Medical Image Analysis 21 (2015) 40–58

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Robust whole-brain segmentation: Application to traumatic brain injury

Christian Ledig [a,*], Rolf A. Heckemann [e,b,c], Alexander Hammers [b,c,d], Juan Carlos Lopez [e], Virginia F.J. Newcombe [g], Antonios Makropoulos [a], Jyrki Lötjönen [f], David K. Menon [g], Daniel Rueckert [a]

[a] Biomedical Image Analysis Group, Department of Computing, Imperial College London, UK
[b] The Neurodis Foundation, CERMEP, Lyon, France
[c] Division of Brain Sciences, Faculty of Medicine, Imperial College London, UK
[d] The PET Centre, Division of Imaging Sciences and Biomedical Engineering Kings College London, St Thomas Hospital, London, UK
[e] MedTech West, Institute of Neuroscience and Physiology, University of Gothenburg, Sweden
[f] Knowledge Intensive Services, VTT Technical Research Centre of Finland, Tampere, Finland
[g] University Division of Anaesthesia, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK

ARTICLE INFO

Article history:
Received 8 April 2014
Received in revised form 14 December 2014
Accepted 15 December 2014
Available online 24 December 2014

Keywords:
Traumatic brain injury
Magnetic resonance imaging
Multi-atlas segmentation
Brain image segmentation
Expectation–maximisation

ABSTRACT

We propose a framework for the robust and fully-automatic segmentation of magnetic resonance (MR) brain images called "Multi-Atlas Label Propagation with Expectation–Maximisation based refinement" (MALP-EM). The presented approach is based on a robust registration approach (MAPER), highly performant label fusion (joint label fusion) and intensity-based label refinement using EM. We further adapt this framework to be applicable for the segmentation of brain images with gross changes in anatomy. We propose to account for consistent registration errors by relaxing anatomical priors obtained by multi-atlas propagation and a weighting scheme to locally combine anatomical atlas priors and intensity-refined posterior probabilities. The method is evaluated on a benchmark dataset used in a recent MICCAI segmentation challenge. In this context we show that MALP-EM is competitive for the segmentation of MR brain scans of healthy adults when compared to state-of-the-art automatic labelling techniques. To demonstrate the versatility of the proposed approach, we employed MALP-EM to segment 125 MR brain images into 134 regions from subjects who had sustained traumatic brain injury (TBI). We employ a protocol to assess segmentation quality if no manual reference labels are available. Based on this protocol, three independent, blinded raters confirmed on 13 MR brain scans with pathology that MALP-EM is superior to established label fusion techniques. We visually confirm the robustness of our segmentation approach on the full cohort and investigate the potential of derived symmetry-based imaging biomarkers that correlate with and predict clinically relevant variables in TBI such as the Marshall Classification (MC) or Glasgow Outcome Score (GOS). Specifically, we show that we are able to stratify TBI patients with favourable outcomes from non-favourable outcomes with 64.7% accuracy using acute-phase MR images and 66.8% accuracy using follow-up MR images. Furthermore, we are able to differentiate subjects with the presence of a mass lesion or midline shift from those with diffuse brain injury with 76.0% accuracy. The thalamus, putamen, pallidum and hippocampus are particularly affected. Their involvement predicts TBI disease progression.

1. Introduction

With an estimated annual global incidence of 6.8 million cases, traumatic brain injury (TBI) imposes a significant burden on patients, their families, and health services (Irimia et al., 2012). Usually caused by sudden acceleration/deceleration or focal impacts, the lesions caused can be focal as in the case of contusions or more diffuse (diffuse axonal injury (DAI)) (Meythaler et al., 2001; Warner et al., 2010b). It is common for patients to have a combination of these. After the acute injury secondary processes including complex metabolic cascades, alterations in cerebral blood flow and raised intracranial pressure may occur contributing to the burden of injury. It is well recognised that complex pathophysiological processes including secondary Wallerian-type degeneration continue to occur months to years after the initial insult (Meythaler et al., 2001; Ding et al., 2008; Warner et al., 2010a). In order to improve treatment stratification and patient outcomes, as well as more accurately predict outcome, we need

* Corresponding author at: Department of Computing, Imperial College London, 180 Queen's Gate SW7 2AZ, UK.
E-mail address: christian.ledig@imperial.ac.uk (C. Ledig).

