brock 2010, Brock et al 2017). These metrics, however, require manually generated contours or landmark definitions and are only accurately defined at high image contrast regions. Moreover, these methods are prone to human observer variability, which may negatively affect the performance estimate of the DIR (Mencarelli et al 2012). Although these methods are suitable for the commissioning of a DIR algorithm, these metrics are too time-consuming for patient-specific DIR evaluation in routine clinical practice.

Other metrics of quantitative DIR evaluation are based on mathematical analysis of the transformation matrix and can be grouped into metrics that account for anatomical or physiologically realistic transformations and numerically robust solutions of the DIR. The former use biomechanical tissue properties and are generally directly derived from the deformation field. An often-used entity to evaluate the DIR for unrealistic tissue folding is the Jacobian determinant of the deformation field (Chen et al 2008, Vercauteren et al 2009, 2013). The DIR error has also been analyzed by the energy state of the deformation grid (Vercauteren et al 2013, Veiga et al 2014, Weistrand and Svensson 2015, Vercauteren et al 2009, Zhong et al 2007). Examples include the harmonic energy (HE) (Vercauteren et al 2009), unbalanced energy (Zhong et al 2007), or Dirichlet energy (Weistrand and Svensson 2015), and all account for the local smoothness of the deformation field. These metrics have also been used as regularization terms in the DIR optimizer.

The numerical stability of the DIR can be derived using functional compositions of registration circuits. In a scenario without DIR uncertainties, the composition of each transformation within a circuit produces the same deformed image as the reference image. Any DIR uncertainties are then illustrated by increased error values. The most simple circuit assesses the property of invertibility of the registration and was previously introduced as the inverse consistency error (ICE), which uses a functional composition of the forward and backward registration between two images (Christensen and Johnson 2001, Chen et al 2008, Yang et al 2008, Bender et al 2009). Increased ICE values indicate inverse inconsistent registrations that can occur when tissue is only present in one of the images. The generalized ICE was introduced as the first circuit including >2 images and referred to as the transitivity error (TE) (Christensen and Johnson 2003, Bender et al 2009). More recently, the AQUIRC (Datteri et al 2015) and distance discordance metric (DDM) (Saleh et al 2014, 2016) were introduced comprising multiple registration loops within the circuit. The DDM requires mutual registrations between at least four image sets and reports the DIR error as the mean distance between corresponding voxels when
registered to a reference image (Saleh et al. 2014). They reported a correlation of $r = 0.68$ between the mean DDM and the true registration error as defined by contour volume ratios from an inter-patient DIR of computed tomography (CT) scans of head and neck (HN) cancer patients (Saleh et al. 2014). In a more recent study by Saleh et al., the intra-patient DIR was studied in pelvic organs for a series of prostate cancer patients (Saleh et al. 2016). The DDM values ranged from 1.0 – 13.0 mm for bladder and rectum, with considerable DIR variability across subjects. So far, however, DDM values for intra-patient DIR of HN and lung cases have not been investigated.

Although various quantitative DIR error metrics have been proposed and tested in several anatomical sites, publications describing deformable registrations generally limit the DIR evaluation to a few metrics. To the best of our knowledge, a comprehensive evaluation of all relevant quantitative DIR error metrics has not been performed in CT scans of HN and lung cancer patients. In this work, we evaluate DIR accuracy using complementary measures that account for physiologically realistic registrations and numerical robust DIR solutions. The former includes the determinant of the Jacobian, HE, and the octahedral shear strain (OSS). The latter includes the ICE, TE, and DDM. These DIR error metrics have been derived from intra-patient deformable registrations of 26 HN cancer patients and 12 lung cancer patients. The goal of this study was to develop a method based on the most predictive DIR error metrics to identify voxels with a DIR error tolerance of 2.0 mm.

