Estimating the registration error of brain MRI data based on regression U-Net

Leandro Nascimento[1,2][0009−0009−0019−0370], Quentin François[2], Bertrand Duplat[2], Sinan Haliyo[3][0000−0003−4587−381X], and Isabelle Bloch[1][0000−0002−6984−1532]

[1] Sorbonne Université, CNRS, LIP6, Paris, France
leandro.nascimento@lip6.fr
[2] Robeauté, Paris, France
[3] Sorbonne Université, CNRS, ISIR, Paris, France

Abstract. Some neurosurgery procedures require precise information on the region of interest, and a quantitative control of the overall uncertainty. Such procedures often rely on image registration, which is an essential step in many of these workflows. However, the problem of registration error estimation (REE) remains a challenge, due to its lack of ground-truth. In this work, we establish different criteria to evaluate REE methods and we propose a regression U-Net, a supervised convolutional neural network approach, that is able to compute the REE for the case of deformable brain MRI mono-modal registration. The model is trained and tested separately on four modalities. The best result is for T1 images, with a root mean square deviation (RMSD) on the test set of 0.17 mm for images with 1 mm^3 isotropic voxels. We also tested the model generalization and transfer learning capabilities on a separate MRI data set with two modalities. For the T1 images, the direct inference has a RMSD of 0.62 mm and the transfer learning method leads to a RMSD of 0.19 mm, also for volumes of 1 mm^3 voxel size. These results demonstrate the feasibility of our approach and the possible use of a U-Net based model for REE in brain MRI registration. The proposed methods enable a better quantitative control of procedure uncertainty in neurosurgeries and open the way to closed loop robotic control in these procedures.

Keywords: Image registration · Registration error · Uncertainty estimation · Brain MRI · Convolutional neural networks · U-Net.

1 Introduction

Neurological procedures can be very sensitive to the overall uncertainty of the positioning of a critical element or the targeting of region of interest. Examples are the positioning of electrodes for stereoelectroencephalography (SEEG) or deep brain stimulation (DBS), and for targeting tumors on stereotactic radiosurgery (SRS). Medical image registration plays an indispensable role in these applications [1]. Indeed, the medical staff often exploits several images which need to

be aligned. In this context, the quality of alignment is key to guarantee a correct interpretation of the images. More importantly, knowing quantitatively the registration error is important to reduce the uncertainty and improve the quality of care as well as reduce side effects, specially for medical robotic operations [11].

In the last years, the registration algorithms have largely improved their performance thanks to machine learning methods. Nevertheless, the registration evaluation techniques have not benefited from the same improvement and stay qualitative in most use cases. It has been shown that measuring the quality of the registration is rather complex [14], especially for deformable image registration (DIR), i.e., when the transformation estimated during registration is non linear.

There are some proposed methods that address the registration error estimation (REE) problem. Nowadays, the most widely used algorithms are either based on segmentation, with criteria such as Dice-Sorensen coefficient (DSC) or Hausdorff distance (HD), or the target registration error (TRE) [4]. The segmentation based methods provide only a global indicator of a correct registration, which can also be mixed up with potential segmentation errors [15]. The TRE, which consists in measuring the Euclidean distance between corresponding points, typically anatomical landmarks, is often performed manually, and it is therefore prone to operator variability, that is not negligible compared to the small brain structures.

The main contribution of this work is the development of an automatic method that is able to measure the TRE for all voxels inside the region of interest (ROI). This method is based on a convolutional network (CNN), namely U-Net [17], used as a regression model for TRE prediction. The method is illustrated on 3D images of the brain, acquired with magnetic resonance imaging (MRI). We also show that the U-Net is capable to predict REE well when tested on another dataset, demonstrating its generalization capabilities. Finally, we show that transfer learning is efficient for improving REE results across two different data sets and across modalities. This method allows for better control of the uncertainty of the operation during neurological procedures, which improves the quality of care.