http://dx.doi.org/10.1016/j.media.2014.12.003
1361-8415/

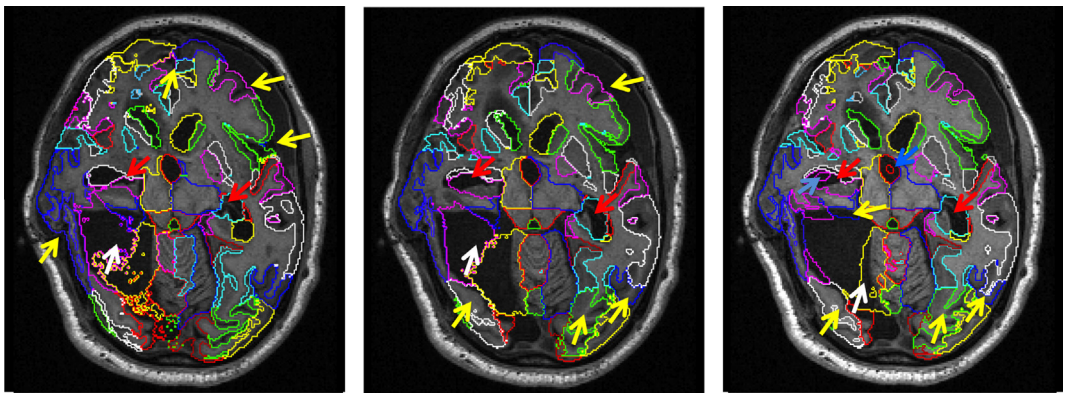

Fig. 1. Example of segmentation results obtained on a subject with highly abnormal brain configuration. Segmentations calculated with MAPER using majority voting (left, Heckemann et al. (2010)) and SyN (Avants et al., 2008) from the ANTs toolkit using either majority voting (middle) or the joint label fusion (right, Wang et al. (2013)). Red arrows: substantial oversegmentation of the hippocampus; yellow arrows: inaccurate cortex segmentation due to gross brain deformation; blue arrows: ventricles incorrectly labelled as background; white arrows: region of missing tissue prohibits reasonable one-to-one mapping of the atlases. Segmentation contours are shown in a colour scheme that provides good colour contrast between neighbouring structures. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

to better understand the complexity and heterogeneity of TBI both in the acute and chronic stages.

Although patterns of abnormalities have been shown to be predictors of outcome, such use of imaging data is mainly based on expert interpretation of visually inspected X-ray computed tomography (CT) images. Standard models to predict the outcome of TBI patients remain unavailable (Irimia et al., 2012). To assist the understanding of TBI disease progression, accurate quantitative assessment of the structural changes occurring during and after TBI is crucial. Segmentation of structural magnetic resonance (MR) images offers a potential way to gain more insight. For example, in Bendlin et al. (2008) brain volume loss following TBI has been identified using tissue segmentation techniques on structural MR images (MRIs) and diffusion tensor imaging (DTI). In Irimia et al. (2011) an intra-patient time point comparison has been performed on three representative TBI patients using semi-automatic methods for tissue and lesion classification and 3D model generation. Ramlackhansingh et al. (2011) used structural MRI and positron emission tomography (PET) to demonstrate inflammatory processes that remain active for months or years following brain trauma. An overview of existing structural MRI findings in mild TBI is provided in Shenton et al. (2012). Most of the few existing studies (Strangman et al., 2010; Warner et al., 2010a,b) that analyse structural morphometric measures are based on the segmentation techniques available in FreeSurfer (Fischl et al., 2002) and investigate small patient cohorts (Warner et al., 2010a,b). In Warner et al. (2010b) the authors investigate the correlation between structural brain atrophy of 25 patients with DAI and functional outcome. Several brain structures showed significantly increased structural atrophy when compared to a control group 8 months post injury (Warner et al., 2010b). In Strangman et al. (2010), fifty patients that sustained TBI were enrolled in a memory rehabilitation program and their individual progress recorded. The study investigated the predictive value of structural brain volumes with respect to the outcome of the rehabilitation (Strangman et al., 2010). Both studies (Strangman et al., 2010; Warner et al., 2010b) identified several structures, including the thalamus and hippocampus that are particularly affected by TBI and are of significant value when predicting clinical outcome.

The automatic structural segmentation of MR brain scans of TBI patients remains, however, a difficult endeavour as most existing methods lack robustness towards TBI-related changes in anatomy (Irimia et al., 2011, 2012). In the acute phase contusions, the presence of blood, hydrocephalus and/or oedema can greatly affect the ability to accurately segment a brain. In more chronic scans gliosis and atrophy are also often poorly dealt with using currently available segmentation methods. It is this high variability and extent of brain change following a moderate or severe TBI that makes the segmentation task so demanding. An exemplar subject with highly abnormal brain configuration is shown with overlaid automatic segmentations in Fig. 1 to illustrate the difficulty of the segmentation task.

A popular class of automatic segmentation algorithms is multi-atlas label propagation with origins in Rohlfing et al. (2004b) and Heckemann et al. (2006). In multi-atlas label propagation, each of the semi-automatically or completely manually annotated atlases is individually aligned with the unsegmented target image. The propagated segmentations are then merged into a consensus label at each voxel in the target image. Voxelwise label conflicts can be resolved using either simple, unweighted approaches (Rohlfing et al., 2004a; Heckemann et al., 2006; Aljabar et al., 2009) or by weighting individual contributions locally based on the intensity information from the atlas and target images (Artaechevarria et al., 2009; Sabuncu et al., 2010). Alternative fusion strategies based on statistical optimisation have been proposed, with the most popular representative being STAPLE (Warfield et al., 2004) and its modifications (Asman and Landman, 2011, 2013; Landman et al., 2012; Cardoso et al., 2013a). A more detailed overview of atlas-based methods is provided by Cabezas et al. (2011). A particular successful strategy called joint label fusion was recently proposed by Wang et al. (2013). In this state-of-the-art approach, as evaluated in (Landman and Warfield, 2012), segmentation bias is reduced by estimating joint segmentation errors of different atlas pairs (Wang et al., 2013).

