Which Metrics Should Be Used in Non-linear Registration Evaluation?

Andre Santos Ribeiro[1], David J. Nutt[1], and John McGonigle[1]

Centre for Neuropsychopharmacology, Division of Brain Sciences,
Department of Medicine, Imperial College London, UK
j.mcgonigle@imperial.ac.uk

Abstract. Non-linear registration is an essential step in neuroimaging, influencing both structural and functional analyses. Although important, how different registration methods influence the results of these analyses is poorly known, with the metrics used to compare methods weakly justified. In this work we propose a framework to simulate true deformation fields derived from manually segmented volumes of interest. We test both state-of-the-art binary and non-binary, volumetric and surface -based metrics against these true deformation fields. Our results show that surface-based metrics are twice as sensitive as volume-based metrics, but are typically less used in non-linear registration evaluations. All analysed metrics poorly explained the true deformation field, with none explaining more than half the variance.

1 Introduction

Analysis of medical data often requires a precise alignment of subjects' scans with a common space or atlas. This alignment allows the comparison of data across time, subject, image type, and condition, while also allowing its segmentation into different anatomical regions, or find meaningful patterns between different groups [7]. Due to high inter-subject variability, a non-linear deformation of a subject's scan is typically applied to best conform this image to a reference standard. This type of deformation allows for complex modulation, such as elastic or fluid deformations.

The quantification of the accuracy of a non-linear registration is intrinsically complex. The main reason being the lack of a ground truth for the validation of different methods [8]. While in a rigid registration only three non-colinear landmarks are required, in non-linear registration a dense mesh of landmarks is needed [12]. To provide ground truth data the EMPIRE 10 challenge [9] provided correspondences of 100 annotated landmark pairs to distinguish between registration algorithms. However, the precision of the evaluation is still limited by the number of correspondence points. Furthermore, this study focused on intra-subject thoracic CT which may not fully apply across inter-subject studies or other modalities and structures.

Two other main approaches exist to evaluate non-linear registration methods: one based on the simulation of deformation fields; and the other in the evaluation of manually segmented regions in different subjects.

N. Navab et al. (Eds.): MICCAI 2015, Part II, LNCS 9350, pp. 388–395, 2015.
DOI: 10.1007/978-3-319-24571-3_47

In the former, several types of deformation field simulations are suggested, based in the deformation of control points, biomechanical models [2],[3],[13] or Jacobian maps [6],[11]. These techniques can struggle in that they are not capable of simulating the totality of inter-subject variation, nor the artefacts present in medical images.

In the latter, a database of different subjects with both an anatomical image and manually segmented volumes of interest (VOI) is typically used, in which each subject is registered to a randomly chosen subject, or average image, and the calculated deformation applied to the VOI map [4],[15]. In this way, the registered VOI map of the source subject can be compared with the VOIs of the target subject. Typically, to evaluate the different methods a labelling metric is used: Dice coefficients [7]; volume overlap [7],[14]; Jaccard index [10],[12]. This analysis assumes that the metrics evolve in the same way as the unknown deformation field. To our knowledge no study has been performed that effectively compare such metrics with a ground truth such as the true deformation field.

In this study we simulate pairs of deformation fields target images, and evaluate the relation between the different similarity metrics and the true deformation field. We explore what proportion of the variance of the ground truth can be explained by each metric, and which is more suitable when ground truth is not available (such as in inter-subject registration).

2 Methods

2.1 Data Acquisition

20 individual healthy T1 weighted brain images along with cortical and subcortical manual segmentations were obtained from the Open Access Series of Imaging Studies (OASIS) project[1]. The manually edited cortical labels follow sulcus landmarks according to the Desikan-Killiany-Tourville (DKT) protocol, with the sub-cortical labels segmented using the NVM software[2]. The resulting maps provide a maximum of 105 regions for each individual.

2.2 True Deformation Field Simulation

The first step of the framework is the simulation of the true deformation field. This field should attempt to maintain the characteristics of the native image such as the overall shape of the head, and absent of foldings. The former is required to remove the effect of global transformations such as affine or purely rigid transformations, while the latter is important as current methods limit the Jacobian determinants (here on referred to as Jacobian) to be positive to provide reasonable fields.