### 6.2 Methods

#### 6.2.1 Patients and imaging data

This study included CT scans of 26 patients with HN cancer and 12 patients with lung cancer. HN cancer patients were diagnosed with stage II-IV HN cancer and were treated with curative intent. A reference CT $I_{ref}(x)$ was acquired before treatment and repeated CT imaging was performed on a weekly basis (7 weeks) during the course of radiotherapy. Patients were immobilized during imaging and radiotherapy with a supine headrest and a 5-point head-shoulder mask. The reference CT scans were performed with an intravenous iodinated contrast agent. The CT scans ranged from the base of the skull to the superior part of the lungs. The following reconstruction settings were used: a slice thickness of 2.0 mm, a field of view of 500 mm, and an image size of 512 x 512 pixels.

The lung cancer patients were diagnosed with stage III non-small-cell lung cancer and treated with curative intent. During imaging and radiotherapy, the patients were positioned on a lung-board with arms supported above the head. A 4D-CT scan was
acquired from which ten breathing phases were reconstructed for estimation of target motion. The image reconstruction parameters were similar to those of the HN scans except a slice thickness of 3.0 mm.

6.2.2 Deformable image registration
Each image set was rigidly aligned with the \( I_{\text{ref}}(x) \) followed by a deformable registration using a cubic B-spline transformation model as implemented in the elastix toolbox (version 4.8) (Klein et al 2010). The deformable registration algorithm used a 3-step multi-resolution procedure and an advanced Mattes mutual information image similarity metric. The number of grey level histogram bins in each resolution level was 32, the B-spline grid spacing was 5.0 mm, and number of optimization iterations was 500. To avoid unrealistic deformations, a regularization penalty term on the bending energy of the deformation field was applied (relative weight = 0.05). All deformable registrations were performed with a fixed image mask consisting of the patient body contour plus 10 mm.

6.2.3 Synthetic CT with known deformation
To calculate the DIR error per voxel, for each image set, a synthetic CT image \( I_s(x) \) with a known deformation with respect to the reference CT was created from a deformed reference CT \( I_{\text{ref}}(x) \) using a known “ground-truth” deformation field \( T(x)_{GT} \) (in x, y, z). To obtain a realistic ground-truth deformation field, \( T(x)_{GT} \) was derived from a deformable registration between the reference CT \( I_{\text{ref}}(x) \) and another CT of the same patient \( I_F(x) \), using an independent deformation algorithm DIR\(_{GT}\). For the HN cases, the \( I_F(x) \) was a repeat CT acquired in week 6 to include substantial deviation from the reference image. For the lung cases, the 0% and 40% reconstructions of a 4D dataset were used as \( I_{\text{ref}}(x) \) and \( I_F(x) \), respectively. It is expected that the B-spline algorithm will overestimate the performance of synthetic images generated with a B-spline algorithm. Therefore, an independent and commercially available free-form deformable algorithm (“super-fine”, RTx v1.6, Mirada Medical Ltd. Oxford, UK) was used as DIR\(_{GT}\). The “super-fine” algorithm uses an adaptive multi-resolution registration scheme with tri-linear interpolation where the deformation field is at lower resolution. Hence, for generalizability and to minimize the bias towards the use of the “super-fine” algorithm, we convolved the \( T(x)_{GT} \) with a symmetrical box shaped kernel of 5.0 mm, which resulted in lower spatial frequencies in the ground-truth deformation field.

6.2.4 Ground truth deformation error
The synthetic CT image \( I_s(x) \) was deformably registered with the reference image \( I_{\text{ref}}(x) \) using the B-spline DIR algorithm (DIR\(_{BS}\)), resulting in transformation \( T(x)_{BS} \). The known deformation vector field \( T(x)_{GT} \) was used to calculate the Euclidean length of the ground-truth deformation error as:
\( \varepsilon(x)_{\text{GT}} = \| T(x)_{\text{GT}} - T(x)_{\text{BS}} \| \)

which was used to benchmark the DIR accuracy metrics. The deformable registration evaluation workflow is described in figure 6.1.