Atlas propagation techniques rely on the accurate registration of the atlas and unsegmented MR image to determine the spatial transformation of the atlas labels into the target space. This can be difficult if the target image differs from the available atlases due to the presence of pathology.

Recently, Liu et al. (2014) presented a promising approach based on low-rank matrix decomposition to register multiple images of TBI patients simultaneously to a reference image. In Niethammer et al. (2011), the authors formulated a geometric metamorphosis model to address the challenges arising in the registration of images from TBI, tumour or stroke patients. Other approaches iteratively register and segment the images simultaneously to identify missing correspondences (Periaswamy and Farid, 2006; Chitphakdithai and Duncan, 2010). Based on a seed,

Zacharaki et al. (2009) simulated tumour growth in an atlas image before registering it to a tumour patient. Next to this, more standard approaches often rely on a mask to ignore abnormal regions during the registration process (Brett et al., 2001; Stefanescu et al., 2004; Andersen et al., 2010). However, methods that rely on strong prior knowledge such as masks or tumour growth models are in general not applicable for the segmentation of subjects with heterogeneous pathologies as they are often present in TBI patients. Bauer et al. (2013) provide a comprehensive overview on image analysis in the context of brain tumours.

In addition, there have been several methods proposed to address the challenge of registering abnormal adult brain images of Alzheimer's Disease (AD) patients. In Wolz et al. (2010a) a manifold of atlases and unsegmented MR images is learned to robustly propagate the atlas label sets to all unsegmented images within the manifold. Another approach, "Multi-Atlas Propagation with Enhanced Registration" (MAPER) (Heckemann et al., 2010, 2011), employs automatically calculated tissue classification into the registration process to enable robust image alignment, even if the target image shows severe brain atrophy. Recently, methods based on nonlocal patch-based label fusion have been proposed (Coupé et al., 2011; Rousseau et al., 2011) that rely on affine alignment with a template library and thus relax the requirement for accurate nonrigid registrations. Patch-based methods have been developed further, often with focus on a particular application. For example Tong et al. (2013) used dictionary learning and sparse coding for hippocampal segmentation in patients with AD, while Wang et al. (2014) applied patch-driven level sets to tissue segmentation in neonates. However, while in AD brain changes are consistent with disease progression, MR brain images of patients with TBI can show inconsistent and gross pathological change as demonstrated in Fig. 1. Fig. 1 further shows that established registration techniques such as MAPER (Heckemann et al., 2010) or SyN (Avants et al., 2008) struggle to establish a plausible mapping between the available atlas images and an image of an abnormal brain. Both the presence of gross deformation and the potential absence of brain tissue prevent an accurate anatomical correspondence estimation. Even the application of a state-of-the-art label fusion technique (Wang et al., 2013) is not able to correct the substantial and consistent errors of alignment.

Atlas-based segmentation can be further improved by incorporating intensity information from the unseen image through a Gaussian mixture model (GMM) (Van Leemput et al., 1999; Fischl et al., 2002). The resulting optimisation problem is often solved using expectation–maximisation (EM) (Van Leemput et al., 1999; Lötjönen et al., 2010; Cardoso et al., 2011; Cardoso et al., 2013b; Ledig et al., 2012) or graph cuts (van der Lijn et al., 2008; Wolz et al., 2010b). While the EM approach enables simultaneous probabilistic segmentation of multiple brain structures, graph-cut based methods yield binary labels for individual structures. Also approaches that propagate atlas labels over a graph (Wolz et al., 2010a; Cardoso et al., 2013c) or within clusters (Ribbens et al., 2014) enjoy increasing attention in the community. In the context of brain tissue segmentation it has been further shown that the relaxation of anatomical atlas priors can improve segmentation quality (Cardoso et al., 2011, 2013b).

As an alternative or complement to either approach, Wang et al. (2011) proposed to use machine learning techniques to learn systematic segmentation errors that are then corrected in a post-processing step.

Atlas-based approaches require a number of brain atlases that, in the ideal case, have been generated by expert manual delineation. In this work we use an atlas set that was the basis of a recent whole-brain segmentation challenge (MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling, Landman and Warfield (2012)). We can thus provide a direct comparison to the methods that were evaluated in this competition. In general, other whole-brain atlases are equally suitable for the proposed approach.

The main motivation for our work was to devise a fully automatic and robust segmentation method that allows accurate measurement of various brain structures in MR images in the presence of severe pathologies, as exemplified in Fig. 1. Specifically, we are interested in the analysis of MR brain scans acquired of patients that had sustained traumatic brain injury.

The method described in this work addresses a key need in the management of TBI, that was identified by Irimia et al. (2012): "[...] the key methodological hurdle that must be overcome in order to make structural neuroimaging a powerful tool for predicting TBI outcome is the current paucity of automated image processing methods that can allow researchers to analyse large numbers of TBI CT/MRI volumes without the need for excessive user input or intervention."