To generate the ground truth deformation fields used in this work the manually segmented images were used to provide an anatomically driven deformation.

[1] `http://mindboggle.info/data.html`

[2] `http://neuromorphometrics.org:8080/nvm/`

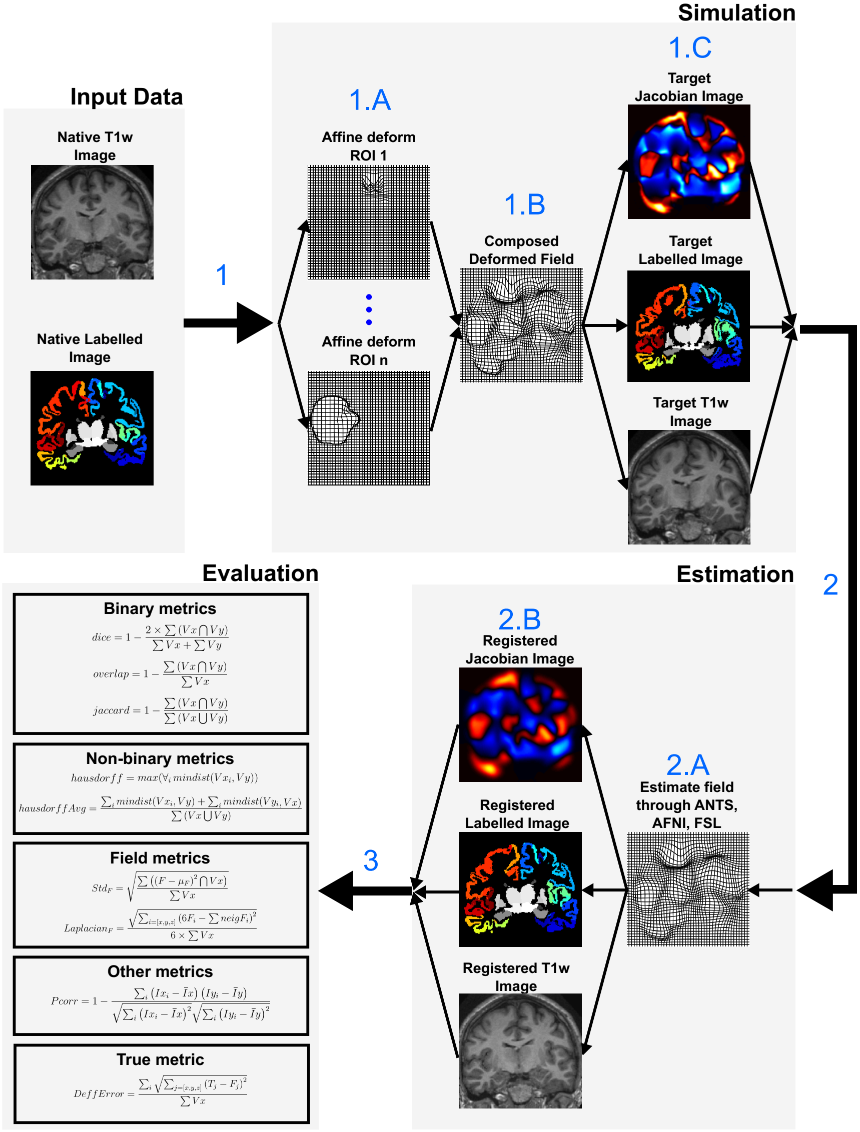
Input Data
Native T1w Image
Native Labelled Image
1
Simulation
1.A
Affine deform ROI 1
Affine deform ROI n
1.B
Composed Deformed Field
1.C
Target Jacobian Image
Target Labelled Image
Target T1w Image
2
Estimation
2.A
Estimate field through ANTS, AFNI, FSL
2.B
Registered Jacobian Image
Registered Labelled Image
Registered T1w Image
3
Evaluation
Binary metrics
$dice = 1 - \frac{2 \times \sum (Vx \bigcap Vy)}{\sum Vx + \sum Vy}$
$overlap = 1 - \frac{\sum (Vx \bigcap Vy)}{\sum Vx}$
$jaccard = 1 - \frac{\sum (Vx \bigcap Vy)}{\sum (Vx \bigcup Vy)}$
Non-binary metrics
$hausdorff = max(\forall_i \, mindist(Vx_i, Vy))$
$hausdorffAvg = \frac{\sum_i mindist(Vx_i, Vy) + \sum_i mindist(Vy_i, Vx)}{\sum (Vx \bigcup Vy)}$
Field metrics
$Std_F = \sqrt{\frac{\sum ((F - \mu_F)^2 \bigcap Vx)}{\sum Vx}}$
$Laplacian_F = \frac{\sqrt{\sum_{i=[x,y,z]} (6F_i - \sum neigF_i)^2}}{6 \times \sum Vx}$
Other metrics
$Pcorr = 1 - \frac{\sum_i (Ix_i - \bar{I}x)(Iy_i - \bar{I}y)}{\sqrt{\sum_i (Ix_i - \bar{I}x)^2}\sqrt{\sum_i (Iy_i - \bar{I}y)^2}}$
True metric
$DeffError = \frac{\sum_i \sqrt{\sum_{j=[x,y,z]} (T_j - F_j)^2}}{\sum Vx}$