**Figure 6.1. Flow chart of the deformable registration evaluation process.**

The floating image \( I(x) \), reference image \( I_{\text{ref}}(x) \), and the “super-fine” deformable registration \( \text{DIR}_{\text{GT}} \) were used to create the synthetic image \( I_{\text{syn}}(x) \). The image stack \( I_{x1..n}(x) \) represents the images required to calculate the transitivity error (TE) and distance discordance metric (DDM). Abbreviations: ICE = inverse consistency error; HE = harmonic energy; OSS = octahedral shear strain; JAC = determinant of Jacobian.

### 6.2.5 DIR accuracy metrics

**Physiologically realistic transformations**

The main goal of deformable registrations is to align two images on a voxel level, preferably with preservation of topology. The latter can be achieved using a regularization of the deformation field, for example by minimizing an energy term of the deformation field (Weistrand and Svensson 2015, Vercauteren et al 2009, Zhong et al 2007). Another previously described entity regarding physically realistic deformations is the Jacobian determinant of the deformation field. The Jacobian accounts for tissue expansion (and shrinkage) and is defined as the determinant of the deformation gradient (Ashburner 2007, Christensen and Johnson 2001, Yang et al 2008). When incorporated into the optimizer, the Jacobian determinant prevents the deformation field from unrealistic folding which results in invertible deformation fields (Vercauteren et al 2013). Another metric, which has been introduced is the harmonic energy, which is inversely proportional to the smoothness of the deformation field (Vercauteren et al 2009, Varadhan et al 2013). A third measure, which accounts for the biomechanical properties of tissue is the shear strain (Christen et al 2012). The shear strain provides additional information to what extend a deformation was realistic in different tissue types (i.e. bone or soft tissue) and was derived using the expression of the OSS (Christen et al 2012, McGarry et al 2011). In this paper we determined the determinant of the Jacobian, HE and OSS to investigate whether the DIR transformation was physiological realistic. These measures were derived from \( T(x)_{\text{BS}} \).
Numerically Robust Transformations

This paragraph describes the DIR error metrics that we used to quantify whether a deformable registration results into a numerically robust transformation. If needed, transformations were combined using so-called functional compositions, in which the resulting transformation of one DIR is applied as input to the following DIR. The ICE was then derived as the Euclidean length of the functional composition of the forward and reverse transformations. The TE, on the other hand, required a registration circuit using more than two image sets (Bender et al. 2009). In this study we calculated the TE by using DIR\textsubscript{BS} on three images. For the HN cases, a repeat CT acquired during the third week of treatment was used (additional to \( I_{ref}(x) \) and \( I_s(x) \)). For the lung cases, another breathing phase of the 4D CT images was used.

The DDM requires registrations between at least four image sets and uses the variability in the distance between corresponding voxels when registered to a reference image (Saleh et al. 2014). The DDM value then displays the mean dispersion between these voxels. Accurate registrations with a small distance between the voxels then correspond to a low DDM value, whereas a poor registration with a large distance gives a larger DDM value. It was stated that the DDM performance improves with number of image sets included (Saleh et al. 2014). This is, however, at the cost of intensive computational power, as it requires \((n – 1)!\) registrations with \(n\) the number of image sets. Therefore, we calculated the DDM using DIR\textsubscript{BS} on five image sets. In addition to the images used for the TE calculations, a repeat CT acquired in week 5 and 7 during treatment (HN cases) or two more breathing phases (lung cases) were included. In contrast to the ICE and TE, the calculation of the DDM also requires the inverse transformation. Since the exact inverse of a deformable registration is non-existing, an approximate inverse transformation was derived by minimizing \(\|T(x) – x\|^2\) at each location \(x\). An overview of the DIR error metrics that were investigated in this paper is provided in table 6.1.