2 Requirements and Related Work

In this section, we define important criteria necessary for quantitatively assessing DIR algorithms within clinical procedures. We also present the reasons why the method has to fulfill these requirements, and to which extent existing methods satisfy them. Let us consider a 3D image $I : \Omega \to \mathbb{R}$, $\mathbf{x} \mapsto I(\mathbf{x})$, where $\Omega \subset \mathbb{R}^3$. The criteria are the following:

1. Automation (**A**): the method has to be fully automatic, so we can guarantee a better reproducibility when compared to manual methods, and less time consumption.
2. Density (**De**): the method has to output a "dense" error measure, which is defined as "for a given ROI, $\forall \mathbf{x} \in$ ROI if we know the error measure for $\mathbf{x}$,

then we also know the error for all its 26-neighbor voxels that are inside the ROI" (clearly, a voxel on the border of the ROI will have less neighbors). This is necessary because DIR needs a very high number of estimation points to be correctly evaluated [6]. In our case, we choose the brain as the ROI.
3. Unit (**U**): the method is required to measure the error in millimeters, so it can be used for stereotactic procedures, where surgeons measure errors in millimeters, for example, in deep brain stimulation (DBS) post-operative electrode positioning validation [9]. Adimensional coefficients are not suited for these clinical applications.
4. Direct Measure (**DM**): the method should provide a direct error measure to avoid inaccuracies from additional steps, which is especially important for stereotactic applications.

With these criteria, we can analyze the most used REE methods. First, the Dice similarity score does not provide a guaranteed correlation between REE and segmentation metrics, but it remains the most used method to evaluate new registration algorithms [4]. The advantage of Dice or other segmentation-based scores is that it can be calculated automatically with the segmentation models. It is, however, a surrogate method, since in reality it is a segmentation metric, hence it fails to satisfy the **DM** criterion.

Manual TRE is also extensively used in the literature since it provides a reference through the involvement of physicians' expertise. The main issue with TRE is the need of human interaction which fails to satisfy criterion **A**. Semi-automatic methods can reduce these errors but are still time consuming for an expert [12, 14]. As it was reviewed by Bierbrier, Gueziri and Collins [3], there are only a few works using this approach and they consist of corresponding landmarks detection [12, 19].

More recently, some machine learning (ML) based solutions have been proposed to address the limitations of the traditional methods. However, they face the same challenge of lack of ground truth [14]. This makes the training of supervised ML models more difficult. The creation of artificial transformations is mentioned as one possible solution to create a training reference [3, 14]. The only other relevant strategy listed in [3] is the use of ML methods to reduce limitations of TRE methods such as an automatic TRE estimator that is trained with manually annotated anatomical landmarks [18].

Another method is proposed in [7], which is capable to automatically estimate TRE for all voxels in the image in the case of mono-modal, non-linear lung CT REE. To achieve that, a supervised convolutional network is trained with artificial non-linear transformations created by thin plate splines. Analyzing this model according to our criteria, we see that it is a fully automatic method; it can generate a dense REE for all voxels in the ROI and also gives a measure in millimeters in a direct manner. However, the CNN model used is outdated with respect to the current state-of-the-art, and CT images are very different from MR images, which makes the method not directly applicable in our context.

In conclusion, the literature shows a lack of work on REE for brain image data that satisfies our constraints. Therefore, we propose a model that is able

to predict deformable mono-modal REE for brain MRI. Moreover, we assess the model performance when inferring on images belonging to a different data set, and how transfer learning can improve the accuracy and reduce the training time.

3 Data

3.1 BraTS-Reg

The first data set is from the 2022 BraTS-Registration challenge data set [2], also called BraTS-Reg. It is a multi-site, multi-modal MRI brain data set that comprehends a total of 250 patients, from which 140 are made available for the public. Each patient has four pre-operative MRIs with a brain tumor and four follow-up ones after the tumor removal. The four sequences are T1 weighted, T2 weighted, T1 contrast enhanced (T1CE) and fluid attenuated inversion recovery (FLAIR). The data set has many advantages, such as a preprocessing pipeline with skull stripping, interpolation to isotropic voxels of size 1 mm^3, and volume dimensions set to $240 \times 240 \times 155$. All the volumes are rigidly registered to the SRI24 atlas anatomical space [16]. Another advantage is the presence of manual landmarks annotated by six experts. For each pair of MRIs (pre and post-operative) of a given patient, from 6 to 50 anatomical points are annotated, which can serve as an additional criterion to evaluate the quality of our REE model. In our experiments, we first pseudo-randomly shuffle the 140 images, then we separate 30 for the test, and for the remaining 110, we re-split them into 90 samples for training and 20 for validation.

