RegQCNET: Deep Quality Control for Image-to-template Brain MRI Affine Registration

Baudouin DENIS de SENNEVILLE[1], **José V. MANJÓN**[2], **Pierrick COUPÉ**[3]

[1] CNRS, University of Bordeaux, "Institut de Mathématiques de Bordeaux" (IMB), UMR5251, F-33400 Talence, France

[2] ITACA, Universitat Politècnica de València, Camino de Vera, s/n, 46022, Valencia, Spain

[3] CNRS, University of Bordeaux, Bordeaux INP, "Laboratoire Bordelais de la Recherche Informatique" (LaBRI), UMR5800, F-33400 Talence, France

E-mail: bdenisde@math.u-bordeaux.fr, jmanjon@fis.upv.es, pierrick.coupe@u-bordeaux.fr

July 2020

Abstract. Affine registration of one or several brain image(s) onto a common reference space is a necessary prerequisite for many image processing tasks, such as brain segmentation or functional analysis. Manual assessment of registration quality is a tedious and time-consuming task, especially in studies comprising a large amount of data. An automated and reliable quality control (QC) becomes mandatory. Moreover, the computation time of the QC must be also compatible with the processing of massive datasets. Therefore, an automated deep neural network approaches appear as a method of choice to automatically assess registration quality.

In the current study, a compact 3D convolutional neural network (CNN), referred to as RegQCNET, is introduced to quantitatively predict the amplitude of an affine registration mismatch between a registered image and a reference template. This quantitative estimation of registration error is expressed using metric unit system. Therefore, a meaningful task-specific threshold can be manually or automatically defined in order to distinguish usable and non-usable images.

The robustness of the proposed RegQCNET is first analyzed on lifespan brain images undergoing various simulated spatial transformations and intensity variations between training and testing. Secondly, the potential of RegQCNET to classify images as usable or non-usable is evaluated using both manual and automatic thresholds. During our experiments, automatic thresholds are estimated using several computer-assisted classification models (logistic regression, support vector machine, naïve bayes and random forest) through cross-validation. To this end we used expert's visual quality control estimated on a lifespan cohort of 3953 brains. Finally, the RegQCNET accuracy is compared to usual image features such as image correlation coefficient and mutual information.

Results show that the proposed deep learning QC is robust, fast and accurate to estimate affine registration error in processing pipeline.

Keywords: Quality Control, Image-to-template registration, Deep Neural Network.

1. Introduction

A wide variety of processing pipelines have been proposed in the literature to make automatic brain image analysis possible. Spatial and intensity normalizations are usually necessary prerequisites for functional (Cook et al. 2006) (Song et al. 2011) or structural studies (Wei et al. 2002) (Chaogan & Yufeng 2010). These are commonly achieved using suitable algorithms designed for image-to-template registration (Tustison et al. 2014) (Jenkinson et al. 2012) (Collins et al. 1994), inhomogeneity correction (Tustison et al. 2010) (Sled et al. 1998), or intensity normalization (Nyúl et al. 2000) (Friston et al. 1995). A visual human inspection of the data after each step of the processing pipeline is commonly employed to detect possible problems in the outputs. This visual quality control (QC) is unfortunately not feasible when a huge amount of imaging data is involved (typically more than several thousands scans). Consequently, with the rise of large-scale datasets, recent efforts are dedicated to the development of reliable QC methods to detect pipeline failures (Alfaro-Almagro et al. 2018) (Kim et al. 2019). Although out of the scope of the current study, an increasing interest in registration QC is noticed in radiation therapy (Brock et al. 2017) (Paganelli et al. 2018).

Fig. 1 summarizes the context of the current paper. Our study focuses on the image registration step, this step is a necessary prerequisite to co-register one or several brain scans onto a common space defined by a template image. In practice, mis-registered data are inherently encountered and thus a registration QC is needed (red box in Fig. 1) (Avants et al. 2011). A manual assessment is generally employed for QC and thus automatic methods have been developed to achieve this task. However, such manual strategy is time consuming. Random forest (Hessam et al. 2019) and convolutional neural network (CNN) (Eppenhof & Pluim 2018) have been proposed to quantify registration accuracy for both parametric (*i.e.*, rigid, affine) and deformable registrations in chest CT scans. In the context of neuroimgaging, several methods have been proposed for MRI brain registration to template space. In (Fonov et al. 2018) a cross-entropy loss function is used as an objective function to train a deep neural network on a serie of 2D control images. This method produces qualitative estimation (*i.e.*, good or not good) of rigid registration accuracy. In (Bannister et al. 2019) and (Dubost et al. 2019), a DICE metric between transformed and original organ contours is proposed as a surrogate of registration quality. The use of an indirect metric (*i.e.*, DICE) estimated on an auxiliary task (*i.e.*, segmentation) does not provide direct quantitative information on registration accuracy. Moreover, this information can be corrupted by segmentation error that is a complex task by itself. Finally, these three methods produce metrics (binary decision or auxiliary DICE) that cannot express registration error in metric unit system. Consequently, no meaningful task-specific threshold on the misalignment amplitude can be defined by a user.

Our contribution is four-fold:

(i) A compact 3D CNN is introduced to quantitatively estimate the quality of an affine alignment between a brain MRI and a template. The proposed QC network, referred

to as RegQCNET, is quantitative and can be expressed using metric system units. Moreover, an efficient and robust training procedure, based on simulated affine transformations, is proposed. The inputs of the designed CNN are: the registered image and the reference template. Besides, it is demonstrated that RegQCNET meets computational requirements related to massive processing.

(ii) The robustness of the proposed RegQCNET is analyzed on lifespan brain images undergoing various simulated spatial transformations and intensity variations between training and testing.

(iii) The potential of RegQCNET to classify images as usable or non-usable is evaluated using both manual and automatic thresholds. Automatic thresholding is evaluated using several computer-assisted classification models (logistic regression, support vector machine, naïve bayes and random forest) through a cross-validation procedure. To this end we used expert's visual quality control estimated on a lifespan cohort of 3953 brains as a gold standard.

(iv) The RegQCNET accuracy is compared to usual image features such as image correlation coefficient and mutual information.

