CONSISTENT AND ROBUST 4D WHOLE-BRAIN SEGMENTATION: APPLICATION TO TRAUMATIC BRAIN INJURY

Christian Ledig[⋆] *Wenzhe Shi*[⋆] *Antonious Makropoulos*[⋆] *Juha Koikkalainen*[†]
Rolf A. Heckemann[‡,◇,∗] *Alexander Hammers*[∗,◇] *Jyrki Lötjönen*[†] *Olli Tenovuo*[○] *Daniel Rueckert*[⋆]

[⋆] Department of Computing, Imperial College London, UK
[†] VTT Technical Research Centre of Finland, Tampere, Finland
[‡] MedTech West, Institute of Neuroscience and Physiology, University of Gothenburg, Sweden
[◇] Division of Brain Sciences, Faculty of Medicine, Imperial College London, UK
[∗] The Neurodis Foundation, CERMEP, Lyon, France
[○] Turku University Hospital, Turku, Finland

ABSTRACT

We propose a consistent approach to automatically segmenting longitudinal magnetic resonance scans of pathological brains. Using symmetric intra-subject registration, we align corresponding scans. In an expectation-maximization framework we exploit the availability of probabilistic segmentation estimates to perform a symmetric intensity normalisation. We introduce a novel technique to perform symmetric differential bias correction for images in presence of pathologies. To achieve a consistent multi-time-point segmentation, we propose a patch-based coupling term using a spatially and temporally varying Markov random field. We demonstrate the superior consistency of our method by segmenting repeat scans into 134 regions. Furthermore, the approach has been applied to segment baseline and six month follow-up scans from 56 patients who have sustained traumatic brain injury (TBI). We find significant correlations between regional atrophy rates and clinical outcome: Patients with poor outcome showed a much higher thalamic atrophy rate ($4.9 \pm 3.4\%$) than patients with favourable outcome ($0.6 \pm 1.9\%$).

***Index Terms*—** brain MRI, longitudinal segmentation, EM optimisation, temporal consistency, traumatic brain injury

1. INTRODUCTION

The worldwide incidence of traumatic brain injury (TBI) cases is estimated at 6.8 million annually, representing a substantial public health burden [1]. Although the need for reliable assessment tools was already expressed 30 years ago [2], prognostic assessment remains a challenge, and standardised models to predict outcome of patients with moderate TBI remain unavailable [1]. To assist the development of such models, an accurate assessment and understanding of the structural changes happening during and after TBI is crucial. In [3] indications for brain volume loss following a TBI have been identified using tissue segmentation techniques on structural magnetic resonance (MR) scans and diffusion tensor imaging.

Several methods have been published that address the problem of quantifying brain changes over time based on MR scans. One popular representative is "CLASSIC" [4], a method that uses adaptive clustering for tissue segmentation of MR scans taken at multiple time-points. Another longitudinal segmentation method, presented by Lorenzo et al. [5], is based on expectation-maximization (EM) optimisation and applied to 4D cardiac sequences. This asymmetric approach uses stationary temporal Markov random fields (MRF) and affine alignment with a probabilistic spatiotemporal atlas.

Several other methods have been proposed that address primarily brain tissue segmentation such as [6] which uses level sets. Wolz et al. [7] simultaneously segment the hippocampus of longitudinal scans using graph cuts. Recent research suggests that a fully symmetric process can significantly reduce bias, e.g. in atrophy measurement [8]. In addition the advantages of a longitudinal segmentation of several time points over multiple, single time point segmentation are well established [4, 7].

In this work we propose a novel method based on [9] for consistent segmentation of serial images into many anatomical regions, rather than tissue classes. Different from previous methods, we perform image alignment, intensity normalisation and differential bias field correction in a symmetric fashion. To achieve consistency, we present an approach to perform differential bias field correction in the presence of significant pathology in the images. We further introduce a novel way to determine a spatially and temporally varying, fully data driven temporal coupling of the longitudinal segmentation based on MRF. To our knowledge, our method is the first consistent segmentation approach that segments longitudinal whole-brain scans into a large number of anatomical structures while being robust to pathology. Our experiments show quantitative evidence for improved segmentation consistency while maintaining high accuracy. We also demonstrate on a cohort of TBI patients that the proposed method is robust and has the potential to quantify imaging biomarkers, specifically atrophy, that correlate well with clinical outcome.