6.2.6 Analysis

For each registration the DIR error metrics were calculated on a voxel level and the correlation with the \(\varepsilon_{GT}\) was determined using Pearson’s correlation coefficient. Linear regression analysis was performed to find a relationship between the DIR error measures and the \(\varepsilon_{GT}\) for those measures that showed the largest correlation with the \(\varepsilon_{GT}\). Since linear regression analysis requires a constant variance over the prediction values and a relatively large percentage of the \(\varepsilon_{GT}\) values ranged from 0.0 – 1.0 mm, the predictors and response variable (i.e. the \(\varepsilon_{GT}\)) were log-transformed before analysis. The regression analysis was performed with 100\(m\) independent and randomly selected samples, where \(m\) is the number of patients. The sampling and modeling procedure was repeated 100 times, of which the average regression coefficients were determined. The linear models were
then validated using a 10-fold cross-validation of which the root-mean-square (RMS) error was defined. For each fold, different patients \( m \) were selected for training and testing the model. The performance of the selected DIR error metrics was further assessed by the sensitivity, specificity, positive- and negative predictive value and the area under the curve (AUC) at a 2.0 mm threshold for \( \varepsilon_{GT} \). This value was considered as clinically relevant given a typical radiotherapy dose grid size of \((1.0 - 3.0)^3 \) mm\(^3\) (see also recommendations in the AAPM TG-132 report: for 95% of the volume DIR errors should be within 2.0 mm) (Brock et al. 2017). Threshold values for the selected DIR error metrics were proposed to identify voxels with a DIR error \( > 2.0 \) mm with a negative predictive value of 90% and 95% (i.e. the ratio of voxels correctly identified as DIR error \( \leq 2.0 \) mm). Note that the voxels outside the body contour were omitted from the analyses.

<table>
<thead>
<tr>
<th>Table 6.1. Overview of fully spatial DIR error metrics.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Physiologically realistic DIR</strong></td>
</tr>
<tr>
<td>Jacobian determinant (JAC)</td>
</tr>
<tr>
<td>Measure of unrealistic folding of the deformation field</td>
</tr>
<tr>
<td>Derived from the deformation gradient</td>
</tr>
<tr>
<td>Harmonic Energy (HE)</td>
</tr>
<tr>
<td>Measure of transformation regularity</td>
</tr>
<tr>
<td>Derived from the deformation gradient</td>
</tr>
<tr>
<td>Octahedral Shear strain (OSS)</td>
</tr>
<tr>
<td>Measure of tissue shearing (different for bone and soft tissues)</td>
</tr>
<tr>
<td>Derived from the deformation gradient</td>
</tr>
<tr>
<td><strong>Numerically robust DIR</strong></td>
</tr>
<tr>
<td>Inverse consistency error (ICE)</td>
</tr>
<tr>
<td>Forward and backward registration between 2 image sets required</td>
</tr>
<tr>
<td>Identifies the invertibility of the deformation field.</td>
</tr>
<tr>
<td>Transitivity error (TE)</td>
</tr>
<tr>
<td>Circuit registration with single loop</td>
</tr>
<tr>
<td>Registrations between at least 3 image sets required</td>
</tr>
<tr>
<td>Distance discordance metric (DDM)</td>
</tr>
<tr>
<td>Circuit registration comprising multiple loops</td>
</tr>
<tr>
<td>Registrations of at least 4 image sets required</td>
</tr>
<tr>
<td>Calculation time with current implementation is time-consuming</td>
</tr>
<tr>
<td>(between 70 – 135 min)</td>
</tr>
</tbody>
</table>

6.3 Results

Deformable registration errors \( \varepsilon_{GT} \) up to 10 mm were observed in the HN and lung cases. A cross-section of the log-transformed \( \varepsilon_{GT} \) and DIR error metrics of a representative HN and lung case are shown in figure 6.2 and figure 6.3, respectively. On average (\( \pm \)sd), the percentage of voxels with an \( \varepsilon_{GT} > 2.0 \) mm was 20.6 \( \pm \) 10.8% for the HN cases. The most pronounced registration errors were observed within the oral cavity and the neck region. The average percentage of voxels with an \( \varepsilon_{GT} > 2.0 \) mm in the lung cases was 21.5 \( \pm \) 13.6%. The largest \( \varepsilon_{GT} \) values were observed near the chest-wall and the diaphragm.
The distribution of the $\varepsilon_{GT}$ and the DIR error metrics are shown in figure 6.4. All DIR error measures, except the ICE, were larger for the lung cases.