3.2 CERMEP

The second data set comes from the CERMEP institute [13]. It contains paired MRI exams from 37 healthy patients. The modalities are FLAIR and T1 weighted. The images are available in the original voxel size which is $1.2 \times 1.2 \times 1.2$ mm^3 and in a resampled size of 1 mm^3. The resampled images are also normalized to the MNI atlas space and are of size $207 \times 243 \times 226$. Comparing to the BraTS-Reg data set, the images do not come with the skull strip pre-processing and do not have pre-annotated anatomical landmarks. The CERMEP data set is used for testing the generalization of the model trained with BraTS-Reg and its capability of fine-tuning to another data set. For the fine-tuning experiment, we divide the 37 patients into 23 for training, 7 for validation and 7 for testing.

4 Methods

With the goal of obtaining a mono-modal REE model that is able to predict the error for every voxel inside the image ROI, we propose a CNN network that is trained on artificially transformed images. The REE problem is formulated here as a supervised 3D image regression task.

In a registration context we denote the fixed image by $I_F : (\Omega \subset \mathbb{R}^3) \to \mathbb{R}$, the moving image by $I_M : (\Omega \subset \mathbb{R}^3) \to \mathbb{R}$, and the registered image by $\hat{I}_R : (\Omega \subset \mathbb{R}^3) \to \mathbb{R}$, with $\hat{I}_R(\mathbf{x}) = I_M[\hat{\mathbf{T}}(\mathbf{x})]$, where the transformation $\hat{\mathbf{T}} : \Omega \to \Omega$ is obtained by a registration algorithm, and is supposed to be invertible.

We assume that there is an ideal transformation $\mathbf{T}(\mathbf{x})$ (also supposed to be invertible), that maps all the corresponding voxels from I_F to I_M so that $I_F(\mathbf{x}) = I_M[\mathbf{T}(\mathbf{x})]$. Since usually $\mathbf{T}(\mathbf{x})$ is different from $\hat{\mathbf{T}}(\mathbf{x})$, there is a residual transformation $\mathbf{T}_{\text{res}}$ defined as $\mathbf{T}(\mathbf{x}) = \mathbf{T}_{\text{res}}(\hat{\mathbf{T}}(\mathbf{x}))$, i.e., with the invertibility assumption $\mathbf{T}_{\text{res}}(\mathbf{x}) = \mathbf{T}(\hat{\mathbf{T}}^{-1}(\mathbf{x}))$. $\mathbf{T}_{\text{res}}$ represents the deformation vector field (DVF) between the registered image $\hat{I}_R$ and the fixed image I_F, $I_F(\mathbf{x}) = \hat{I}_R[\mathbf{T}_{\text{res}}(\mathbf{x})]$.

In order to train our supervised model, we choose samples from the data set as the fixed images I_F, and we apply artificial transformations (random B-splines in our experiments) to obtain a simulation of $\hat{I}_R$ images. These transformations are the equivalent of $\mathbf{T}_{\text{res}}^{-1}$, because $\hat{I}_R(\mathbf{x}) = I_F[\mathbf{T}_{\text{res}}^{-1}(\mathbf{x})]$. From $\mathbf{T}_{\text{res}}^{-1}$, we can calculate a registration error map $E(\mathbf{x}) = ||\mathbf{T}_{\text{res}}^{-1}(\mathbf{x})|| = ||\mathbf{T}_{\text{res}}(\mathbf{x})||$. E is our ground truth that enables the supervised learning process and it is called target error map (TEM). Figure 1 depicts the training workflow as it has been described.