2. METHOD

2.1. Spatial priors and symmetric longitudinal image alignment

Assuming preprocessed, brain extracted and bias corrected images we derive subject specific probabilistic labels from M available atlases using a multi-atlas label propagation approach. For the unsegmented images at n time points, $\mathbf{I^t}$ at $t = 0 \dots n-1$, we calculate individual transformations $\phi_m^t,\ m = 1 \dots M$ by registering M manually labelled atlases to $\mathbf{I^t}$. For the image alignment, we employ MAPER [10], an approach that incorporates tissue probability maps into the registration and relies on a non-rigid registration

This work is partially funded under the 7th Framework Programme by the European Commision (http://cordis.europa.eu/ist/).

based on free-form deformations (FFD) [11, 12]. Using the resulting transformations we map the M label maps to the space of image $\mathbf{I^t}$ and create, for each time point t, a probabilistic atlas using a locally weighted multi-atlas fusion strategy based on a Gaussian weighted sum of squared differences (GSSD). In the following, π_{ik}^t denotes the probabilistic prior for label k at voxel i and at time point t.

As shown in [8, 13], asymmetric image registration can introduce significant bias into longitudinal image analyses. We thus choose a symmetric registration setup, 6 degree-of-freedom global rigid registration similar to [13], and transform images, $\mathbf{I^t}$, and priors, $\pi^{\mathbf{t}}$, of all time points, to a common intermediate space [8].

2.2. 4D EM segmentation

We then estimate the probabilistic segmentation of the aligned multi-time-point scans, residing in their common rigid intermediate space, simultaneously. Based on the subject and time-point specific priors, we employ a 4D extension of the EM algorithm [9], which was originally presented by van Leemput et al. [14]. We employ separate Gaussian distributions, described by Φ^t, for each time point. Spatial smoothness of the final segmentations is enforced through a global and stationary MRF [15]. To also enforce longitudinal consistency, we propose novel spatially and temporally varying constraints that are dependent on the patch similarity between voxels. This enables us to formulate the strength of the temporal coupling as a function of the local intensity similarity of the aligned images. This is crucial both to allow for flexibility in pathological regions and to enforce consistency in regions with no or very little change. We assume a Gaussian mixture model for the probability of observing an intensity y_i^t at a voxel i at time point t

$$f(y_i^t|\Phi^{t,(q)}) = \sum_k f(y_i^t|z_i^t = k, \Phi^{t,(q)}) \frac{\pi_{ik}^t e^{-U_{\text{MRF}}(k|p_{\mathcal{N}_i}^{(q)})}}{\sum_{j=1}^K \pi_{ij}^t e^{-U_{\text{MRF}}(j|p_{\mathcal{N}_i}^{(q)})}} \tag{1}$$

where $\mathcal{N}_i$ is the set of first-order neighbours of voxel i, k a label index and z_i^t the labelling of voxel i at time point t.

We represent a MRF energy U_{MRF} that is dependent on the probabilistic label estimates $\mathbf{p}$ at iteration q as:

$$U_{\text{MRF}}(k|p_{\mathcal{N}_i}^{(q)}) = \sum_{j=1}^K G_{kj} v_j \ , \ \ v_k = \sum_{l \in \mathcal{N}_i^{x,y,z}} p_{lk}^t + \sum_{l \in \mathcal{N}_i^t} \gamma_i^{t,l} p_{ik}^l \tag{2}$$

Here $\mathbf{G}$ denotes a $K \times K$ matrix defining the connectivity between class k and j. The column vector $\mathbf{v}$ of size K denotes the accumulated class probabilities in the 8-neighborhood $\mathcal{N}_i$ of voxel i. The class probabilities of adjacent time points are weighted by temporal weighting factors, $\gamma_i^{t,t\pm1}$. The following sections describe the symmetric intensity normalisation, the calculation of $\gamma_i^{t,t\pm1}$, and the differential bias correction. These steps are performed once in the given order after the first EM steps.