The DIR error metrics within the brain did not agree with the $\varepsilon_{GT}$ (all metrics showed a correlation of $r < 0.25$). Therefore, the following HN results excluded the brain volume. The relationship between the DIR error metrics and the $\varepsilon_{GT}$ is depicted by scatter plots (figure 6.5). Generally, the correlation for the HN cases was lower than for the lung cases (see heat maps in figure 6.6, excluding the brain volume). For the HN, the highest correlation was found between the DDM and $\varepsilon_{GT}$ ($r = 0.50$) followed by the HE ($r = 0.44$), OSS ($r = 0.41$), and the ICE/TE ($r = 0.39$). Similarly for the lung cases, the highest correlation with the $\varepsilon_{GT}$ was found for the DDM ($r = 0.56$) followed by the ICE ($r = 0.51$), TE ($r = 0.50$), HE ($r = 0.49$) and the OSS ($r = 0.45$). Overall, the ICE, TE, and DDM underestimated the absolute registration error. Figure 6.5(g) shows a strong correlation ($r = 0.90$) between the HE and OSS values, which was expected due to the nature of the HE and OSS formulations.

Furthermore, a moderate correlation was found between the DDM and HE (HN: $r = 0.52$; Lung: $r = 0.60$) and the DDM and OSS (HN: $r = 0.47$; Lung: $r = 0.54$) as shown in figure 6.5(h-i). It was observed by visual inspection that the HE was not in agreement with $\varepsilon_{GT}$ at large errors near the diaphragm. The DDM was a better predictor in that region. The Jacobian determinant was positive (distributed around a value of 1.0) in all cases, indicating no folding of the deformation field (figure 6.4 and figure 6.5(d)). Furthermore, a poor correlation with the $\varepsilon_{GT}$ was found.
Figure 6.3. A coronal cross-section of a CT scan of a representative lung cancer patient with overlays of the log-transformed ground-truth error map, the deformation field and the log-transformed (except the det. Jacobian) DIR error metrics.

Abbreviations as in figure 6.2.

Table 6.2. Performance of DIR uncertainty metrics at 2.0 mm tolerance threshold.

<table>
<thead>
<tr>
<th></th>
<th>DDM</th>
<th>HE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative predictive value</td>
<td>0.95</td>
<td>0.90</td>
</tr>
<tr>
<td>Positive predictive value</td>
<td>0.28</td>
<td>0.37</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.94</td>
<td>0.68</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.33</td>
<td>0.71</td>
</tr>
<tr>
<td>AUC</td>
<td>0.75</td>
<td>0.77</td>
</tr>
<tr>
<td>Threshold [mm]</td>
<td>0.16</td>
<td>0.49</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.33</td>
<td>1.60</td>
</tr>
<tr>
<td>Gradient</td>
<td>0.48</td>
<td>0.43</td>
</tr>
<tr>
<td>RMS Error [mm]</td>
<td>2.34</td>
<td>2.42</td>
</tr>
</tbody>
</table>
Table 6.2 shows the performance of the DDM and HE at a DIR tolerance threshold of 2.0 mm. The corresponding DDM and HE thresholds were based on a negative predictive value of 0.95 and 0.90. A lower negative predictive value resulted into a higher specificity at the cost of falsely identified negatives (i.e. voxels with $\varepsilon_{GT} > 2.0$ mm that were not identified by the DDM and HE thresholds). Thresholds of DDM = 0.49 and HE = 0.014 were found under the assumption that it is acceptable that a maximum of 10% of the voxels that were identified as having a DIR error $\leq 2.0$ mm appeared to have an $\varepsilon_{GT} > 2.0$ mm (negative predictive value = 0.90; vertical line in figure 6.7). The AUC and RMS error for the DDM and HE were comparable. We further found that a combination of DDM and HE or any other studied metrics did not improve the DIR uncertainty estimates.
Figure 6.5. Scatter plots of the ground-truth registration error $\varepsilon_{\text{GT}}$ with the deformable registration error metrics (data log-transformed except Jacobian determinant) of 200 randomly sampled voxels of (circles) 26 head and neck cases and (squares) 12 lung cases (A-F). The relationship between the distance discordance (DDM), harmonic energy (HE), and octahedral shear strain (OSS) is depicted in subplots G–I. The solid (head and neck) and dashed (lung) lines indicate the linear regression for each situation.
Figure 6.6. Heat maps of the Pearson correlation coefficients of the deformable registration uncertainty metrics and ground-truth registration error $\varepsilon_{GT}$ derived from the head and neck (A) and lung data (B).