Fig. 1. Scheme of the proposed evaluation framework. 1 - The manually labelled images are input to the simulation block to derive the ground truth deformation fields (1.B), target T1 images and target labelled images (1.C). To generate this field the labels are affine transformed (1.A), combined and regularized (1.B). 2 - The T1 native and target images are passed to the estimation block to estimate the deformation field (2.A), and registered T1 and labelled images (2.B). 3 - The registered and target images are finally given to the evaluation block to evaluate the different metrics against the true deformation field error metric - DeffError. $mindist(x_i, y)$ - minimum Euclidean distance between point x_i and set y. Note that the deformations shown were greatly enhanced for visualization purposes only.

As such, for each segmented image, each of the VOIs was allowed to randomly deform in an affine manner with 12 degrees of freedom. The affine transform of each VOI was normally distributed and centred at the identity transform. The variance for the translation, rotation, scale, and shear parameters were respectively, 1, 1, 0.1, and 0.1. Due to the free deformation of each VOI, the resulting field was regularized by locally smoothing the image such that every voxel presented a Jacobian within a range between 0.125 and 8, with a maximum deformation gradient of 0.5 in every direction. The deformation fields were further limited to retain skull invariance.

For each subject, 20 different ground truth deformation fields were produced, for a total of 400 simulations. For each deformation field a target T1 weighted image was created by deforming the original acquired image with the simulated field.

2.3 Estimation of the Deformation Field

To generate expected deformation fields through non-linear registration algorithms, ANTS-SyN, FSL-FNIRT, and AFNI-3dQwarp were used to register the original T1 to the simulated T1 image.

It should be noted that these methods are being used solely in the evaluation of the different metrics and no comparison between them is made in this work.

For each simulation, a registration was performed by each method and the resulting deformed T1 image and respective segmentation obtained, for a total of 1200 simulations. The Jacobian for each obtained deformation field was also calculated.

2.4 Evaluation of Registration Metrics

For each VOI of the simulations, each metric is calculated to investigate the similarity between the estimated and simulated fields.

Due to their popularity in the evaluation of registration the following metrics were analysed[3]: Dice coefficients (Figure 1 - Dice); Jaccard index (Figure 1 - Jaccard); target overlap (Figure 1 - overlap); Hausdorff distance (Figure 1 - Hausdorff); and Pearson correlation (Figure 1 - Pcorr). Due to the sensitivity of the Hausdorff distance [5], a modified average Hausdorff distance, also called Mean Absolute Distance - MAD [1], is further analysed (Figure 1 - Hausdorff average). This metric uses the average of all the closest distances from the registered to the target VOI, and the target to the registered VOI, instead of relying on a single point to derive the distance metric.