2.3. Intensity normalisation

To enable an accurate comparison of corresponding patches at different time points, we employ a symmetric intensity normalisation. We perform this normalisation based on the probabilistic segmentation estimates obtained in the first EM steps. Based on these current probabilistic label estimates, we calculate for multiple time points n the mean intensities μ_s^t $(s = 1 \dots 5)$ for the five global structures white matter (WM), deep grey matter (dGM), cortical grey matter (cGM), ventricles (vent), background/cerebrospinal fluid (CSF). Mean intensity values for these five "super structures" are obtained by merging the labels of the individual structures of which they are composed. We then estimate all μ_s^t based on fuzzy label estimates and normalise corresponding structures at all time points n symmetrically. Specifically, we normalise intensities at time point t and from super structure s with $\sqrt[n]{\prod_{l=0\dots n-1} \mu_s^l}/\mu_s^t$. Again, we consider the voxel-wise probabilistic label estimates when normalising the intensity of a certain voxel.

2.4. Voxel-wise pathology index for temporal coupling

To reduce segmentation bias, we propose coupling the segmentation of corresponding images from multiple time points through a spatially and temporally varying MRF. While the spatial MRF weights are stationary, we determine the temporal weights based on the similarity of temporally adjacent intensity normalised images. At time point t we calculate local patch similarities between time point t and $t+1$, or t and $t-1$ respectively. Specifically, we calculate the Gaussian weighted sum of absolute differences (GSAD) where the kernel width determines the "patch size". From the GSAD value at a given voxel $w_i^{t,t\pm1}$, we derive the pathology indices $\gamma_i^{t,t\pm1}$ for the temporal coupling based on a logistic function of the form:

$$\gamma_i^{t,t\pm1} = c - \frac{c}{1.0 + a \cdot \exp(-b \cdot w_i^{t,t\pm1})} \tag{3}$$

Here c determines the maximal value of γ_i while a and b determine the shape of the S-curve. $\gamma_i^{t,t\pm1}$ is close to c in areas where the data suggests no presence of pathologies or significant structural changes (small w_i) and close to 0 otherwise.

2.5. Differential Bias Correction

Lewis et al. [16] propose a straightforward way to estimate the differential bias fields of aligned longitudinal image pairs by applying a median filter to the difference image, calculated between log-transformed images. This approach is extended in [8] to multiple time points. However, the basic assumption of this approach, that the differential bias field is given as the low frequency component of the difference image, does not hold for pathological images. A violation of this assumption is exemplified by the image pair in Figure 1. Here, significantly enlarged ventricles or pathological changes appear as low frequency intensity change and are thus falsely regarded as bias field by the approach presented in [8, 16].

We present a novel way to adapt bias field estimation that increases robustness towards pathologies. We extend the formulation presented in [8] and estimate pairwise differential bias fields $\mathbf{dbf}_{\text{poly}}^{t,t'}$, for time points t, t', based on a polynomial model. After estimating a differential bias field $\mathbf{dbf}^{t,t'}$ based on the difference image, we fit a second order polynomial to $\mathbf{dbf}^{t,t'}$ using weighted least squares optimisation to calculate $\mathbf{dbf}_{\text{poly}}^{t,t'}$. As weights we exploit the quantities γ that were previously derived from patch similarities. Employing the pathology indices γ, we can reduce the influence of pathological areas to the bias field estimation. We then perform a symmetric bias field correction similar to [8] using all pairwise estimates $\mathbf{dbf}_{\text{poly}}^{t,t'}$.

3. EXPERIMENTAL RESULTS

3.1. Material and Parameters

As atlases we used the data set provided as part of the "MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling". It

consists of 35 T_1-weighted MR brain images with isotropic voxel sizes of $1 \times 1 \times 1$ mm from the OASIS database [17]. The corresponding 134 anatomical labels have been created manually by experts.[1] The scans have been acquired from 30 subjects (20 female, 10 male, mean±SD age: 34.3±20.7 years) where 5 subjects have been scanned twice. The duplicates were acquired within 90 days of the original scan [17].

In a novel data cohort of patients who had sustained TBI, we segmented 56 intra-subject pairs of MR brain scans acquired six months apart (mean 6.7±1.2). Acquisition parameters were identical for all TBI scans: resolution $240 \times 256 \times 170$; spacing $1 \times 1 \times 1$ mm; sagittal slices.