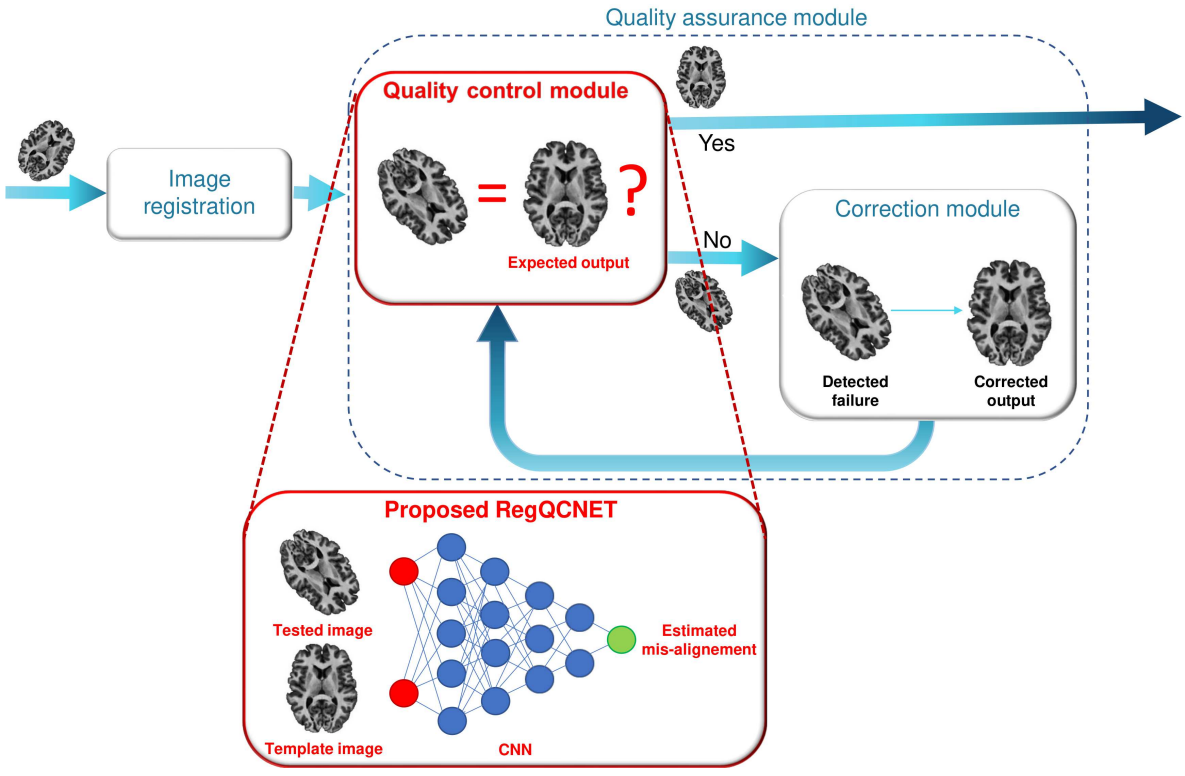

Figure 1: General principle of the proposed Quality Control (QC) on image-to-template registration. The current study aims at providing a QC module (red bock) designed to detect misaligned images. Note that this module can potentially be fed into an additional correction module (outside the scope of the current study).

2. Materials and Methods

2.1. Datasets

Figure 2 details the processing sequence designed to generate the datasets involved in our experiments. Throughout this study, we used 3 datasets 1 for training and 2 for testing referred to as "Simulated training dataset", "Simulated testing dataset" (these last two were built using a lifespan dataset, for which synthetic affine transformations were applied) and "Real testing dataset". First, all the native images have been downloaded from public databases. While considered as the native images, these images may contain specific preprocessing (i.e. NDAR includes defacing). Afterwards, all these images went through our preprocessing pipeline as described in the following.

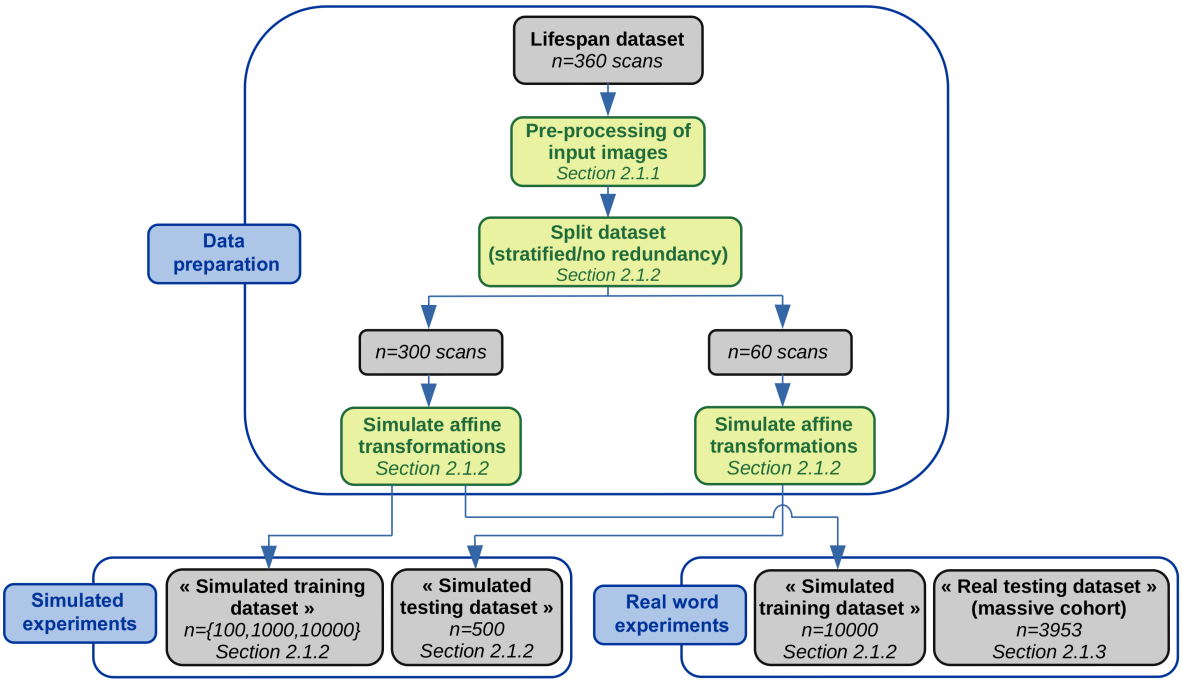

Figure 2: Processing sequence designed to generate datasets involved in experiments. Each generated dataset is displayed as a gray block. Image processing tasks are reported with green blocks.

2.1.1. Preprocessing of input images. To ensure spatial and intensity normalization between images, we used a preprocessing pipeline. Consequently, all the scans involved in this study were first preprocessed beforehand using the volBrain pipeline (Manjón & Coupé 2016). This pipeline is based on the following steps: i) denoising (Manjón et al. 2010), ii) inhomogeneity correction (Tustison et al. 2010), iii) affine registration into the template space ($181\times217\times181$ voxels at $1\times1\times1$ mm^3, the ICBM 152 Atlas template was taken as reference for registration (Fonov et al. 2011)), iv) manual human assessment of the registration as described in (Coupé et al. 2017) (Coupé et al. 2019), and

v) tissue-based intensity normalization (Manjón et al. 2008). Finally, image intensities were normalized using z-scoring using the mean and standard deviation from the complete image field-of-view.

2.1.2. Simulated training and testing datasets. 360 T1-weighted MRI of cognitively normal subjects were randomly selected under constraints from the dataset used in our previous BigData study on normal aging (Coupé et al. 2017). This dataset was based on 9 datasets publicly available (C-MIND, NDAR, ABIDE, ICBM, IXI, OASIS, AIBL, ADNI1 and ADNI2). From 1 to 90 years we selected 2 females (F) and 2 males (M) for each age (*i.e.*, 2F and 2M of 1 year old, 2F and 2M of 2 years old and so on). Therefore, we obtained a balanced group with 50% of each gender uniformly distributed from 1 to 90 years. This balanced selection is done to limit bias introduction in training and testing datasets and to make our QC method robust to different age and gender.

All the 360 MRIs underwent a human quality control. Consequently, all these images were considered as correctly aligned with negligible residual registration mismatch. The RMSE was thus considered to be equal to 0. We are aware that these images are not perfectly aligned and that negligible errors might remain. Considering these remaining errors equal to epsilon or zero do not impact the rest of the study.

At this point, we have a set of 360 scans including spatial and intensity normalizations. This set was split into two separate datasets (stratified/no redundant data): 300 scans were used to build the "Simulated training dataset" and 60 scans were used to build the "Simulated testing dataset". This split was done under contraint to ensure well-balance of age and gender between both.

The "Simulated training dataset". One can expect that the training set has to be populated densely enough in terms of both anatomical inter-subject variability and simulated spatial deformations. During our experiments, RegQCNET was trained using N simulated scans (randomly selected with replacement from the 300 above-mentioned scans). These scans were simulated using affine spatial transformations. In order not to favor large or small transformations, we force the RMSE distribution resulting from the simulated transformation to be uniform. We tested N-values in the following set: $\{100, 1000, 10000\}$. Note that the obtained training set may include several transformations for each patient when $N > 300$.

The "Simulated testing dataset". This set was composed of 500 scans (randomly selected with replacement from the 60 subjects selected to build the testing data). In this way, several transformations were applied to each patient (≈ 8 in average).