Abbreviations: ICE = inverse consistency error; TE = transitivity error; DDM = distance discordance metric; HE = harmonic energy; OSS = octahedral shear strain; JAC = determinant of Jacobian.

Figure 6.7. Scatter plot of the distance discordance metric (DDM) [A] and harmonic energy (HE) [B] against the ground-truth error $\varepsilon_{GT}$ of the combined datasets of the head and neck and lung cases.

The solid line indicates the regression line. The horizontal dashed line is the deformable registration tolerance threshold of 2.0 mm. The vertical line is the threshold at a negative predictive value of 0.90. Note the logarithmic scales.
6.4 Discussion

This work evaluates several DIR error metrics that account for physiologically realistic and numerically robust DIR in CT image sets of HN and lung cancer patients. Among all cases, the DDM and HE showed the highest correlations with the ground-truth registration error. The HE can directly be derived from the deformation field whereas the DDM requires functional compositions of multiple and time-consuming registrations. It was further found that a combination of the DDM and HE model did not improve the DIR error estimates, which can be explained by a moderate correlation between the DDM and HE.

The deformation fields of the HN and lung cases are generally characterized by a certain level of smoothness (also due to the rigidity term in the regularization). Since the HE is inversely proportional to the smoothness of the deformation field the relatively high performance of the HE as a measure of DIR accuracy was therefore expected. It is however unknown whether the HE is applicable on deformation fields derived from image sets with significant volume changes (i.e. loss of topology). In this study, the HE was only evaluated in HN and lung cases. However, its performance in other target areas is unknown.

Our results show that the DIR error metrics did not correlate well in the brain. We observed that the B-spline DIR and DIR\textsubscript{GT} resulted in different deformation fields, which was likely caused by limited contrast in that region. The DIR error metrics relied on B-spline transformations only and therefore did not capture the ground-truth error. Further analysis of DIR performance in the brain is therefore required.

Since it is recommended to review all deformable registrations for the purpose of dose mapping, and also from a practical point of view, the DIR analysis should be fully automated. Predefined thresholds for the DIR evaluation metrics are therefore required. In this study, the performance of the DDM and HE was based on a DIR tolerance threshold of 2.0 mm [according to the TG132 report (Brock \textit{et al} 2017)], on a ground-truth error, and a negative predictive value of 0.90 or 0.95. We found that a substantial amount of voxels of the DDM and HE-maps were incorrectly flagged as error (false positives). An increased rate of false positives directly translates into an increased workload since DIR errors in regions of a high dose should be corrected using e.g. a contour-guided DIR. An acceptable proportion of false positives should therefore be defined. The threshold values of DDM = 0.49 and HE = 0.014 with a negative predictive value of 0.90 and a specificity of 0.71 (DDM) and 0.52 (HE) seem to be the most optimal choice.
We found that the ICE and the TE were less suitable to detect DIR errors than the DDM. These results were similar to those reported by Saleh and colleagues (Saleh et al. 2014). They introduced and compared the DDM with the ICE and TE using checkerboard phantoms with simulated deformations and clinical CT images of HN and prostate cancer patients. In the HN cases, they derived the DDM from an inter-patient registration setting (Saleh et al. 2014). Since multiple images are available for patients receiving adaptive radiotherapy, we evaluated the DDM on an intra-patient registration basis. As expected, the DDM values from our intra-patient registrations were substantially lower than the DDM values from the inter-patient registrations reported by Saleh et al.