We evaluate the proposed method in two ways: First, we evaluate consistency and accuracy in the 5 pairs of intra-subject repeat scans. Second, we apply our method to the TBI cohort to evaluate clinical relevance of the novel segmentations.

To segment all five pairs of repeat scans, which are part of the atlas cohort, we used the subset of 15 atlases that served as training data in the Grand Challenge. These $M = 15$ scans were distinct from the 5×2 repeat datasets. For the segmentation of the TBI data we used all $M = 30$ atlases, excluding the five repeat scans.

Parameters: Parameters were set mainly based on results from previous experiments. For the GSSD in the locally weighted atlas fusion we set $\sigma_{\text{GSSD}} = 2.5$. As spatial MRF weights, we set entry (k, k) of the connectivity matrix $\mathbf{G}$ to 0. Entry (k, j) was set to 1.0 if structures k and j shared a boundary, and to 1.5 otherwise. For the logistic function in Sec. 2.4, we heuristically chose $c = 2.0$ in the order of the spatial MRF weights, $a = 100$ and $b = 5.0/\text{median}[w_i^{t,t+1}]$. b was modelled dependent on the median of the occurring GSAD values w_i ($\sigma_{\text{GSAD}} = 1.0$) within the images. This adapts the logistic function to different similarity levels of image pairs in a data-driven fashion. As in [16], a median filter of kernel size 5 was used to estimate the differential bias field.

3.2. Test-retest reliability

We employed the intraclass correlation coefficient (ICC) to quantify test-retest reliability [18]. The ICCs for both manual and automated segmentations are summarized in Table 1. Overall, the proposed method yields a significantly higher ICC than the reference methods.

Table 1. ICC (mean±SD) based on label volumes averaged over 5 image pairs for all 36 subcortical and all 98 cortical regions obtained manually, with MAPER, and our method without and with the temporal constraints as described in Sec. 2.3-2.5. Bold = significantly higher than others (paired, two-sided Student's t-test, $p < 0.005$).

ICC	manual	MAPER	no coupling	proposed
subcort.	.80±.28	.95±0.08	.96±.07	**.98±.03**
cort.	.78±.25	.95±0.09	.94±.07	**.98±.02**

3.3. Subject-atlas label overlaps

Enforcing identical segmentations at both time points leads to high ICC values, but high ICCs do not necessarily reflect a high segmentation accuracy which needs to be assessed with measures of overlap. The accuracy of segmenting both time points with and without the temporal constraints was assessed quantitatively. Table 2 shows overlap results based on the Dice coefficient (DC). We calculated the DC on the 10 repeat scans by comparing to the gold standard. Enforcing temporal consistency does not alter segmentation accuracy.

[1] provided by Neuromorphometrics, Inc. under academic subscription (http://Neuromorphometrics.com/).

Table 2. Average DC values comparing the proposed approach without or with temporal constraints to the gold-standard segmentations of the 10 images of the 5 pairs of repeat scans.

	36 subcort.	98 cort.	134 structures
no coupling	81.5	72.5	74.9
temp. coupling	81.6	72.4	74.9

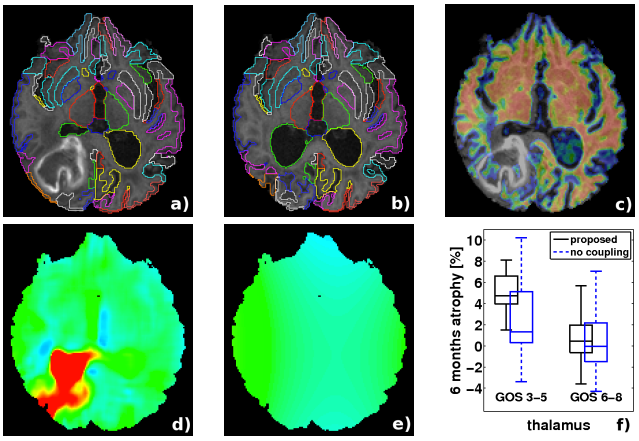

Fig. 1. Visual results for a TBI subject (GOS=3) with significant pathologies obtained with the proposed method. (a) T_1-weighted baseline and (b) 7.9 month follow-up scan with overlaid segmentation, (c) temporal weights according to Eq. 3 (high weights in red), (d) differential bias field estimated as presented in [16], (e) proposed polynomial fitted bias field. (f) Boxplot of thalamic volume change calculated with (black) / without (blue) the proposed temporal coupling as described in Sec. 2.3-2.5 for GOS groups 3-5 vs. 6-8.