Simulated spatial affine transformations. To train and test the proposed method, 3D spatial transformations — composed by translation, rotation and scale variations — were simulated. Ranges for X-, Y-, Z-translations, rotations around X-, Y-, Z-axis, X-, Y-, Z-scaling factors are detailed in the experimental section below. The RMSE was

calculated for each simulated transformation. Let `MAX_RMSE` be the upper limit of the simulated RMSE. A set of spatial transformations with a uniform RMSE distribution in the interval $[0, \texttt{MAX\_RMSE}]$ voxels was built. To this end, 3D spatial transformations composed by translation, rotation and scale variation, were simulated as follows:

- X-, Y-, Z-translations were randomly selected separately in the interval $[-100, 100]$ voxels,
- Rotations around X-, Y-, Z-axis were randomly selected in the interval $[-45, 45]$ degrees,
- X-, Y-, Z-scaling factors were randomly selected in the interval $[0.5, 1.5]$ (a factor of 1 being equivalent to no scaling),

These transformations were then applied using b-spline interpolation to the images (see Fig. 4).

2.1.3. The "Real testing dataset". The performance of the proposed RegQCNET was evaluated on a massive database ($N = 3953$) including cognitively normal patients, patients with Alzheimers Disease (AD) and Mild Cognitive Impairment (MCI). These 3953 MRIs were the remaining subjects from the large-scale cohort used in (Coupé et al. 2019) after removal of the 360 cognitively normal subjects used to build the simulated training and testing dataset. Consequently, the real testing dataset contained pathological alterations unseen in the training dataset. A visual assessment was done by checking screen shots of one sagittal, one coronal and one axial slice in middle of the 3D volume using the volBrain reports (Manjón & Coupé 2016). Therefore, a human-based QC was available for all the scans and was used as qualitative ground truth.

2.2. Proposed RegQCNET

2.2.1. Implemented quantitative metric. In this study, we aim at quantifying the residual misalignment (noted T) between two given images via the Root Mean Square Error criterion (RMSE) computed as follows (Maurer et al. 1997):

$$\text{RMSE} = \frac{1}{|\Omega|} \sum_{\vec{r} \in \Omega} \sqrt{u(\vec{r})^2 + v(\vec{r})^2 + w(\vec{r})^2} \tag{1}$$

$\vec{r} = (x, y, z)$ being the voxel coordinates, Ω the image coordinates domain, $|\Omega|$ the number of voxels in Ω ($|\Omega| = 181 \times 217 \times 181$ in the current study) and $T = (u, v, w)$ the voxelwise 3D residual displacement vector field.

The proposed RegQCNET is thus designed to predict registration RMSE using two given images: a reference template and a registered one.

2.2.2. Implemented deep neural network. Figure 3 describes the architecture of the proposed quantitative CNN-based QC for image-to-template registration. Input images were first down-sampled by a factor 4 (note that a down-sampling factor 2 was also tested

and discussed). We used a convolutional encoder followed by a 3 regression layers per resolution level using a basis of 24 filters of 3×3×3 (*i.e.*, 24 filters for the first layer, 48 for the second and so on). Each block was composed of batch normalization, convolution and ReLU activation. We employed the following parameters: batch size = 1, optimizer = Adam with default parameters, epoch = 100, loss = Mean Square Error (MSE) and dropout = 0.5 after each block. We used 2 input channels (the down-sampled T1w and the template images).

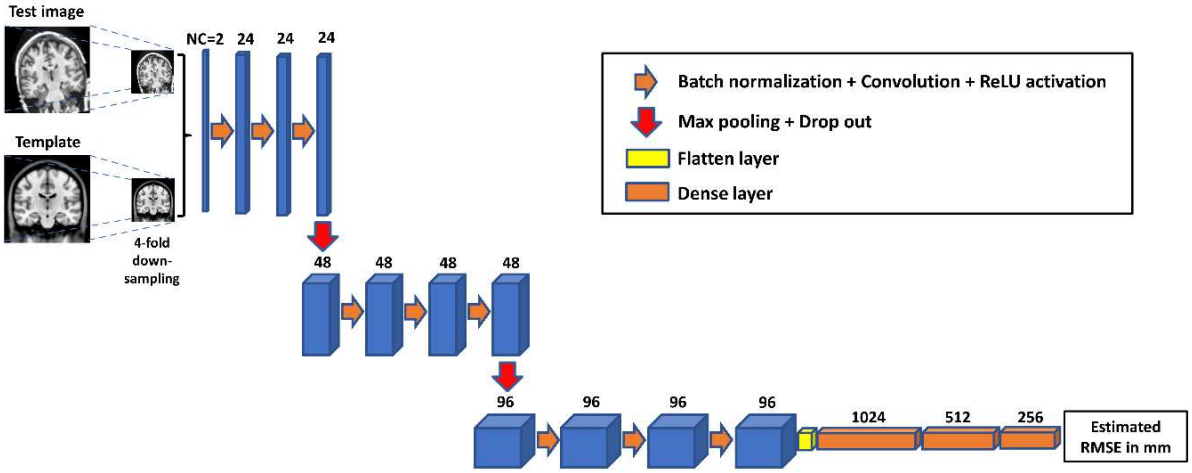

Figure 3: Architecture of the proposed RegQCNET for image-to-template registration. Each block is composed of batch normalization, convolution and ReLU activation. The number of input channels (NC) as well as the number of $3 \times 3 \times 3$ filters are indicated on the top of each block.

2.3. Experimental setup.

2.3.1. Assessment of RegQCNET on the "Simulated testing dataset". For this dataset, RegQCNET estimations were challenged against real RMSEs by evaluating R^2, slope and Y-intercept of a linear regression. While the R^2 provides information about the precision of the proposed QC, the slope and Y-intercept quantifies its accuracy. We compared results obtained using different sizes for the training dataset (*i.e.*, with N=100, 1000 and 10000, respectively).

To assess the robustness of our method to inaccuracy in image normalization and inhomogeneity correction we performed two experiments. First, we used a given uniform intensity shift to simulate normalization inconsistencies between images. Second, we used a non-uniform intensity bias to simulate inaccurate inhomogeneity correction within images.

Robustness against uniform intensity shift. To evaluate the robustness of RegQCNET to incorrect image normalization, RegQCNET was challenged against uniform intensity variation applied on all scans included in the testing dataset. For this purpose,

the test scans were identically disturbed as follows: Let `MAX_INTENSITY` be the maximum intensity of an image I, and J be the image I after application of the spatially homogeneous intensity bias. For all voxel location $\vec{r}$, the intensity $I(\vec{r})$ was multiplied by a factor 2, while being restricted in the interval original intensity range $[0, \texttt{MAX\_INTENSITY}]$:

$$J(\vec{r}) = \begin{cases} I(\vec{r}) \times 2 & \text{if } I(\vec{r}) \times 2 < \texttt{MAX\_INTENSITY} \\ \texttt{MAX\_INTENSITY} & \text{otherwise} \end{cases} \tag{2}$$

Robustness against non-uniform intensity bias. To evaluate the robustness of RegQCNET to error during inhomogenety correction, RegQCNET was challenged against a non-uniform intensity bias applied on all scans included in the testing dataset. For this purpose, the test scans were identically disturbed as follows: Let $\vec{r}_0 = (x_0, y_0, z_0)$ be the voxel coordinates at the central position of an image I, and K be the image I after application of the spatially heterogeneous intensity bias. For all voxel location $\vec{r}$, the intensity $I(\vec{r})$ was weighted by a voxel-wise exponential decay as follow:

$$K(\vec{r}) = I(\vec{r}) \times \exp\left(-\frac{\|\vec{r} - \vec{r}_0\|_2^2}{2\sigma^2}\right) \tag{3}$$

Practically, intensities in voxels located close to the central position $\vec{r}_0$ were less disturbed than those located further. In the scope of this study, we used $\sigma = 60$.

2.3.2. Assessment of RegQCNET on the "Real testing dataset". First, the 3953 brain images of the "Real testing dataset" were visually inspected to build a gold standard. 13 mis-registered brain images were detected and considered as a "negative case" for the rest of the manuscript. The "Simulated training dataset" with $N = 10000$ was here employed for training. The accuracies, area under the ROC curve (AUROC), sensitivity, specificity, positive predictive values (PPVs) and negative predictive values (NPVs) were recorded for the following experiments:

Manually defined threshold. RegQCNET was used to differentiate scans with RMSE higher than a user-defined threshold noted δ (δ was expressed in millimeters). We tested δ-values in the following set: $\{5, 10, 20, 50\}$ mm. The AUROC was computed using the method detailed in (Cantor & Kattan 2000).