We pursue this objective on three levels. First, we combine the best features of state-of-the-art atlas-based segmentation tools into a new framework, MALP-EM, by building on MAPER and adding the benefits of joint label fusion and an intensity-based refinement using EM. Second, we adapt this method for the challenges posed by highly abnormal brain configurations. To achieve this, we use a prior relaxation scheme that corrects anatomical atlas priors in regions where accurate alignment of the images is impossible due to missing brain tissue or severe deformation. We further employ a data-driven and locally adaptive weighting scheme to combine anatomical atlas prior probabilities and intensity-refined posterior probabilities for maximum benefit. Third, we use the modified MALP-EM algorithm to segment 125 MR brain scans of a heterogeneous population of 101 subjects who had sustained TBI. To assess segmentation accuracy on this TBI cohort, we devised a specific protocol that was independently followed by three blinded raters. This protocol enables an expert to rate hippocampus, thalamus, putamen and occipital pole segmentation quality in the absence of manual reference segmentations.

We then derive volumetric biomarkers based on an index that quantifies asymmetry between structures appearing both in the left and right brain hemisphere (absolute asymmetry index, AAI). We show the potential of single time-point MR imaging based variables to correlate with and predict outcome-relevant clinical variables.

2. Material and methods

2.1. Material

2.1.1. Traumatic brain injury (TBI) database

We obtained T_1-weighted MR brain images of 101 TBI patients provided through the University Division of Anaesthesia, Cambridge University, UK. The images were acquired using an MPRAGE sequence on a Siemens MAGNETOM TrioTim Syngo with parameters: TR 2300 ms, TE 2.98 ms, TI 900 ms, flip angle 9°, matrix size $256 \times 240 \times 176$ and an isotropic voxel size of $1.0\,\text{mm} \times 1.0\,\text{mm} \times 1.0\,\text{mm}$. Patients underwent MRI in the acute-phase, as part of the follow-up, or both. As only 24 patients had MR scans at both time points, we focus on a cross-sectional analysis in this work. In total we had 125 datasets available, including 61 acute and 64 follow-up MR brain scans. Information about the patients' gender and age distributions, the elapsed time between scanning date and injury and the Glasgow Coma Score (GCS) is summarised in Table 1. The GCS is a clinical score that quantifies a patient's level of consciousness in the acute stage of the injury (Teasdale and Jennett, 1974). The datasets were further grouped by clinical scores using Marshall Classification (MC, Marshall et al. (1991)) and the Glasgow Outcome Score (GOS,

Table 1
Overview of the available MR images with patient gender, patient age, scan time relative to injury, and Glasgow Come Score (GCS) (Teasdale and Jennett, 1974).

	Acute-phase MR image	Follow-up MR image
# of subjects	61	64
Gender (# male/ # female)	48/13	40/24
Age (mean ± standard deviation)	36.6 ± 14.9 years	36.1 ± 14.9 years
Time since injury (mean±standard deviation)	3.7 ± 4.2 days	10.0 ± 7.2 months
GCS (median [min; max])	5 [3; 15]	7 [3; 15]

Table 2
Clinical variables of the 61 acute-phase TBI images (MC, GOS) and the 64 follow-up MRIs (GOS). See Appendices B and C for details of the definition of the Marshall Classification and Glasgow Outcome Scale. n/a: not available.

MC	n/a	DI I (1)	DI II (2)	DI III (3)	DI IV (4)	EML (5)	NEML (6)
# of subjects per group (acute-phase)	3	4	29	2	0	16	7
GOS	n/a	D (1)	VS (2)	SD (3)	MD (4)	GR (5)	
# of subjects per group (acute-phase)	6	8	1	18	17	11	
# of subjects per group (follow-up)	0	0	0	21	28	15	

Jennett and Bond (1975)), as shown in Table 2. MC is a score based on the worst acute computed tomography (CT) image within 24 h of injury. MC takes into account brain pathology such as lesion load, the presence of oedema and midline shift caused by the injury. In contrast, the GOS is a clinical measure categorising the outcome of TBI and is assessed 6 months after injury or once the TBI outcome is considered stable. In Section 3.2.3 we describe correlations of these clinical scores with asymmetry biomarkers derived from brain MR scans. Details of the definition of the Marshall Classification and Glasgow Outcome Scale are provided in Appendices B and C.

2.1.2. Atlases

The atlas cohort used in this study consisted of 35 manually annotated MR brain images of 30 subjects of the OASIS database (Marcus et al., 2007). The manual segmentation into 138 anatomical structures has been carried out by experts according to publicly available protocols[1] and were provided by Neuromorphometrics, Inc. (http://Neuromorphometrics.com/, last accessed: 8 December 2014) under academic subscription. The same atlas cohort was used in the recent "MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling" (Landman and Warfield, 2012). As suggested by Landman and Warfield (2012), the small regions (order of 100 voxels each) "vessel" and "cerebral exterior" were excluded in our experiments in both the left and the right hemisphere, so that we effectively investigated 134 structures. In five of the subjects, repeat scans were acquired in a second session within 90 days of the original scan (Marcus et al., 2007).

The 134 atlas labels comprise 63 anatomical structures which have symmetric counterparts in their opposite hemisphere, in total 126 labels (see Appendix A). The remaining eight unpaired structures are: 3rd ventricle, 4th ventricle, brain stem, CSF, optic chiasm, cerebellar vermal lobules I–V, cerebellar vermal lobules VI–VII, cerebellar vermal lobules VIII–X.

[1] http://www.cma.mgh.harvard.edu/manuals/segmentation/ and http://www.braincolor.org (last accessed: 8 December 2014).