Although not generally used to compare different methods, field smoothness metrics are used to assess whether a particular method provides reasonable deformation fields. The standard deviation, and the Laplacian of both the deformation field, and the derived Jacobian (Figure 1 - std and laplacian) were further calculated. Landmark-based metrics were not included due to the lack of annotations (one-to-one correspondences) in the analysed dataset.

[3] All metrics were transformed such as the best value is 0.

Due to the highly folded nature of cortical gray matter, a good surface overlap is usually required. Volume-based metrics may not be suitable for this particular problem. Therefore, all the previously described metrics, with the exception of the Pearson correlation, were further applied to the surfaces of each VOI (i.e. the boundary voxels of the VOI).

For each metric (M) the resultant similarity values were randomly divided into 50 sub-groups (G) of $X = 2000$ VOIs, and linearly and non-linearly correlated with the true deformation field metric (T) (Figure 1 - DeffError) as described below:

1. Select a random subset of X VOIs from the total pool of VOIs;
2. Extract the analysed metrics (M), and the true metric (T), for each X;
3. Apply Pearson's linear correlation (r), and Spearman's rank correlation (ρ) between each M and T;
4. Square both r and ρ to obtain the proportion of shared variance in a linear fit and between the two ranked variables, respectively.
5. Repeat 1 to 4 for each of G sub-groups.
6. Compare the distribution of r^2 and ρ^2 for each M.

In this work the total number of VOIs was (number of subjects × number of simulations per subject × number of regions per subject × number of registration methods) $\approx$ 126,000.

Pearson's r was calculated as it provides a view of the linear relation between the metrics and the true field, while Spearman's ρ only assumes monotonicity and therefore extends the analysis to non-linear correlations.

A full schematic of the proposed framework is presented in Figure 1.

3 Results

In Figure 2 the explained variance (r^2 and ρ^2) of the ground truth deformation field for each of the analysed metrics is shown. All the metrics explain only a fraction (all below 50% for both r^2 and ρ^2) of the total variance present in the true deformation field. Specifically, the surface-based metrics were around twice as sensitive as their volume-based equivalents.

From the binary metrics the Jaccard index showed the highest r^2 (volume: $r^2 = 0.17 \pm 0.02$, surface: $r^2 = 0.32 \pm 0.02$), while in the non-binary metrics the Hausdorff average showed the highest r^2 (volume: $r^2 = 0.29 \pm 0.03$, surface: $r^2 = 0.40 \pm 0.04$). The original Hausdorff distance showed much lower results for both volume and surface metrics (volume: $r^2 = 0.11 \pm 0.02$, surface: $r^2 = 0.11 \pm 0.02$).

For the field metrics, all poorly explained the variance, with the best metric being the standard deviation of the deformation field ($r^2 = 0.08 \pm 0.01$). ρ^2 presented a similar trend to r^2, except between volume and surface versions of the Dice and Jaccard indexes. For these metrics the same ρ^2 was observed (volume: $\rho^2 = 0.22$, surface: $\rho^2 = 0.43$).

Paired t-tests, corrected for multiple comparisons (Bonferroni), were further performed between all metrics. This test was performed as the samples are not independent (i.e. the methods were applied over the same regions), making the distribution seen on Figure 2 only an indication of the difference between methods. The results were in agreement with the previous figure, suggesting that the Hausdorff average is more sensitive overall to changes in the deformation field. The only tests that did not presented significant differences were between the surface Jaccard and Pearson corr, and between the surface Dice and volume Hausdorff average.