Automatically defined threshold. A 10-fold-stratified cross-validation was used to evaluated the performance of RegQCNET when using an threshold automatically tuned by machine learning algorithm. The dataset of 3953 brains was randomly partitioned into training (90%) and testing (10%) subsets (making sure that at least one positive and one negative case were included in each subset). For this purpose, the following classification algorithms were applied using the commercial software Matlab (©1994-2020 The MathWorks, Inc.)/"Statistics and Machine Learning" toolbox: logistic

regression (LR), support vector machine (SVM), naïve bayes (NB) and random forest (RF). Default hyper-parameters in Matlab implementations were employed for RF and SVM (RF: Classification method/100 bagged decision trees, SVM: supports sequential minimal optimization/box constraint/linear kernel) (Kohavi 1995). The cross-validation steps were repeated 1000 times with shuffling of the folds. Finally, average metrics, standard deviations and confident intervals were calculated.

Comparison with usual image features. Our automatically defined threshold experiment was also conducted using correlation coefficient (CC) and mutual information (MI) for comparison.

2.4. Hardware and implementation details.

We evaluated the computational burden of our proposed method using an Intel Xeon E5-2683 2.4 GHz (2 Hexadeca-core) with 256 GB of RAM equipped by a GPU Nvidia Tesla V100 with 16 GB of memory. The computation time during the testing session was evaluated without and with the use of the GPU. Our implementation was using Tensorflow 1.4 and Keras 2.2.4.

3. Results

Fig. 4 shows typical images generated using synthetic affine transformations. Middle transversal, coronal and sagittal slices are reported for several 3D volumes. The template used as reference for affine image registration (see section 2.1.1) is displayed in the first row. The second (scan #1) row shows 3D brain images from the original lifespan dataset (RMSE considered equal to 0 mm). Lower rows (scan $\#[3-5]$) show examples of 3D scans obtained after the application of simulated spatial transformations of various amplitudes, as described in section 2.1.2. Note that several images underwent different masking (see (f) on the second row) due to defacing (e.g., NDAR dataset).

3.1. Assessment of RegQCNET on the "Simulated testing dataset".

Fig. 5 reports the precision and the accuracy of RegQCNET obtained on the "simulated testing dataset" (as described in section 2.1.2) using a training dataset of 100 (5a), 1000 (5b) and 10000 (5c) scans. As one can expect, the precision (rated by the R^2 of the linear fit) improves when the size of the training dataset increased. The R^2 converged slowly toward 1 along with the size of the training dataset increased (R^2 equal to 0.84, 0.95 and 0.99 were obtained using 100, 1000 and 10000 images, respectively). The accuracy (rated by the slope and the Y-intercept of the linear regression) followed the same trend. As long as N increased, the slope and the Y-intercept converged toward optimal values (*i.e.*, 1 and 0, respectively).

Fig. 6 shows the impact uniform intensity shift and spatially varying intensity bias (as described in section 2.1.2) on the performance of the proposed RegQCNET. While

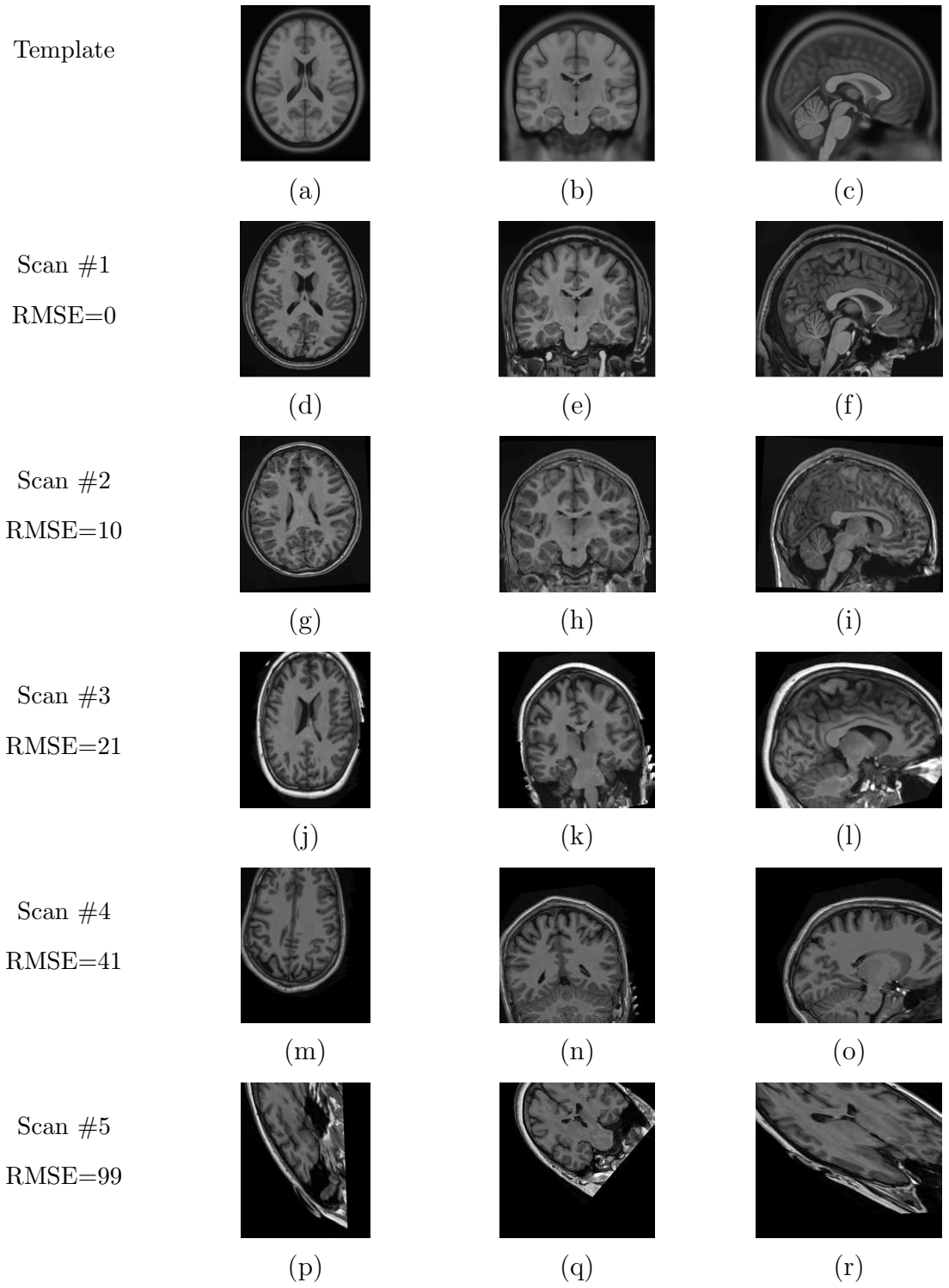

Figure 4: Typical images generated using synthetic affine transformations. Transversal (left column), coronal (middle column) and sagittal (right column) central slice are reported for the template scan (a-c) and for different subjects/various RMSE values. Each generated image is referred to as "Scan $\#[1-5]$".

a uniform intensity shift did not deteriorate the precision (R^2=0.99) and the accuracy (slope=0.98/Y-intercept=-1.08) (6c) this observation did not hold when a non-uniform bias (6e): in this last case, both precision (R^2=0.98) and accuracy (slope=0.89/Y-intercept=7.54) were slightly worse.

3.2. Assessment of RegQCNET on the "Real testing dataset".

Fig. 7 shows RMSE estimated by RegQCNET on the 3953 tested brain MRIs. Transversal cross-section of mis-registered images, as detected by visual inspection, are reported above and below the graph. These images can be visually compared to the corresponding template cross-section reported in Fig. 4a. In two images (referred to as case #9 and case #11 in Fig. 7) of patients with Alzeihmer's Disease, very large lateral ventricle slightly disturbed RegQCNET (estimated RMSE $<$ 20 mm). In the 9 other images, huge mis-registrations are observable, which have been detected by RegQCNET (estimated RMSE $>$ 50 mm).