2.2. Multi-Atlas Label Propagation with Expectation–Maximisation based refinement (MALP-EM)

2.2.1. Notation

To present our framework called "Multi-Atlas Label Propagation with Expectation–Maximisation based refinement" (MALP-EM), we employ the following notation:

We label an unsegmented T_1-weighted MR image $\mathbf{I_u}$ into $K = 134$ structural regions. We index $\mathbf{I_u} = \{y_1, y_2, \ldots, y_n\}$ where $y_i \in \mathbb{R}$, with $i = 1, \ldots, n$, denotes the intensity value of the ith voxel. To incorporate expert knowledge into the segmentation process, we employ M manually annotated brain atlases denoted by $\mathbf{A_m}$ with $m = 1, \ldots, M$. ϕ_m denotes the calculated transformation from the atlas space of $\mathbf{A_m}$ in the coordinate system of $\mathbf{I_u}$ and $\mathbf{A_m^\phi}$ denotes the propagated atlas. Employing multi-atlas label fusion we then create a subject specific probabilistic segmentation $\Pi = \{\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_n\}$ that is, based on image intensities, relaxed to Π^R and used as spatial prior in the EM framework. Using an EM approach, we then estimate an intensity-refined probabilistic segmentation of $\mathbf{I_u}$, denoted by $Z = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\}$. Here $\boldsymbol{\pi}_i, \boldsymbol{\pi}_i^R$ and $\mathbf{z}_i$ are vectors of size K and the k^{th} component represents the probability that a voxel i belongs to a region k. Thus Π denotes the subject specific probabilistic segmentation before intensity-based refinement and Z after intensity-based refinement respectively. We abbreviate the normal distribution $\mathcal{N}(\mu_k, \sigma_k)$ with $\mathcal{N}_k$ where μ_k is the mean and σ_k the standard deviation of the intensity distribution within label k.

2.2.2. Registration and label fusion

For each unsegmented image $\mathbf{I_u}$, we obtain M transformations ϕ_m by registering M manually generated atlases to the coordinate space of $\mathbf{I_u}$. In this study we employ the enhanced registration approach that has been developed as part of MAPER (Heckemann et al., 2010). MAPER incorporates tissue probability maps into a nonrigid registration scheme based on free-form deformations (Rueckert et al., 1999; Modat et al., 2010).

A probabilistic map $\boldsymbol{\pi_k}$ of each anatomical structure k is then formed from the M transformed atlases $\mathbf{A_m^\phi}$ using the joint label fusion strategy presented by Wang et al. (2013). We employed the publicly available implementation at https://www.nitrc.org/projects/picsl_malf/ (Version 1.2, last accessed: 8 December 2014) with standard parameters. We have used joint label fusion as it has been shown to be a leading label fusion technique (Landman and Warfield, 2012). This procedure is illustrated in Fig. 2.

2.2.3. Relaxation of probabilistic priors

Segmentation refinement based on image intensities relies heavily on the probabilistic priors Π. As a consequence, the segmentation at voxel i cannot be refined if all atlases agreed on a certain label k (π_{ik} close to or equal to 1). In general, this is a sensible constraint assuming that at least a subset of the propagated atlases votes for the correct label. When segmenting MR scans showing significant pathologies or abnormalities, however, there is evidence that this assumption is no longer justified. Especially in regions that undergo large deformations, for example the inferior lateral ventricles when the target region is enlarged due to injury, swelling or atrophy, we observed unanimous bias in all individual segmentations. Label fusion approaches can thus return substantial mislabelling of subcortical grey matter structures such as the hippocampus, as well as of cortical regions. This problem is illustrated in the top left image of Fig. 3. The intensity-based