In contrast to our study, Saleh et al. evaluated the DDM against volume ratio measures derived from contours, whereas our findings were based on synthetic images with a realistic known deformation and corresponding ground-truth deformation fields (Saleh et al. 2014). Similarly, Varadhan et al. used simulated deformations to evaluate the DIR using the Jacobian, HE and ICE (Varadhan et al. 2013). Similar to our findings, the ICE for simulated lung cases was lower than in HN cases. HN images were acquired over the 7 weeks treatment course and relative large anatomical changes were seen between images (e.g. tumor regression and/or weight loss). This may result in transformations that are not invertible. In 4D images of the lung cases, however, the anatomy is generally preserved between the images, leading to more invertible DIR and therefore lower ICE values.

Functional compositions of multiple registrations performed in a loop, like the DDM and the TE are increasingly investigated. A common property is that all DIR error estimates can be determined automatically, on a voxel level, even in image regions with little contrast, and for a case-specific image set. One approach is based on registrations between test image sets derived from different DIR algorithms (Kirby et al. 2016). From the different registrations, the spatial and dose uncertainty estimates were then based on a Student’s t distribution. One should notice that this approach requires multiple simulated deformations and known ground-truth deformation fields. Another approach uses multiple combinations of the TE and was presented as the AQUIRC method (Datteri et al. 2015). These methods show, similar to the DDM that DIR uncertainties are more likely to occur at low contrast regions. Moreover, the calculation of these metrics is computationally intensive and time-consuming. The calculation of the DDM with five image sets of a representative HN and lung case took approximately 70 – 135 min on an Intel Xeon 2.10 GHz 32 GB computer. The use of parallel computing can substantially reduce this calculation time.

The registration error is most commonly assessed by landmark comparisons. However, definition of landmarks is time-consuming and prone to human observer variability (Mencarelli et al. 2012). Moreover, landmarks can only be accurately defined at high
contrast regions. However, with regard to dose mapping, the highest DIR accuracy and precision is required within and especially near the tumor (i.e. within the dose gradient). It was, however, demonstrated by Mencarelli et al. that, due to limited contrast, B-spline-based DIR was less precise in tumor tissue than in normal tissues (Mencarelli et al 2014). Fully spatial DIR error metrics that also assess the registration quality in low contrast regions, such as the DDM and HE, are therefore preferred.

Due to the limitations of the landmark-based evaluation, we evaluated the DIR error metrics against a ground-truth registration error derived from a synthetic CT image. We acknowledge that this “ground-truth” deformation field is not completely equal to a real physiological deformation. However, the synthetic image $I_s(x)$ closely resembled the additional image (floating image in figure 6.1), indicating that the resulting deformation field was close to clinical reality. Another limitation is the use of one DIR system (DIR$_{GT}$) to generate the transformation $T(x)_{GT}$. Therefore, and to generalize our findings, we smoothed the $T(x)_{GT}$ from which the synthetic image was created. The effect of the convolution kernel on the results was not part of this study. Future work will focus on the application of the DDM and HE for different DIR algorithms. Moreover, the effect of the DIR error on the dose distribution will be investigated.

In conclusion, several DIR error metrics that account for DIR robustness and physiologically realistic deformation fields have been implemented and evaluated in CT images of HN and lung cancer patients. The log-transformed DDM and HE show the highest correlation with the simulated ground-truth error. The DDM and HE, with thresholds of 0.49 mm and 0.014 respectively, can be used in an automated procedure to identify voxels with DIR errors $> 2.0$ mm with a specificity of 0.71 (DDM) and 0.52 (HE) and a negative predictive value of 0.90.
6.5 References


Ashburner J 2007 A fast diffeomorphic image registration algorithm Neuroimage 38 95–113


Weistrand O and Svensson S 2015 The ANACONDA algorithm for deformable image registration in radiotherapy Med. Phys. 42 40–53