3.2.1. Manually defined threshold. Table 1 reports classification scores of RegQCNET using different manually defined thresholds. A classification threshold δ = 10 mm provided best scores: accuracy=99.6%, AUROC=1.0, sensitivity=99.6%, specificity=100.0%, PPV=100.0%, and NPV=44.8% (note that NPV=44.8% means here that 16 good-registered images were considered as mis-registered). Good- and mis-registered images were thus correctly identified in 3924/3940 and 13/13 brain images, respectively.

3.2.2. Automatically defined threshold. The RegQCNET output served as a metric in all tested machine learning classifiers (see Table 2). In particular, using a logistic regression classifier, QC scores were: accuracy=96.0%, AUROC=1.0, sensitivity=95.9%, specificity=100.0%, PPV=97.5%, and NPV=93.7%.

3.2.3. Comparison with usual image metrics. Very poor scores were obtained using CC (best classifier=naïve bayes) and MI (best classifier=logistic regression): a good detection of correctly registered images was achievable by accepting a large amount of false-negative cases, as shown in Fig. 8. Conversely, a perfect detection of mis-registered images was only achievable by accepting a dramatic impact on the sensitivity (*i.e.*, 1.7% and 30.9% for CC and MI, respectively, as shown in Table 2).

4. Discussion

The proposed method aims at quantifying the amplitude of the spatial affine mismatch between a brain MRI and a template. To this end, we used RMSE as criterion to evaluate affine registration quality. Our experimental results demonstrate that the proposed RegQCNET outperforms traditional intensity-based criteria. Moreover, using

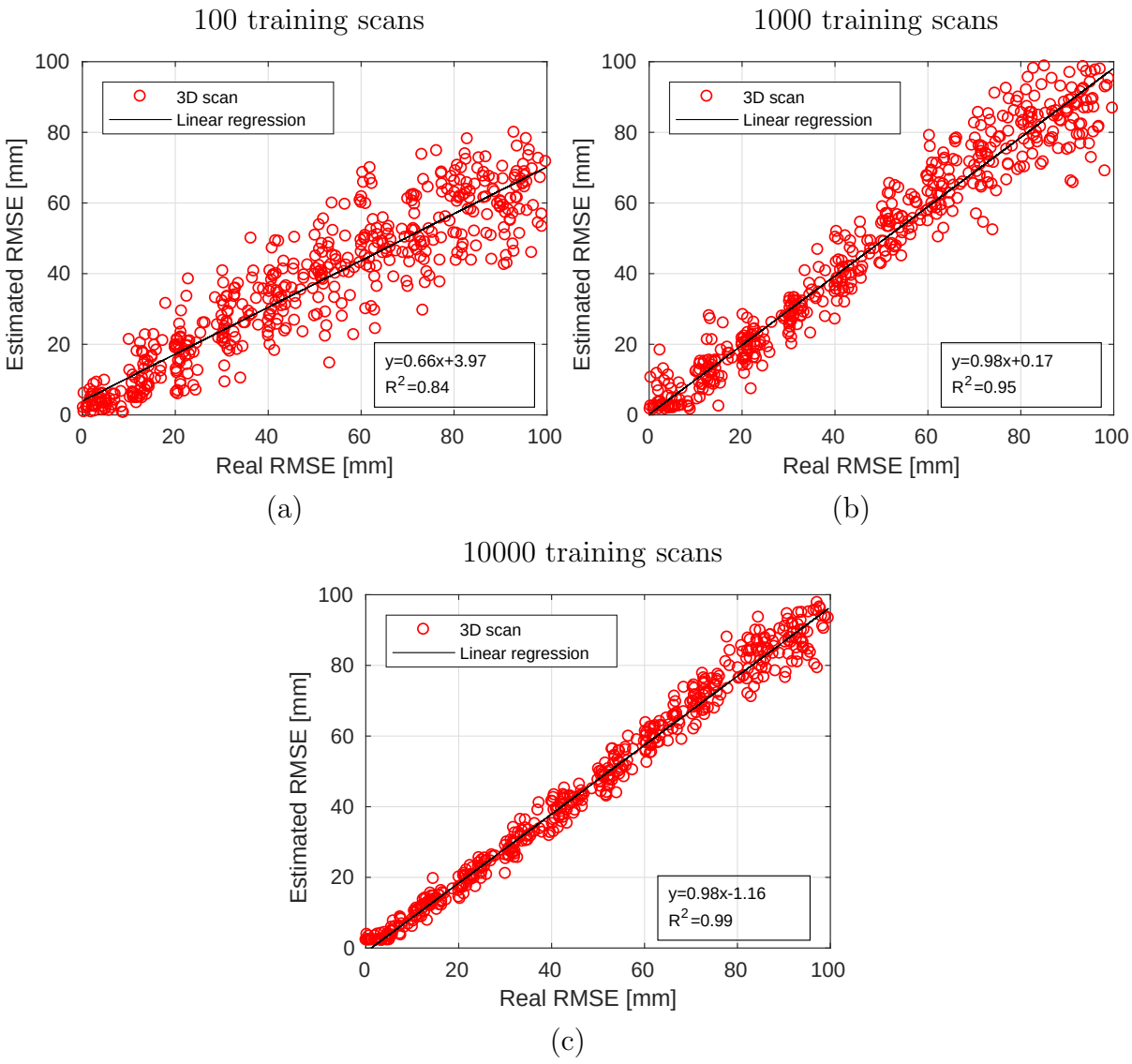

Figure 5: RMSE estimated by RegQCNET on the simulated dataset using training set composed of 100 (a),1000 (b) and 10000 (c) scans. Estimated RMSEs are plotted against real RMSEs and the R^2, slope and Y-intercept of a linear regression are reported in the insert of each graph.

automatic threshold, our approach delivers reproducible results and minimizes operator dependency.

It can be observed that good scores were obtained thanks to the use of i) a well-balanced training population covering the entire lifespan and ii) a well-balanced distribution of the simulated transformations (*i.e.*, uniform distribution of the resulting RMSEs). It is interesting to highlight that comparable precision and accuracy were obtained on simulated data (as described in section 2.1.2) using a subsampling factor 2 on the images (instead of the subsampling factor 4 used in the presented results).

One can distinguish two potential contributions on the apparent image-to-template