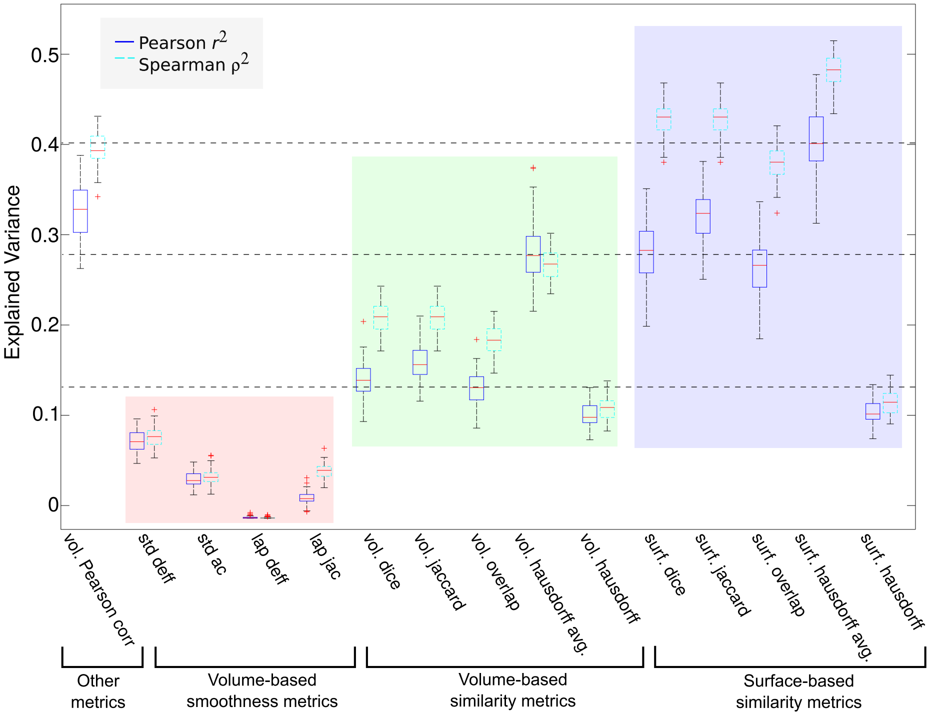

Fig. 2. Boxplot of the explained variance (derived through r^2 and ρ^2) of the true deformation field for each of the analysed metrics. Dark blue boxes - Pearson r^2; Light bashed blue boxes - Spearman ρ^2. Dashed lines serve as reference for the r^2 of the volume overlap metric presented in [7], the best volume-based metric, and the best surface-based metric.

4 Discussion

These results suggest that surface-based metrics are more sensitive to the true deformations than volume-based metrics, both for binary and non-binary metrics. One explanation for such behaviour is that the volume enclosed by a surface does not provide sufficient additional information regarding whether the region is overlapping or not, yet decreases the sensitivity of the metric.

Binary surface metrics are more prone to erroneous evaluations, as a simple shift of the two surfaces (still maintaining a high volume overlap) will lead to an almost null surface overlap.

Non-binary/distance metrics attempt to solve this problem by calculating the distance between the two sets, with the small shift identified either by the distance of each voxel in one set to the closest voxel in the other set, or simply by the center of mass of both sets. This leads to typically more robust metrics than binary ones, as is seen by the Hausdorff average distance. The original Hausdorff distance, however, showed low results for both volume and surface metrics, yet this was expected due to its sensitivity to outliers [5].

Although the Dice and Jaccard indexes differ in the r^2 they showed the same results for the ρ^2. This was also expected as they have the same monotonicity, yet differ in how they are normalized. As a linear trend is usually desirable, the Jaccard index should be used instead of the Dice coefficients.

Interestingly, in these results the Pearson correlation (applied only to the volume-based metrics) showed a much higher sensitivity compared to the volume-based metrics, and was similar to surface-based metrics. Yet this metric may be influenced by noise in the T1 images, and may not be suitable in registering images of different contrasts, such as in inter-modality analysis.

In general these results show that none of the metrics examined here explain more than 50% of the variance of the deformation field. Further, the best metric observed was the Hausdorff average distance (a modified version of the original Hausdorff distance). This poses the question of whether current assumptions based on these metrics hold true with regard to the evaluation of non-linear registration methods.

5 Conclusion

In this work we presented a framework to evaluate currently accepted metrics for comparison of non-linear registration algorithms, and showed that they perform poorly at estimating the true deformation field variance. These results suggest that current assumptions regarding “good” and “bad” methods may not be applicable. Furthermore, although surface-based metrics seem to perform better than volume-based metrics, they are typically less used in non-linear registration comparisons.