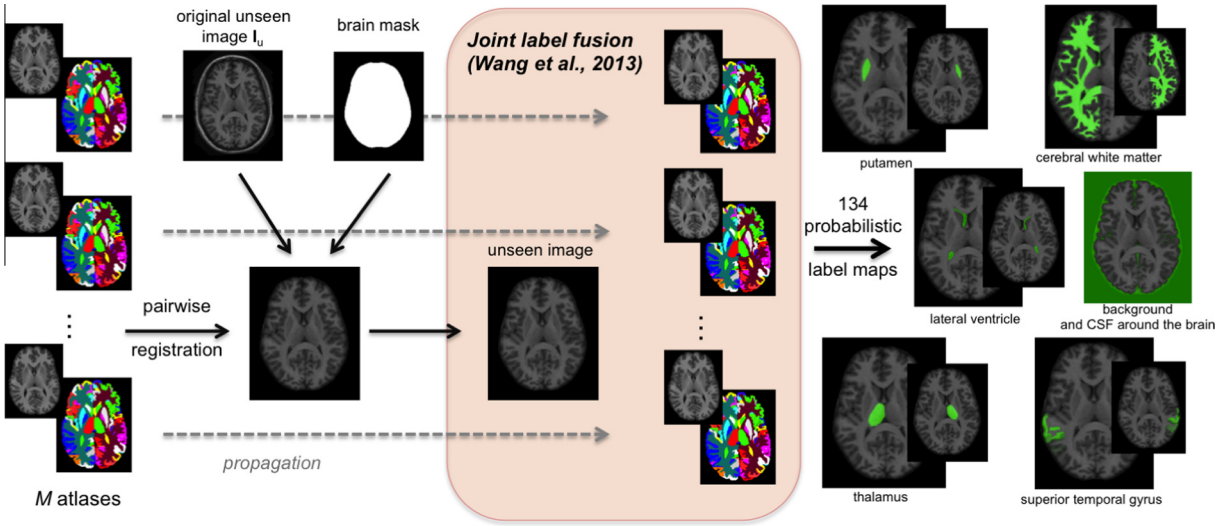

Fig. 2. Schematic process of the calculation of the subject specific spatial priors Π for an unsegmented target image $\mathbf{I_u}$. After brain extraction and bias correction, the available M atlases are registered to the space of $\mathbf{I_u}$. Using these transformations, label maps and corresponding T_1-weighted MR images are mapped to the space of $\mathbf{I_u}$. The label maps are then averaged into probabilistic priors for the individual structures using the joint label fusion (Wang et al., 2013). A subset of the 134 probabilistic labels is shown in green. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

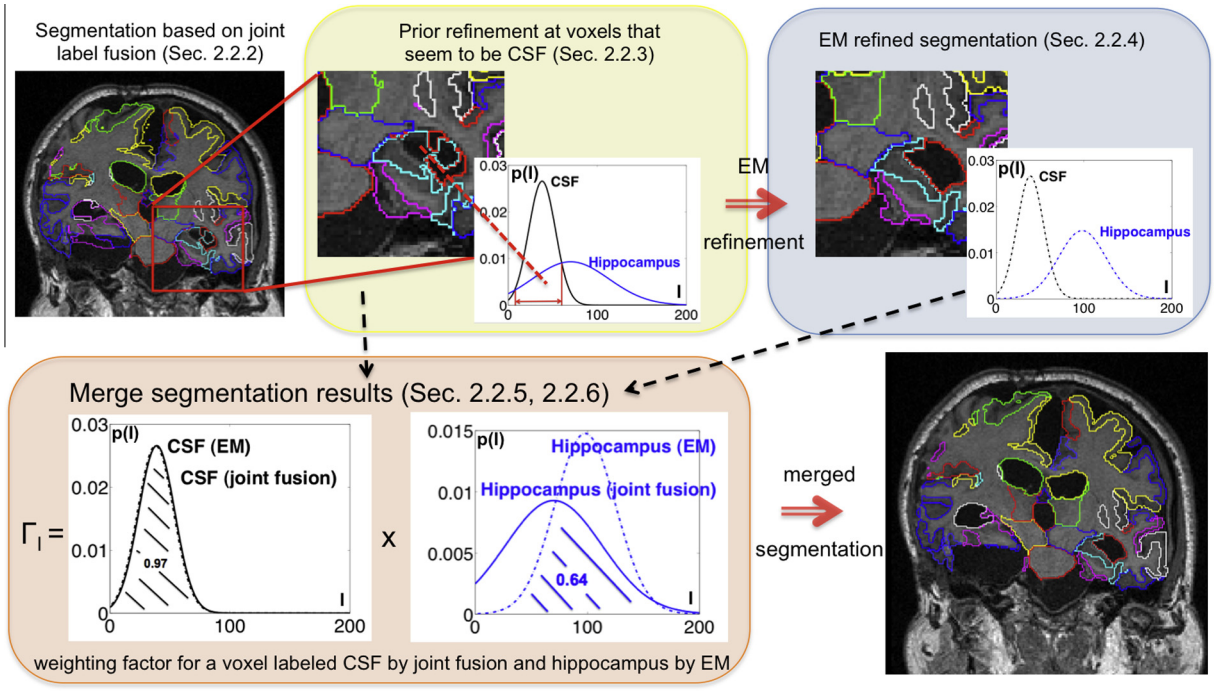

Fig. 3. Schematic process of the segmentation refinement using prior relaxation, EM-optimisation and spatially weighted combination of probabilistic label maps on the example of the hippocampal region. If registration consistently fails joint label fusion tends to label a significant number of voxels belonging to the inferior lateral ventricle as hippocampus. These wrongly labelled low-intensity voxels lead to a high variance of the estimated intensity distribution within the hippocampus label (top left). The red interval (top left intensity distribution) indicates for which voxels prior relaxation will be carried out. EM-refinement then allows correction of the mislabeled CSF voxels leading to a sharper intensity distribution within the hippocampus (top right). The segmentations obtained using label fusion and EM-optimisation are finally merged into a consensus segmentation (bottom right). This combination is based on spatially varying weights that are calculated based on the overlap of intra-label intensity distributions (bottom left). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

EM-refinement is restricted by these high prior label probabilities Π and thus unable to entirely correct this systematic error.