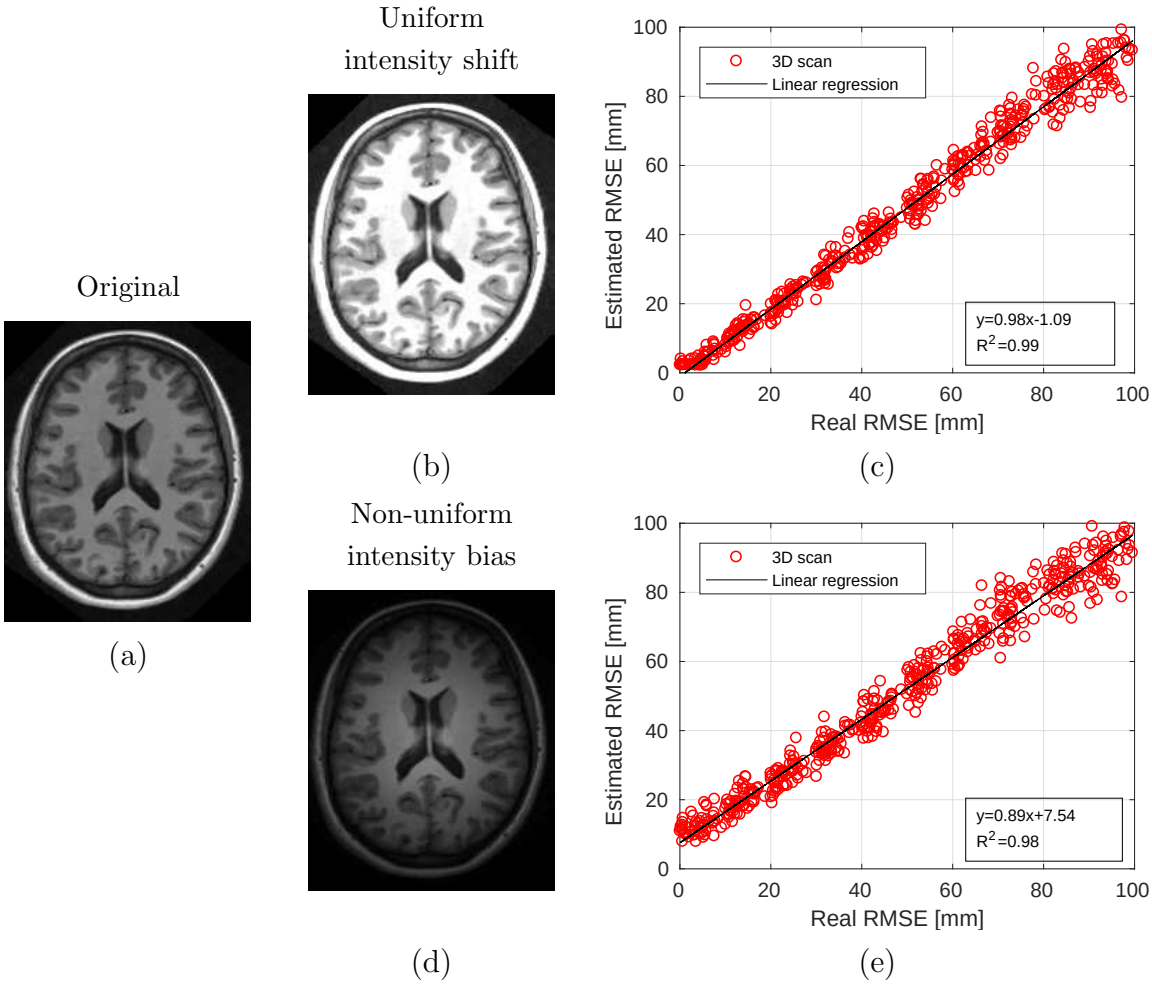

Figure 6: Results obtained on the simulation experiment when an intensity perturbation is applied on the testing dataset. A training dataset composed by 10000 scans (without intensity bias) was used. The axial slice of a brain scan is reported before (a) and after application of an uniform (b) and a non-uniform (d) intensity bias. Estimated RMSEs are plotted against real RMSEs for the uniform (c) and the non-uniform (e) bias, and the R^2, slope and Y-intercept of a linear regression are reported.

mismatch: (i) the effective registration error that we aim to quantify and (ii) the anatomical variability between the actual image and the template (in particular, the size of the brain varies a lot along the lifespan). Any unobserved spatial transformations/brain shapes/image artifacts during the training step may disturb in turn the proposed quantitative CNN-based QC. This phenomenon can be observed in Fig. 5a where an insufficiently populated training dataset was employed (100 images). In turn, a dramatic impact arise on both precision and accuracy. Note that using correlation or mutual information as registration QC, any intensity variation between a given brain MRI and the template is attributed to an image mismatch. The tested machine learning classifiers thus provided very poor results in terms of accuracy,

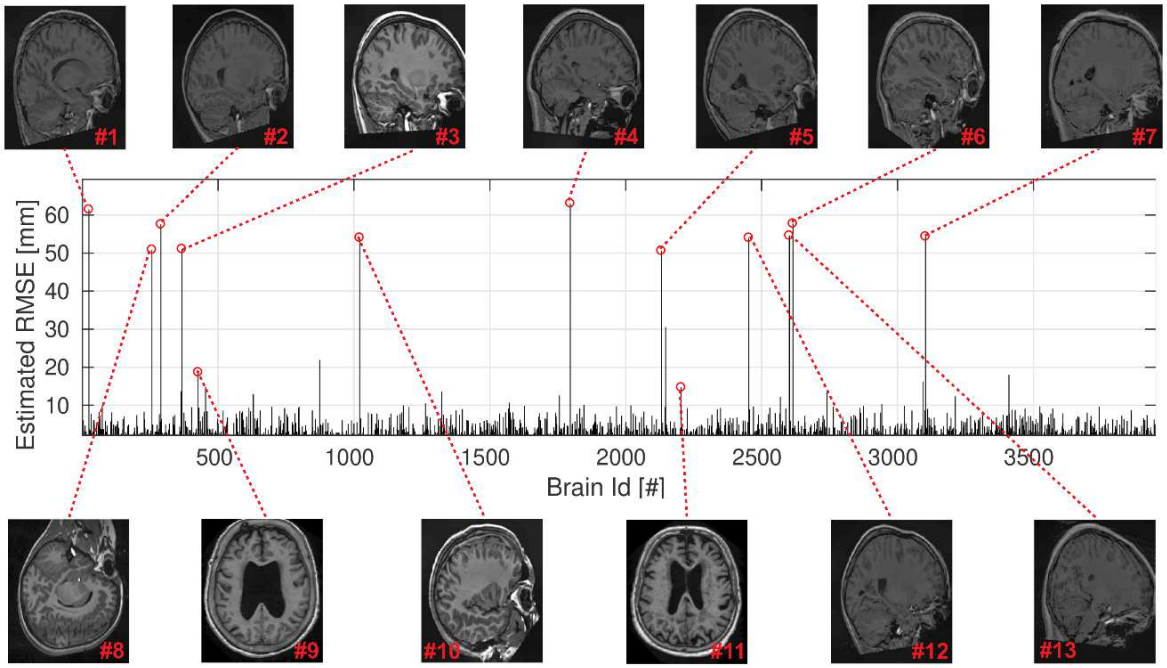

Figure 7: RMSEs estimated by RegQCNET on the 3953 brains of the data base. Mis-registered images (transversal), as detected by visual inspection, are reported above and below the graph. The corresponding template cross-section is shown in Fig. 4a.

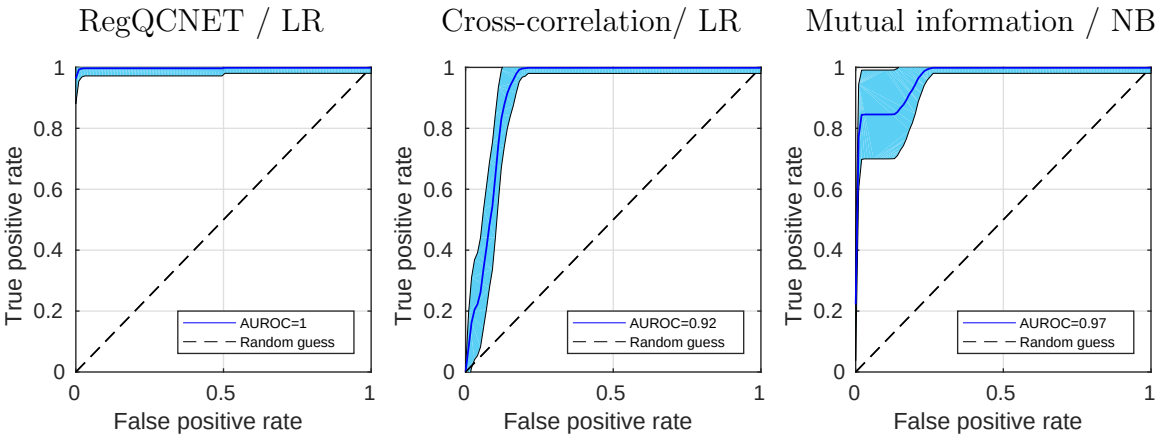

Figure 8: ROC curves obtained using the three tested indicators (RegQCNET (a), CC (b) and MI (c)) as binary classifiers (LR (a), LR (b) and NB (c)) for the two image populations (*i.e.*, correctly registered vs. mis-registered) after 10-fold cross-validation.

sensitivity, PPV and NPV.