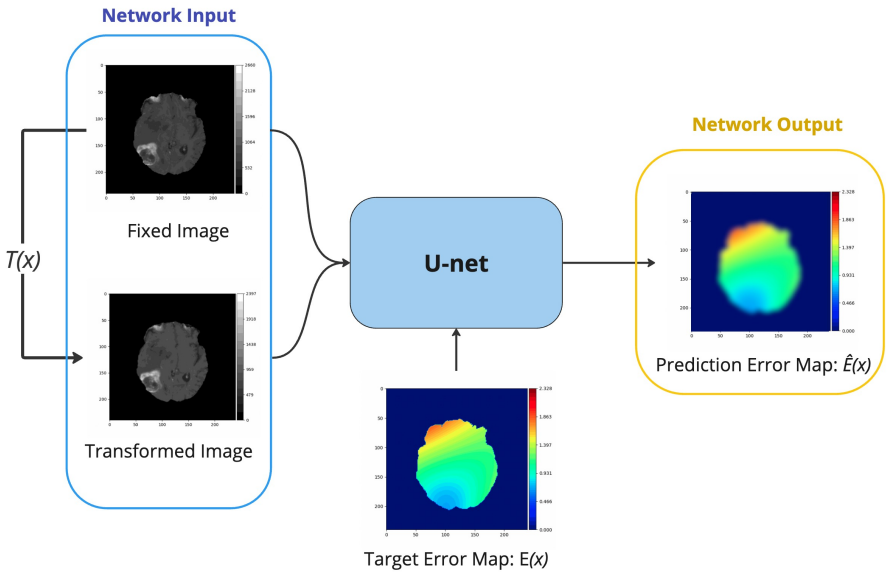

Fig. 1. Diagram displaying the regression training workflow of our mono-modal registration error estimation method. The input are two 3D images, the fixed and moving images. The target error map (TEM) and the prediction error map (PEM) are used for calculating the loss function and hence for training the method.

The proposed method uses as input a pair of 3D images: I_F and $\hat{I}_R$. The output of the model is $\hat{E}$, the predicted error map (PEM). The model is trained by minimizing a *masked* L_1 norm of the difference between the PEM ($\hat{E}$) and the TEM (E):

$$L_1 = \frac{\sum_{\mathbf{x}\in\Omega} M(\mathbf{x})|E(\mathbf{x}) - \hat{E}(\mathbf{x})|}{\sum_{\mathbf{x}\in\Omega} M(\mathbf{x})} \tag{1}$$

where $M(\mathbf{x})$ is an indicator or mask function value that is equal to 1 if $\mathbf{x} \in ROI$ and equal to 0 otherwise.

4.1 Residual transformations

To create the artificial images, we use B-spline transformations with a grid size of $2 \times 2 \times 2$ voxels. To determine the values of our B-spline grid, we sample random uniform values in $[-2, 2]$ for each axis. Since the isotropic voxel size is 1 mm^3, the distortion can be easily converted to distance unit.

For the implementation, we use the `gryds` library that was developed by Eppenhof and Pluim [8]. It has the advantage of being able to run the deformations in a CUDA environment, speeding up the transformations computation and therefore the learning process.

For each image in the training set, at each epoch we apply different transformations, resampling from the $[-2, 2]$ interval, as a data augmentation strategy. For the validation and test sets, the same transformation is applied to each sample so the evaluation remains consistent throughout the epochs.

Once we have the distorted image, we can calculate the TEM (E), based on $\mathbf{T}_{\text{res}}^{-1}$ as explained above. The ROI is chosen as the fixed image brain region, since it is the reference image. We apply the ROI mask to the TEM because otherwise the model would be forced to learn REE from zero information.

4.2 Model architecture

For the model architecture, we employ a modified version [10] of the U-Net model [17], as implemented in the MONAI framework [5]. The modifications are: (i) the possibility of using residual units in the encoding layers, and (ii) the strided convolution is the first layer in the residual unit, instead of the last one as in the original implementation. Although U-Net was created for segmentation problems, we choose its architecture because it has been shown in a previous study that the model can be used for regression tasks [8]. In order to adapt the prediction layer for the regression task, a Rectified Linear Unit (ReLU) is incorporated to ensure that the predictions are constrained to non-negative real values, given that the output represents a norm of the DVF ($\mathbf{T}_{\text{res}}^{-1}$).