3.4. Longitudinal analysis of TBI patients

To evaluate the proposed method in a realistic setting, we determined the size change of individual regions over time and assessed their potential as biomarkers by comparison with a standard clinical outcome score, the Glasgow Outcome Scale (GOS, [2]). We preprocessed all 56×2 TBI brain images using the N4 algorithm to correct for intensity inhomogeneities [19] and employed an in-house method for brain extraction. We then applied the method to the 56 image pairs. Clinical GOS scores were available (GOS 3: 4 subjects, GOS 4: 3, GOS 5: 4, GOS 6: 11, GOS 7: 15, GOS 8: 19). Figure 1 shows visual results calculated on one of the most pathological subjects.

As a preliminary experiment that underlines the potential of the proposed method, we followed [2] and split the TBI subjects into two groups with GOS 3-5 (severe or lower moderate disability) and GOS 6-8 (upper moderate disability, good recovery). We calculated atrophy rates for all 134 structures and employed a two-sided unpaired Student's t-test to identify structures for which the size change rate was significantly different between the groups. We combined

paired structures appearing both in the left and right brain hemisphere. The four most significant subcortical structures were thalamus ($p < 10^{-6}$), lateral ventricles ($p < 10^{-4}$), CSF ($p \approx 0.001$) and pallidum ($p \approx 0.01$). Without temporal coupling group differences were far less marked, e.g. for the thalamus the significance of the atrophy rate difference was only $p \approx 0.05$. Using MAPER we were not able to detect significant group differences for thalamic atrophy (p=0.65). The boxplot shown in Figure 1 illustrates the substantial reduction of the size of the thalamus (GOS 3-5: $4.9 \pm 3.4\%$ vs. GOS 6-8: $0.6 \pm 1.9\%$). In contrast, ventricles expand faster, due to brain atrophy in more severely injured patients ($19.6 \pm 21.3\%$ vs. $1.6 \pm 9.0\%$). Such a clear group separation is crucial for outcome prediction: we ran 1000 iterations of a 10-fold cross validation separating GOS groups 3-5 and 6-8 based on thalamic atrophy using linear discriminant analysis (LDA). While the classification accuracy without temporal coupling was 61% (Specificity: 66%, Sensitivity: 41%) we observed a high accuracy of 87% (88%, 82%) using the proposed method with temporal coupling.

4. CONCLUSION

We have presented a novel method that enables the symmetric and consistent segmentation of pathological serially acquired MR brain images into a large number of anatomical structures. We found that the segmentation consistency was significantly improved in comparison to manual and state-of-the-art procedures. By application to a novel cohort of patients with TBI, we were able to show that our method allows the derivation of more discriminative prognostic markers in serial-imaging. Our findings also indicate the method's potential to assist in the challenging task of monitoring the development of pathological conditions. In future work, we intend to investigate the potential of the derived metrics as markers of TBI disease progression in greater detail. A further promising avenue is the integration of a deformation model to estimate the structural disease progress over time.

5. REFERENCES

[1] A. Irimia, B. Wang, S. R. Aylward, M. W. Prastawa, D. F. Pace, G. Gerig, D. A. Hovda, R. Kikinis, P. M. Vespa, and J. D. Van Horn, "Neuroimaging of structural pathology and connectomics in traumatic brain injury: Toward personalized outcome prediction," *NeuroImage: Clinical*, vol. 1, no. 1, pp. 1–17, 2012.

[2] B Jennett, J. Snoek, M. R. Bond, and N. Brooks, "Disability after severe head injury: observations on the use of the Glasgow Outcome Scale," *Journal of Neurology, Neurosurgery, and Psychiatry*, vol. 44, pp. 285–293, 1981.

[3] B. B. Bendlin, M. L. Ries, M. Lazar, A. Alexander, R. Dempsey, H. Rowley, J. Sherman, and S. C. Johnson, "Longitudinal changes in patients with traumatic brain injury assessed with diffusion-tensor and volumetric imaging," *NeuroImage*, vol. 42, no. 2, pp. 503–514, 2008.