We tackle this problem by calculating relaxed priors Π^R from the label probabilities Π. Specifically we relax the probabilistic priors based on the probabilistic label fusion estimates and the actual image intensities. Assuming a Gaussian distribution, we estimate a common parameter set $(\mu_{\text{CSF-like}}, \sigma_{\text{CSF-like}})$ of eight "CSF-like" structures. We define the set of structures denoted as "CSF-like" as {background (essentially external CSF), 3rd ventricle, 4th ventricle, CSF, right/left inferior lateral ventricle, right/left lateral ventricle}. We furthermore estimate for each structure k an individual parameter set (μ_k, σ_k) based on the probabilistic prior segmentation Π:

$$\mu_k = \frac{\sum_i \pi_{ik} y_i}{\sum_i \pi_{ik}}, \qquad \sigma_k = \sqrt{\frac{\sum_i \pi_{ik} (y_i - \mu_k)^2}{\sum_i \pi_{ik}}} \tag{1}$$

In the actual relaxation step we then redistribute a fraction α_{ik} of the prior probability, π_{ik}, from a structure k to one of the eight CSF-like structures k_{CSF}. At an image voxel i, we determine k_{CSF} as the CSF-like structure with the highest prior probability or the label that is spatially closest to the voxel i:

$$k_{\mathrm{CSF}} = \begin{cases} \underset{k \text{ is CSF-like}}{\arg\max}\ \pi_{ik} & \text{if } \pi_{ik} \neq 0 \text{ for at least one k} \in \text{CSF-like} \\ \underset{k \text{ is CSF-like}}{\arg\min}\ d(k,i) & \text{else} \end{cases} \tag{2}$$

Here $d(k,i)$ denotes the Euclidean distance of voxel i to the closest point in label k. We then calculate α_{ik} based on the probability that the voxel with intensity y_i comes either from the intensity distribution $\mathcal{N}_k^{\Pi}$ estimated in label k or $\mathcal{N}_{\mathrm{CSF-like}}^{\Pi}$ accordingly. We set α to:

$$\alpha_{ik} = \begin{cases} 0 & \text{if } \mathcal{N}_k^{\Pi}(y_i) \geqslant \mathcal{N}_{\mathrm{CSF-like}}^{\Pi}(y_i) \\ \max(0, \min(0.5 - \pi_{ik_{\mathrm{CSF}}}, \pi_{ik})) & \text{else} \end{cases} \tag{3}$$

We thus do not allow the CSF-like label to exceed 50% probability and correct only voxels that have a higher probability of belonging to the CSF-like label k_{CSF}, as defined in Eq. (2), than to k. Finally the relaxed prior probability Π^R is calculated as:

$$\pi_{ik}^R = \begin{cases} \pi_{ik} + \sum_{l \neq k_{\mathrm{CSF}}} \alpha_{il} & \text{if } k = k_{\mathrm{CSF}} \\ \pi_{ik} - \alpha_{ik} & \text{else} \end{cases} \tag{4}$$

The two images on the left in Fig. 4 show where the voxel priors are relaxed by the proposed method, both for a distinctly abnormal brain and for a brain with a normal configuration.

The whole pipeline including registration, joint label fusion and prior relaxation is illustrated in Figs. 2 and 3.

2.2.4. Intensity-based label refinement through expectation–maximisation

We refine the relaxed probabilistic priors Π^R, calculated using the joint label fusion and the prior relaxation as described in Section 2.2.3, based on the observed intensities of $\mathbf{I_u}$ by employing the widely used EM-optimisation presented by Van Leemput et al. (1999). To be consistent with the published literature (Wells et al., 1996; Van Leemput et al., 1999; Zhang et al., 2001; Cardoso et al., 2011), we assume a normal distribution of the log-transformed intensities of voxels of a given label k. Each class k is then described by its mean μ_k and standard deviation σ_k.

The complete model parameters are $\{(\mu_1, \sigma_1), (\mu_2, \sigma_2), \ldots, (\mu_K, \sigma_K)\}$. A full description of the model can be found in Appendix D.

Smoothness of the final segmentation is enforced with a global and stationary Markov Random Field (MRF), which is integrated using the mean field approximation (Zhang, 1992), following the example of Van Leemput et al. (1999) and Cardoso et al. (2011).

In order to increase samples for small non-cortical brain structures and thus increase the robustness of the parameter estimate, we model symmetric brain structures (e.g. hippocampus left/right)