Other modifications were made with different motivations, namely reducing the number of convolutional filters in the first stage to 16 instead of 64. The first reason is the fact that we are working with a 3D U-Net model, which is more complex than the original 2D U-Net, hence we prefer to balance the number of parameters with less convolutional layers. Moreover, a larger model has a

greater impact on the GPU memory used and also needs more data in order to better train the larger number of parameters. Finally, we intend to test whether a model with less initial layers is enough to produce our intended results of REE. We keep 5 levels in the U-Net and double the number of filters on each, so we have 16, 32, 64, 128 and 256 filters at the successive levels.

The remaining MONAI U-Net hyper-parameters are set to their default values. Most importantly, this means that the kernel size for convolution and transpose convolution is equal to $3 \times 3 \times 3$. Also, there is an instance normalization layer after the convolution and before the activation layer. Lastly, the stride values for all levels are set to 2, i.e., we have to set all the dimension sizes of our input to be divisible by $2^4 = 16$, since we have four stride spatial reduction steps. For that, we pad with zero-valued voxels the axial dimension from 155 to 160 to satisfy this constraint.

4.3 Experiments

To evaluate our approach across various MRI types, we trained separate U-Net models for each modality in the BraTS-Reg dataset (four modalities) and the CERMEP dataset (two modalities), all using the same settings. We also tested how well the models trained on BraTS-Reg performed on CERMEP data, specifically on T1 and FLAIR images. Additionally, we applied a transfer learning approach by re-training a BraTS-Reg model with CERMEP data. This let us compare performances between models trained directly on CERMEP, those trained on BraTS-Reg and tested on CERMEP, and those trained with transfer learning.

4.4 Training

The training of the U-Net is supervised, where the input is the concatenation of two 3D volumes composed by I_F and $\hat{I}_R$. The output is the PEM ($\hat{E}$), the loss function is L_1 as stated in Equation 1. We trained our model during 500 epochs, with the Adam optimizer set with a learning rate of 10^{-4}. The moment estimates hyperparameters were set to 0.9 and 0.999 respectively, which are their default values in the PyTorch library. All these parameters were kept for all experiments because we propose to evaluate the different performances of the same setup across the different modalities and data sets. The only difference is for the transfer learning, where we trained the model for 125 epochs, which proved sufficient to achieve convergence, thanks to the initial weights provided by the model trained on BraTS-Reg.

5 Results and Discussion

5.1 Quantitative Evaluation

To measure the accuracy of the REE model, we compute the root mean square deviation (RMSD) between the PEM and TEM, i.e., for any image in the test

set:

$$RMSD = \sqrt{\frac{1}{|\mathrm{ROI}|} \sum_{\mathbf{x} \in \mathrm{ROI}} [\hat{E}(\mathbf{x}) - E(\mathbf{x})]^2} \tag{2}$$

Additionally, we evaluate the RMSD only on the coordinates of the annotated landmarks available in the BraTS-Reg data set. These anatomical landmarks are commonly used by physicians to evaluate registration accuracy, so it is interesting to evaluate our model within these regions. Furthermore, we can compare the two RMSDs and verify if their values are coherent. This shows how the landmarks error approximates the global error in the ROI. We also compute the signed difference $\hat{E}(\mathbf{x}) - E(\mathbf{x})$ and the histogram of these values over all $\mathbf{x}$ and all test set images; we call it evaluation histogram. We then compute the mean, standard deviation (SD) and skewness of the evaluation histogram for each modality. It is important to mention that the SD of the evaluation histogram can also be seen as the uncertainty of our error predictions. All the results coming from the model instances trained only on BraTS-Reg data set are shown in Table 1.

Table 1. Evaluation of the REE model trained on BraTS-Reg data set.