Another limitation arise when an intensity perturbation occurs between training and testing. While a spatially homogeneous bias did not impact the performance (Fig. 6c), an accurate correction of spatially heterogeneous intensity bias is a necessary prerequisite to obtain good performance when using RegQCNET (Fig. 6e) (Tustison

Classification scores: Manually defined threshold						
Classification threshold (δ) [mm]	Accuracy	AUROC	Sensitivity	Specificity	PPV	NPV
5	91.6	0.96	91.6	100.0	100.0	3.8
10	**99.6**	**1.00**	**99.6**	**100.0**	**100.0**	**44.8**
20	99.9	0.92	99.9	84.6	99.9	84.6
50	99.9	0.92	100.0	84.6	99.9	100.0

Table 1: Classification scores of the proposed RegQCNET on the "Real testing dataset". Quantitative scores (*i.e.*, computer-assisted) are given, assuming the qualitative inspection (*i.e.*, visual) as a gold standard. AUROC: area under the ROC curve; PPV: positive predictive value; NPV: negative predictive value. Accuracies, sensitivities, specificities, PPVs, and NPVs are shown in percentages. Best performance are reported in bold font.

et al. 2010) (Nyúl et al. 2000). It has to be noted that our framework was robust to various degrees of masking due to defacing (*e.g.*, NDAR dataset).

As one can expect, the range of spatial transformations during training has to be carefully determined. That brings us to an inherent limit of the proposed technique. Indeed, a RMSE extrapolation outside training limits is intrinsically not possible using our CNN-based approach and thus a large training range is mandatory. This limitation could be limited by training RegQCNET on a larger range of RMSE.

Using the proposed CNN-based QC, a few tenth of a seconds (700 ms and 200 ms without and with the use of GPU acceleration, respectively) is needed to provide a quantitative prediction of the image-to-template alignment accuracy. This perfectly meets our computational requirements related to the inclusion of this QC step in massive processing.

While our experimental results demonstrate that a compact network is an efficient solution to estimate the quality of affine registration between a T1-MRI and a template, the direct translation of our approach to radiation therapies is not straightforward. In the context of radiation therapies, several performance indicators and registration QC solutions have been proposed (see (Paganelli et al. 2012) and (Brock et al. 2017)). Concerning abdominal organs, it must be underlined that deformations are non-rigid and thus extension of our framework to non-rigid registration would be required. For instance, recent works conducted in the abdomen used biomechanical criteria to assess image registration accuracy (Zachiu et al. 2018) (Zachiu et al. 2020). Such approaches involve the estimation of mechanical stress, which would occur within the observed tissues. The calculated stress can then be compared to plausible physiological limits. We believe that a combination of these two complementary approaches (*i.e.*, the network and the biomechanical strategies) should be investigated in future studies.

Classifier	Accuracy	AUROC	Sensitivity	Specificity	PPV	NPV
Classification scores: Automatically defined threshold						
RegQCNET						
LR	**96.0±17.3**	**1.00±0.06**	**95.9±17.4**	**100.0±0.0**	**97.5±15.7**	**93.7±24.2**
	(94.7-97.2)	**(0.99-1.00)**	**(94.7-97.2)**	**(100.0-100.0)**	**(96.4-98.6)**	**(92.0-95.4)**
SVM	94.8±18.8	1.00±0.04	94.8±18.9	100.0±0.0	97.3±16.2	91.3±28.1
	(93.4-96.1)	(1.00-1.00)	(93.4-96.1)	(100.0-100.0)	(96.1-98.5)	(89.3-93.3)
NB	85.7±34.1	1.00±0.04	85.6±34.2	100.0±0.0	86.8±33.9	84.0±36.6
	(83.3-88.1)	(1.00-1.00)	(83.2-88.1)	(100.0-100.0)	(84.4-89.2)	(81.4-86.6)
RF	84.8±30.7	0.94±0.18	84.8±30.8	100.0±0.0	91.3±28.5	76.2±42.9
	(75.7-93.9)	(0.89-1.00)	(75.6-93.9)	(100.0-100.0)	(82.8-99.8)	(63.5-89.0)
CC						
LR	**0.8±6.0**	**0.92±0.05**	**1.7±12.4**	**100.0±0.0**	**0.7±8.4**	**0.6±5.2**
	(0.4-1.2)	**(0.92-0.92)**	**(0.8-2.5)**	**(100.0-100.0)**	**(0.1-1.3)**	**(0.2-1.0)**
SVM	0.6±3.9	0.76±0.32	0.7±7.4	100.0±0.0	0.9±9.3	0.3±0.0
	(0.4-0.9)	(0.74-0.79)	(0.2-1.3)	(100.0-100.0)	(0.1-1.7)	(0.3-0.3)
NB	0.6±3.2	0.92±0.05	1.1±9.4	100.0±0.0	1.0±9.9	0.3±0.0
	(0.4-0.9)	(0.91-0.92)	(0.4-1.8)	(100.0-100.0)	(0.3-1.7)	(0.3-0.3)
RF	0.3±0.0	0.48±0.06	0.0±0.0	100.0±0.0	0.0±0.0	0.3±0.0
	(0.3-0.3)	(0.47-0.50)	(0.0-0.0)	(100.0-100.0)	(0.0-0.0)	(0.3-0.3)
MI						
LR	30.3±40.8	0.97±0.06	30.1±41.0	100.0±0.0	38.9±48.8	20.8±40.2
	(27.4-33.2)	(0.96-0.97)	(27.2-33.0)	(100.0-100.0)	(35.4-42.4)	(18.0-23.7)
SVM	8.9±24.9	0.54±0.44	8.6±24.9	100.0±0.0	12.6±33.2	5.3±21.7
	(7.2-10.7)	(0.50-0.57)	(6.9-10.4)	(100.0-100.0)	(10.2-15.0)	(3.8-6.9)
NB	**30.9±41.0**	**0.97±0.07**	**30.9±41.3**	**100.0±0.0**	**39.8±49.0**	**21.0±40.4**
	(28.0-33.8)	**(0.96-0.97)**	**(28.0-33.8)**	**(100.0-100.0)**	**(36.3-43.4)**	**(18.1-23.9)**
RF	27.9±40.2	0.75±0.23	27.6±40.4	100.0±0.0	36.7±48.5	20.6±40.3
	(18.9-36.9)	(0.70-0.80)	(18.6-36.7)	(100.0-100.0)	(25.8-47.6)	(11.5-29.6)

Table 2: Classification scores of the various classifiers on the "Real testing dataset". Quantitative scores were derived via evaluation of RegQCNET, correlation coefficient (CC), and mutual information (MI) (after 10-fold cross-validation) by various machine-learning algorithms (LR: logistic regression, SVM: support machine vector, NB: naïve bayes, RF: random forest). Quantitative indicators are shown with standard deviations and 95% confidence intervals in parentheses. Best performance are reported in bold font for each indicator.

5. Conclusion

This study demonstrates that quantitative estimation of registration mismatch between a brain image and a template can be achieved using 3D CNNs. However, to ensure the quality of the estimation, the training dataset have to be carefully designed. To this end, in this study we used: i) a gender and age well-balanced lifespan dataset covering the entire lifespan, ii) an uniformly distributed amplitudes of random spatial transformations to cover registration error from 0 to 100 millimeters, and iii) a sufficient amplitude range of simulated spatial transformations. The proposed tool can be used as quality control for automated image registration of T1-weigthed brain onto a reference template.

Future studies will include the extension of the proposed RegQCNET to complex elastics image deformations, the estimation of 3D RMSE maps, the impact of incomplete, noisy and corrupted brains, as well as the extension of the method to cross-contrast and multi-modal images.

Acknowledgment

Experiments presented in this paper were carried out using the PlaFRIM experimental testbed, supported by Inria, CNRS (LABRI and IMB), Université de Bordeaux, Bordeaux INP and Conseil Régional d'Aquitaine (see https://www.plafrim.fr/). This work benefited from the support of the project DeepvolBrain of the French National Research Agency (ANR-18-CE45-0013). This study was achieved within the context of the Laboratory of Excellence TRAIL ANR-10-LABX-57 for the BigDataBrain project. Moreover, we thank the Investments for the future Program IdEx Bordeaux (ANR-10-IDEX-03-02, HL-MRI Project), Cluster of excellence CPU and the CNRS/INSERM for the DeepMultiBrain project. This study has been also supported by the DPI2017-87743-R grant from the Spanish Ministerio de Economia, Industria Competitividad. The authors gratefully acknowledge the support of NVIDIA Corporation with their donation of a TITAN X GPU used in this research.