References

1. Babalola, K., Patenaude, B., Aljabar, P., Schnabel, J., Kennedy, D., Crum, W., Smith, S., Cootes, T., Jenkinson, M., Rueckert, D.: An evaluation of four automatic methods of segmenting the subcortical structures in the brain. NeuroImage 47, 1435–1447 (2009)
2. Camara, O., Scahill, R., Schnabel, J., Crum, W., Ridgway, G., Hill, D., Fox, N.: Accuracy assessment of global and local atrophy measurement techniques with realistic simulated longitudinal data. Med. Image Comput. Comput. Assist. Interv. 10(2), 785–792 (2007)

3. Camara, O., Schweiger, M., Scahill, R., Crum, W., Sneller, B., Schnabel, J., Ridgway, G., Cash, D., Hill, D., Fox, N.: Phenomenological model of diffuse global and regional atrophy using finite-element methods. IEEE Trans. Med. Imag. 25(11), 1417–1430 (2006)
4. Hellier, P., Barillot, C., Corouge, I., Gibaud, B., Le Goualher, G., Collins, D., Evans, A., Malandain, G., Ayache, N., [...], Johnson, H.: Retrospective evaluation of intersubject brain registration. IEEE Trans. Med. Imag. 22, 1120–1130 (2003)
5. Hyder, A.K., Shahbazian, E., Waltz, E. (eds.): Multisensor Fusion. Springer Science & Business Media (2002)
6. Karaali, B., Davatzikos, C.: Simulation of tissue atrophy using a topology preserving transformation model. IEEE Trans. Med. Imag. 25(5), 649–652 (2006)
7. Klein, A., Andersson, J., Ardekani, B., Ashburner, J., Avants, B., Chiang, M., Christensen, G., Louis Collinsi, D., Geef, J., [...], Parsey, R.: Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. NeuroImage 46, 786–802 (2009)
8. Murphy, K., van Ginneken, B., Klein, S., Staring, M., de Hoop, B.J., Viergever, M.A., Pluim, J.P.W.: Semi-automatic construction of reference standards for evaluation of image registration. Medical Image Analysis 15(1), 71–84 (2011)
9. Murphy, K., Ginneken, B., Reinhardt, J., Kabus, S., Ding, K., Deng, X., Cao, K., Du, K., Christensen, G., [...], Pluim, J.: Evaluation of Registration Methods on Thoracic CT: The EMPIRE10 Challenge. IEEE Trans. Med. Imag. 30(11), 1901–1920 (2011)
10. Ou, Y., Akbari, H., Bilello, M., Da, X., Davatzikos, C.: Comparative evaluation of registration algorithms in different brain databases with varying difficulty: results and insights. IEEE Trans. Med. Imag. 33(10), 2039–2065 (2014)
11. Pieperhoff, P., Sdmeyer, M., Hmke, L., Zilles, K., Schnitzler, A., Amunts, K.: Detection of structural changes of the human brain in longitudinally acquired MR images by deformation field morphometry: methodological analysis, validation and application. NeuroImage 43(2), 269–287 (2008)
12. Rohlfing, T.: Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. IEEE Trans. Med. Imag. 31(2), 153–163 (2012)
13. Schnabel, J., Tanner, C., Castellano-Smith, A., Degenhard, A., Leach, M., Hose, D., Hill, D., Hawkes, D.: Validation of nonrigid image registration using finite-element methods: application to breast MR images. IEEE Trans. Med. Imag. 22(2), 238–247 (2003)
14. Wu, G., Kim, M., Wang, Q., Shen, D.: S-HAMMER: hierarchical attribute-guided, symmetric diffeomorphic registration for MR brain images. Hum. Brain Mapp. 35(3), 1044–1060 (2014)
15. Yassa, M.A., Stark, C.E.L.: A quantitative evaluation of cross-participant registration techniques for MRI studies of the medial temporal lobe. NeuroImage 44, 319–327 (2009)