Modality	Global RMSD (mm)	Landmarks RMSD (mm)	Mean (mm)	SD (mm)	Skewness (mm)
T1	0.17	0.15	0.01	0.17	-0.12
T1CE	0.22	0.20	0.00	0.22	-0.15
T2	0.23	0.20	0.04	0.20	-0.33
FLAIR	0.23	0.22	0.00	0.22	-0.15

Table 1 shows that the best result is obtained for the T1 modality, with **0.17 mm** of RMSD. Similar results are obtained for the other modalities, with all the remaining three having essentially the same performance. The reason for that difference is likely that T1 is usually the modality used for anatomical structural details hence it has smaller voxel size. Since the preprocessing from the BraTS-Reg data set re-scaled all images to 1 mm^3 voxel size, the original images that had voxel sizes larger than that will have less information for the REE model, since the interpolation induces spatial imprecision. Nevertheless, in spite of the existence of this extra step of re-scaling inherent to the dataset, we still obtain a sub-voxel-size accuracy for all modalities, which was the objective for surgical applications.

Moreover, from the results in Table 1, we can observe that all mean values are close to zero and the standard deviations follow values close to the RMSD. The skewness shows that the model has a slight tendency to underestimate the registration error, since they are negative low values.

We present the same results for the experiments with the CERMEP data set, with the exception of the landmarks RMSD, since it is not an annotated dataset. Table 2 shows that the inference from the model trained on BraTS-Reg

images also has a sub-voxel size RMSD on its error prediction (0.62 mm for the T1 for instance). The positive mean on both modalities shows that the inference on CERMEP from the models trained on BraTS-Reg tends to overestimate the error. The experiment with training from scratch evidences that, for the same U-Net architecture and hyper-parameters setup, the REE results are also on sub-voxel size accuracy, although a bit less accurate than the one trained on BraTS-Reg (**0.17** mm against **0.29** mm). Finally, the transfer learning technique proved to be the best result for CERMEP, not only needing less epochs for its training (125 compared to 500 from scratch), but also improving the RMSD for both T1 and FLAIR modalities (from 0.24 mm to 0.19 mm and from 0.29 mm to 0.24 mm, respectively).

Table 2. Evaluation of performances for different experiments on the CERMEP data set.

Type of experiment	Modality	RMSD (mm)	Mean (mm)	SD (mm)	Skewness (mm)
Inference from BraTS-Reg	T1	0.62	0.28	0.56	-0.69
Inference from BraTS-Reg	FLAIR	0.69	0.33	0.61	-0.50
Transfer Learning	T1	0.19	0.00	0.19	-0.44
Transfer Learning	FLAIR	0.29	-0.14	0.25	-0.28
From Scratch	T1	0.24	0.03	0.24	-0.59
From Scratch	FLAIR	0.33	0.00	0.33	-0.44

5.2 Qualitative results

For a qualitative analysis, we evaluate if the general aspect of the PEM reproduces the aspect given by the TEM. Figure 2 depicts one case of the test set. In each line, we show the results for T1CE, T1, T2 and FLAIR modalities of the BraTS-Reg set in order to evidence the ability of the REE model to learn the task for different types of acquisition. In the first line, we have a fixed image slice, used as reference; in the second line, we show the equivalent slice of the TEM; and in the third line, the slice of the PEM.

Figure 2 shows that the U-Net model is capable to predict the REE map only from the difference between the fixed image I_F and the registered image $\hat{I}_R$. However, we can also notice the presence of the brain anatomical structure that comes from the input images. This is mostly evident on the FLAIR and T2 images, that also had the worst results quantitatively. The main reason for that