[4] Z. Xue, D. Shen, and C. Davatzikos, "CLASSIC: consistent longitudinal alignment and segmentation for serial image computing," *NeuroImage*, vol. 30, no. 2, pp. 388–399, 2006.

[5] M Lorenzo-Valdes, G. I. Sanchez-Ortiz, A. G. Elkington, R. H. Mohiaddin, and D. Rueckert, "Segmentation of 4D cardiac MR images using a probabilistic atlas and the EM algorithm," *Medical Image Analysis*, vol. 8, no. 3, pp. 255–265, 2004.

[6] L. Wang, S. Feng, P. Yap, J. H. Gilmore, W. Lin, and D. Shen, "Accurate and Consistent 4D Segmentation of Serial Infant Brain MR Images," 2011, vol. 7012, pp. 93–101, Lecture Notes in Computer Science, MBIA 2011.

[7] R. Wolz, R. A. Heckemann, P. Aljabar, J. V. Hajnal, A. Hammers, J. Lötjönen, and D. Rueckert, "Measurement of hippocampal atrophy using 4D graph-cut segmentation: Application to adni," *NeuroImage*, vol. 52, no. 1, pp. 109–118, 2010.

[8] K. K. Leung, G. R. Ridgway, S. Ourselin, and N. C. Fox, "Consistent multi-time-point brain atrophy estimation from the boundary shift integral," *NeuroImage*, vol. 59, no. 4, pp. 3995–4005, 2012.

[9] C. Ledig, R. Wolz, P. Aljabar, J. Lötjönen, R. A. Heckemann, A. Hammers, and D. Rueckert, "Multi-class brain segmentation using atlas propagation and EM-based refinement," *Proceedings of ISBI 2012*, pp. 896–899, 2012.

[10] R. A. Heckemann, S. Keihaninejad, P. Aljabar, D. Rueckert, J. V. Hajnal, and A. Hammers, "Improving intersubject image registration using tissue-class information benefits robustness and accuracy of multi-atlas based anatomical segmentation," *NeuroImage*, vol. 51, no. 1, pp. 221–227, 2010.

[11] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, and D. J. Leach, M. O.and Hawkes, "Nonrigid registration using free-form deformations: Application to breast MR images," *IEEE TMI*, vol. 18, no. 8, pp. 712–721, 1999.

[12] M. Modat, G. R. Ridgway, Z. A. Taylor, M. Lehmann, J. Barnes, D. J. Hawkes, N. C. Fox, and S. Ourselin, "Fast free-form deformation using graphics processing units," *Computer Methods and Programs in Biomedicine*, vol. 98, no. 3, pp. 278–284, 2010.

[13] P. A. Yushkevich, B. B. Avants, S. R. Das, J. Pluta, M. Altinay, and C. Craige, "Bias in estimation of hippocampal atrophy using deformation-based morphometry arises from asymmetric global normalization: An illustration in ADNI 3T MRI data," *NeuroImage*, vol. 50, no. 2, pp. 434–445, 2010.

[14] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "Automated model-based tissue classification of MR images of the brain," *IEEE TMI*, vol. 18, no. 10, pp. 897–908, 1999.

[15] M. J. Cardoso, M. J. Clarkson, G. R. Ridgway, M. Modat, N. C. Fox, and S. Ourselin, "Load: A locally adaptive cortical segmentation algorithm," *NeuroImage*, vol. 56, no. 3, pp. 1386–1397, 2011.

[16] E. B. Lewis and N. C. Fox, "Correction of differential intensity inhomogeneity in longitudinal MR images," *NeuroImage*, vol. 58, no. 23, pp. 75–83, 2004.

[17] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open access series of imaging studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults," *Journal of Cognitive Neuroscience*, vol. 19, no. 9, pp. 1498–1507, 2007.

[18] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: uses in assessing rater reliability," *Psychological Bulletin*, vol. 86, pp. 420 – 428, 1979.

[19] N.J. Tustison, B.B. Avants, P.A. Cook, Yuanjie Zheng, A. Egan, P.A. Yushkevich, and J.C. Gee, "N4ITK: Improved N3 bias correction," *IEEE TMI*, vol. 29, no. 6, pp. 1310–1320, 2010.