References

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L. R., Griffanti, L., G., D., Sotiropoulos, S. N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., Vidaurre, D., Webster, M., McCarthy, P., Rorden, C., Daducci, A., Alexander, D. C., H., Z., Dragonu, I., Matthews, P. M., Miller, K. L. & Smith, S. M. (2018). Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank, *NeuroImage* **166**: 400–424.

Avants, B. B., Tustison, N. J., Song, G., Cook, P. A., A., K. & Gee, J. C. (2011). A reproducible evaluation of ants similarity metric performance in brain image registration, *NeuroImage* **54**(3): 2033–2044.

Bannister, S. E., Page, D., Standen, T., Dunne, A., Rawling, J. P., Birch-Sykes, C. J., Wilson, M. Z., Holloway, S., McClelland, J. & Peters, Y. (2019). Deep neural networks for quality assurance of image registration.

Brock, K. K., Mutic, S., McNutt, T. R., Li, H. & Kessler, M. L. (2017). Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM radiation therapy committee task group no. 132, *Med Phys* **44**(7): e43–e76.

Cantor, S. B. & Kattan, M. W. (2000). Determining the area under the ROC curve for a binary diagnostic test, *Medical Decision Making* **20**(4): 468–470.

Chaogan, Y. & Yufeng, Z. (2010). DPARSF: a MATLAB toolbox for "pipeline" data analysis of resting-state fMRI, *Frontiers in Systems Neuroscience* **4**: 13.

Collins, D. L., Neelin, P., Peters, T. M. & Evans, A. C. (1994). Automatic 3d intersubject registration of MR volumetric data in standardized talairach space, *Journal of computer assisted tomography* **18**(2): 192–205.

Cook, P., Bai, Y., Nedjati-Gilani, S., Seunarine, K., Hall, M., Parker, G. & Alexander, D. (2006). Camino: open-source diffusion-MRI reconstruction and processing, Vol. 2759, p. 2759.

Coupé, P., Catheline, G., Lanuza, E. & Manjón, J. V. (2017). Towards a unified analysis of brain maturation and aging across the entire lifespan: A MRI analysis, *Human Brain Mapping* **38**(11): 5501–5518.

Coupé, P., Manjón, J. V., Lanuza, E. & Catheline, G. (2019). Lifespan changes of the human brain in alzheimer's disease, *Scientific Reports* **9**(1): 3998.

Dubost, F., de Bruijne, M., Nardin, M., Dalca, A., Donahue, K., Giese, A.K. aband Etherton, M., Wu, O., de Groot, M., Niessen, W. & Vernooij, M. (2019). Automated image registration quality assessment utilizing deep-learning based ventricle extraction in clinical data, **arXiv preprint arXiv:1907.00695**.

Eppenhof, K. A. J. & Pluim, J. P. W. (2018). Error estimation of deformable image registration of pulmonary ct scans using convolutional neural networks, *Journal of Medical Imaging* **5**(2).

Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C. & Collins, D. L. (2011). Unbiased average age-appropriate atlases for pediatric studies, *NeuroImage* **54**(1): 313–327.

Fonov, V. S., Dadar, M. & Collins, D. L. (2018). Deep learning of quality control for stereotaxic registration of human brain MRI, *bioRxiv* .

Friston, K. J., Ashburner, J., Frith, C. D., Poline, J. B., Heather, J. D. & Frackowiak, R. S. J. (1995). Spatial registration and normalization of images, *Human Brain Mapping* **3**(3): 165–189.

Hessam, S., Saygili, G., Glocker, B., Lelieveldt, B. P. F. & Staring, M. (2019). Quantitative error prediction of medical image registration using regression forests, **abs/1905.07624**.

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W. & Smith, S. M. (2012). FSL, *NeuroImage* **62**(2): 782–790.

Kim, H., Irimia, A., Hobel, S. M., Pogosyan, M., Tang, H., Petrosyan, P., Blanco, R. E. C., Duffy, B. A., Zhao, L., Crawford, K. L., Liew, S. L., Clark, K., Law, M., Mukherjee, P., Manley, G. T., Van Horn, J. D. & Toga, A. W. (2019). The LONI QC system: A semi-automated, web-based and freely-available environment for the comprehensive quality control of neuroimaging data, *Frontiers in Neuroinformatics* **13**: 60.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection, pp. 1137–1143.

Manjón et al., J. V. (2008). Robust MRI brain tissue parameter estimation by multistage outlier rejection, *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* **59**(4): 866–873.

Manjón, J. V. & Coupé, P. (2016). volbrain: An online MRI brain volumetry system, *Frontiers in neuroinformatics* **10**: 30.

Manjón, J. V., Coupé, P., L. M.-B. D. L. C. & Robles, M. (2010). Adaptive non-local means denoising of MR images with spatially varying noise levels, *Journal of Magnetic Resonance Imaging* **31**(1): 192–203.

Maurer, C. R. J., Fitzpatrick, J., Wang, M., Galloway, R. L. J., Maciunas, R. & Allen, G. (1997). Registration of head volume images using implantable fiducial markers, *IEEE Transactions on Medical Imaging* **16**: 447–462.

Nyúl, L. G., Udupa, J. K. & Zhang, X. (2000). New variants of a method of MRI scale standardization, *IEEE Transactions on Medical Imaging* **19**: 143–150.

Paganelli, C., Meschini, G., Molinelli, S., Riboldi, M. & Baroni, G. (2018). Patient-specific validation of deformable image registration in radiation therapy: Overview and caveats, *Med Phys* **45**(10): e908–e922.

Paganelli, C., Peroni, M., Riboldi, M., Sharp, G. C., Ciardo, D., Alterio, D., Orecchia, R. & Baroni, G. (2012). Scale invariant feature transform in adaptive radiation therapy: a tool for deformable image registration assessment and re-planning indication, *Physics in Medicine and Biology* **58**(2): 287–299.

Sled, J. G., Zijdenbos, A. P. & Evans, A. C. (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data, *IEEE Transactions on Medical Imaging* **17**(1): 87–97.

Song, X. W., Dong, Z. Y., Long, X. Y., Li, S. F., Zuo, X. N., Zhu, C. Z., He, Y., Yan, C. G. & Zang, Y. F. (2011). REST: a toolkit for resting-state functional magnetic resonance imaging data processing, *PLoS One* **6**(9): e25031.

Tustison, N., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A. & Gee, J. C. (2010). N4itk: Improved n3 bias correction, *IEEE Transactions on Medical Imaging* **29**(6): 1310–1320.

Tustison, N. J., Cook, P. A., Klein, A., Song, G., Das, S. R., Duda, J. T., Kandel, B. M., van Strien, N., Stone, J. R., Gee, J. C. & Avants, B. B. (2014). Large-scale evaluation of ants and freesurfer cortical thickness measurements, *NeuroImage* **99**: 166–179.

Wei, X., Warfield, S. K., Zou, K. H., Wu, Y., Li, X., Guimond, A., Mugler, J. P., Benson, R. R., Wolfson, L., Weiner, H. L. & R., G. C. (2002). Quantitative analysis of MRI signal abnormalities of brain white matter with high reproducibility and accuracy, *Journal of Magnetic Resonance Imaging* **15**(2): 203–209.

Zachiu, C., Denis de Senneville, B., Moonen, C. T. W., Raaymakers, B. W. & Ries, M. (2018). Anatomically plausible models and quality assurance criteria for online mono- and multi-modal medical image registration, *Physics in Medicine and Biology* **63**(15): 155016.

Zachiu, C., Denis de Senneville, B., Raaymakers, B. W. & Ries, M. (2020). Biomechanical quality assurance criteria for deformable image registration algorithms used in radiotherapy guidance, *Physics in Medicine & Biology* **65**(1): 015006.