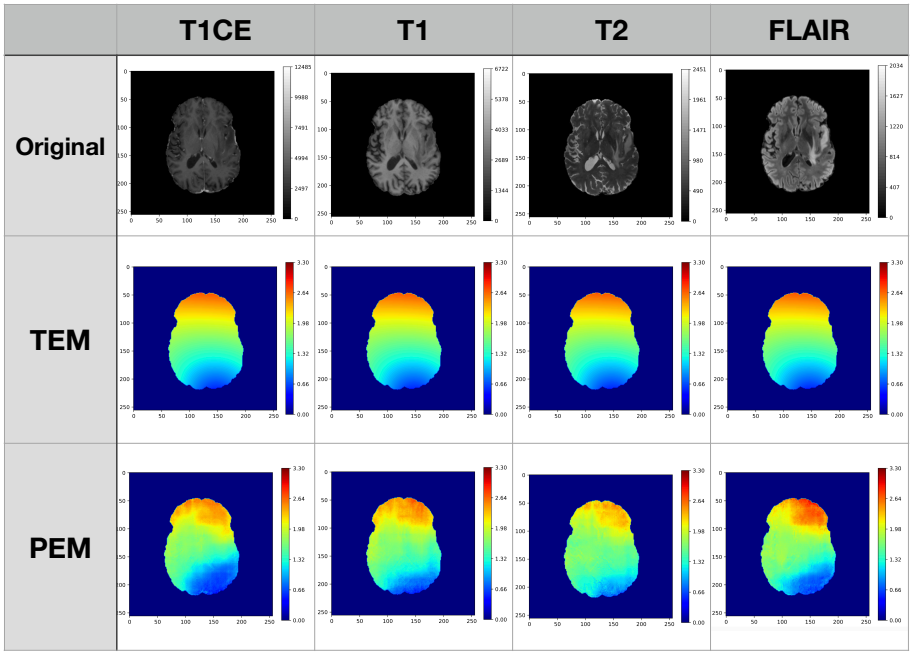

Fig. 2. Qualitative results for one patient of the test set. One slice was taken for each modality for illustration purpose. The first row shows the original image, the second, the target error map (TEM), and the third, the predicted error map (PEM).

could be the skip-connection layers, that transmit the higher level information on the upper levels of the U-Net. This makes the input information be propagated to the final layers of the network and eventually to the output image.

In spite of this element, we can state that our REE model is able to fulfill all of the requirements established in Section 2. We have an automatic method that is able to predict a 3D error map within a few seconds, thanks to the performance of the U-Net during inference. Moreover, we have a direct REE given in millimeters in the brain ROI. Ultimately, our quantitative and qualitative results evidence that, for all modalities, we have sub-voxel accurate REE values and PEMs that reproduce the same shape as the one given by the TEMs.

6 Conclusion

Deformable image registration error estimation is a challenging task. This work presents the capability of a 3D U-Net model to learn non-linear registration error maps for the case of brain MRI mono-modal non-linear registrations, with a sub-voxel-size accuracy. The learning procedure is completely automated as a result of our artificial B-spline transformations which enable the control of the amount of distortion to apply in order to simulate registration errors. Although

we evidence the presence of the brain anatomy on the PEM, the general aspect of the TEM is learned. In addition, the model keeps sub-voxel-size accuracy across a different dataset without re-training, and with transfer learning it improves its performance with less training epochs. In conclusion, in comparison with the traditional methods in the literature for REE (Dice score, TRE, Hausdorff distance, etc.), our model is the only one, to date, that is able to satisfy, for brain MRIs, all the criteria established in Section 2. The next steps will be to develop a similar method for multi-modal deformable registration error estimation, e.g. CT and MRI. Through the REE for each voxel, we aim for supporting highly precise diagnosis, surgery planning and anatomy visualization to improve the uncertainty management in neurosugery, hence leading to a better quality of care. Finally, this work also opens the door to close loop robotic control in neurosurgery by providing medical imaging errors linked to MRI.

Acknowledgments. This work was partially funded by a grant from ANRT Nº 2021/1427. This research study was conducted retrospectively using human subject data made available in open access [2], and provided by the CERMEP [13]. We thank the organizers of the BraTS-Reg challenges and the CERMEP for providing the data. Details regarding ethical approval are stated in the license attached with the open-access datasets and in [13].

Disclosure of Interests. Leandro Nascimento, Quentin François and Bertrand Duplat work as employees at Robeauté.

References

1. Alam, F., Rahman, S.U., Ullah, S., Gulati, K.: Medical image registration in image guided surgery: Issues, challenges and research opportunities. Biocybernetics and Biomedical Engineering **38**(1), 71–89 (2018). https://doi.org/10.1016/j.bbe.2017.10.001
2. Baheti, B., et al.: The brain tumor sequence registration challenge: Establishing correspondence between pre-operative and follow-up MRI scans of diffuse glioma patients (2021). https://doi.org/10.48550/ARXIV.2112.06979, arXiv:2112.06979
3. Bierbrier, J., Gueziri, H.E., Collins, D.: Estimating medical image registration error and confidence: A taxonomy and scoping review. Medical Image Analysis **81**, 102531 (2022). https://doi.org/10.1016/j.media.2022.102531
4. Boveiri, H., Khayami, R., Javidan, R., Mehdizadeh, A.: Medical image registration using deep neural networks: A comprehensive review. Computers & Electronical Engineering **87**, 106767 (2020). https://doi.org/10.1016/j.compeleceng.2020.106767
5. Cardoso, M.J., et al.: MONAI: An open-source framework for deep learning in healthcare (2022). https://doi.org/10.48550/arXiv.2211.02701, arXiv:2211.02701 [cs]
6. Castillo, R., et al.: A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets. Physics in Medicine & Biology **54**, 1849–1870 (2009). https://doi.org/10.1088/0031-9155/54/7/001

7. Eppenhof, K.J., Pluim, J.: Error estimation of deformable image registration of pulmonary CT scans using convolutional neural networks. Journal of Medical Imaging **5**(22), 024003 (2018). https://doi.org/10.1117/1.JMI.5.2.024003
8. Eppenhof, K., Pluim, J.: Pulmonary CT registration through supervised learning with convolutional neural networks. IEEE Transactions on Medical Imaging **38**(5), 1097–1105 (2019). https://doi.org/10.1109/TMI.2018.2878316
9. Geevarghese, R., O'Gorman Tuura, R., Lumsden, D., Samuel, M., Ashkan, K.: Registration accuracy of CT/MRI fusion for localisation of deep brain stimulation electrode position: An imaging study and systematic review. Stereotactic and Functional Neurosurgery **94**(33), 159–163 (2016). https://doi.org/10.1159/000446609
10. Kerfoot, E., et al.: Left-Ventricle Quantification Using Residual U-Net. In: Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges. vol. 11395, p. 371–380 (2019). https://doi.org/10.1007/978-3-030-12029-0˙40
11. Liu, J., Singh, G., Al'Aref, S., Lee, B., Oleru, O., Min, J.K., Dunham, S., Sabuncu, M.R., Mosadegh, B.: Image registration in medical robotics and intelligent systems: Fundamentals and applications. Advanced Intelligent Systems **1**(6), 1900048 (2019). https://doi.org/https://doi.org/10.1002/aisy.201900048
12. Murphy, K., et al.: Semi-automatic construction of reference standards for evaluation of image registration. Medical Image Analysis **15**(1), 71–84 (2011). https://doi.org/10.1016/j.media.2010.07.005
13. Mérida, I., Jung, J., Bouvard, S., Bouillot, C., Redouté, J., Hammers, A., Costes, N.: CERMEP-IDB-MRXFDG: A database of 37 normal adult human brain [18F]FDG PET, T1 and FLAIR MRI, and CT images available for research. EJNMMI Research **11**(1) (2021). https://doi.org/10.1186/s13550-021-00830-6
14. Pluim, J., Muenzing, S., Eppenhof, K., Murphy, K.: The truth is hard to make: Validation of medical image registration. In: ICPR. p. 2294–2300. IEEE (2016)
15. Rohlfing, T.: Image Similarity and Tissue Overlaps as Surrogates for Image Registration Accuracy: Widely Used but Unreliable. IEEE Transactions on Medical Imaging **31**, 153–163 (2012). https://doi.org/10.1109/TMI.2011.2163944
16. Rohlfing, T., Zahr, N., Sullivan, E., Pfefferbaum, A.: The SRI24 multichannel atlas of normal adult human brain structure. Humain Brain Mapping **31**(5), 798–819 (2009). https://doi.org/10.1002/hbm.20906
17. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241 (2015). https://doi.org/https://doi.org/10.1007/978-3-319-24574-4˙28
18. Saygili, G.: Predicting medical image registration error through independent directions. Signal, Image and Video Process. **15**(1), 223–230 (2021). https://doi.org/10.1007/s11760-020-01784-3
19. Saygili, G., Staring, M., Hendriks, E.A.: Confidence Estimation for Medical Image Registration Based On Stereo Confidences. IEEE Transactions on Medical Imaging **35**(2), 539–549 (2016). https://doi.org/10.1109/TMI.2015.2481609