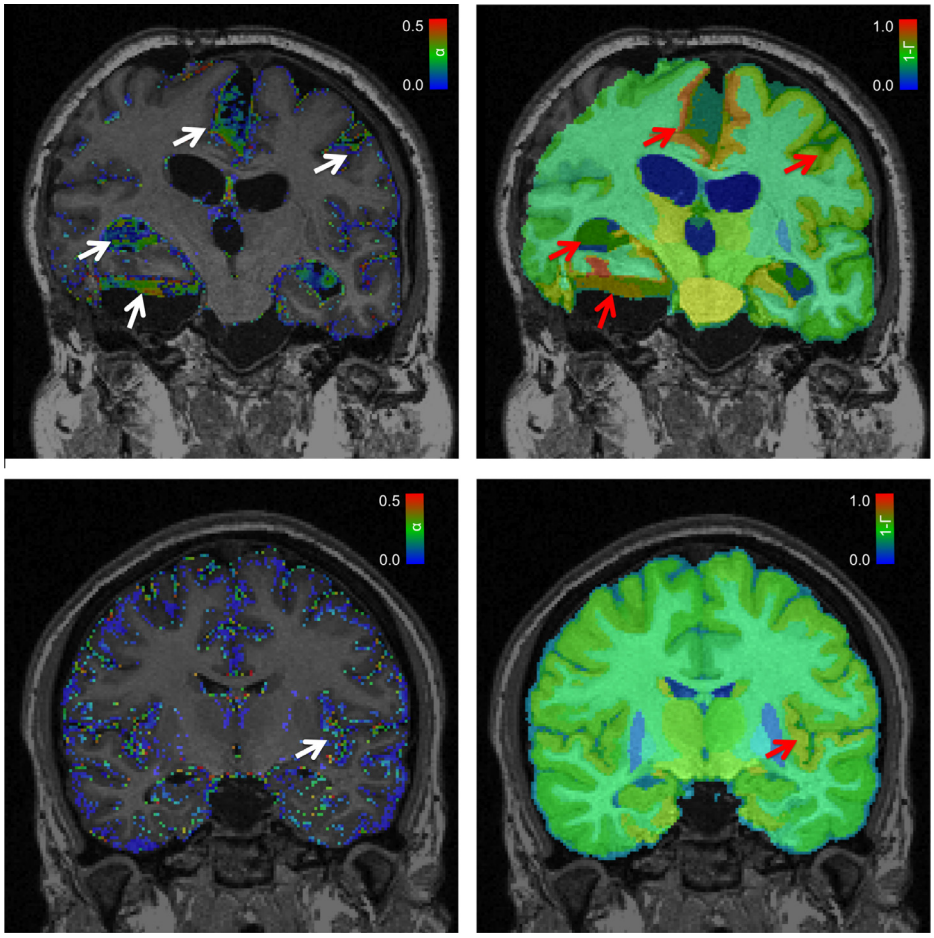

Fig. 4. Illustration of the relaxation weights (left column) and the spatially varying combination weights as $1 - \Gamma_i$ (right column) for the two example subjects in Fig. 9 (top row, subject with an abnormal brain configuration) and Fig. 10 (bottom row, subject with a close-to-normal brain configuration). We note that in cortical regions and regions where the label fusion fails, both increased prior relaxation (white arrows) and combination weights (red arrows) favouring EM-based segmentations are apparent. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

with a single Gaussian distribution. This also allows a more consistent segmentation and thus subsequent volume comparison or symmetry inference. Furthermore, we model all voxels in cortical brain structures with a single Gaussian distribution.

2.2.5. Weighting scheme for merging fusion-based and EM-based segmentations

Adjacent brain structures often have very similar intensity distributions, which makes an intensity-based refinement of their shared boundary difficult. Examples are adjacent cortical brain regions or the boundary between the hippocampus and the amygdala.

Furthermore, the presented intensity-based optimisation approach tends to calculate well-separated intensity distributions by reducing the intraclass variances (σ_k^2) within labels (Ledig et al., 2012). However, this often degrades segmentation results as it does not necessarily reflect manual labelling protocols. For example, the boundary between thalamus and adjacent white matter is determined by geometric characteristics, rather than by intensity characteristics (Hammers et al., 2003). As the boundaries of subcortical structures are not only defined by intensities, their intensity profile tends to have a wider spread (larger σ_k). This has been observed previously in Ledig et al. (2012) for subcortical structures such as the thalamus, caudate or putamen. For instance, the standard EM-optimisation is likely to relabel high intensity, 'thalamus voxels' in the vicinity of the thalamus/white matter boundary as 'white matter'. This results in a reduced intraclass variance for the thalamus class and higher model likelihood. The segmentation accuracy is reduced, however, as the estimated boundary does not represent the manual protocol of the expert rater.

While intensity-based EM-refinement often provides little or no value for subcortical regions of healthy brains, it is a powerful technique to correct consistent registration failures, e.g. in the hippocampal region or in cortical regions. Here, the label estimate of a certain class k often contains several types of brain tissue resulting in large intraclass variance. These intensity distributions can effectively be optimised using the intensity-based refinement as described in Section 2.2.4.

We propose not to rely exclusively on either joint label fusion or EM-refined fusion, but to combine their probabilistic estimates into a common segmentation. Here, we describe the simple global formulation of the proposed model before introducing the spatially variant extension in Section 2.2.6.

Depending on a global weighting factor $\Gamma \in [0;1]$ we combine the spatial priors obtained by joint label fusion, Π, and the posteriors calculated based on intensity-based EM-optimisation, Z.

Specifically we combine the final posterior probability z_{ik} and the spatial prior π_{ik} to calculate a new probabilistic estimate z_{ik}^{merged} as:

$$z_{ik}^{merged} = (1.0 - \Gamma) z_{ik} + \Gamma \pi_{ik} \tag{5}$$

2.2.6. Locally varying weighting parameter Γ

A straightforward extension of this formulation is to model the weighting parameter Γ dependent on the spatial position i. This allows a variable weight for the contribution of either registration-driven (multi-atlas label propagation) or intensity-driven (EM-refinement) label estimates in the final segmentation. With the known characteristics of EM-refined results (cf. Section 2.2.5) in mind, we aim to formulate a model, which favours the geometry-driven and registration-based priors over the intensity-based refinement if there are no indications of substantial registration failures. On the other hand, our model must be flexible and consider the intensity-refined posterior probabilities if it is assumed that the multi-atlas propagation failed. This is often observed if the subject of interest shows severe brain abnormality due to disease related atrophy, traumatic deformation, or surgical resection of brain tissue (cf. Fig. 1).

Our basic assumption for an automatic choice of Γ is that the EM-refined segmentation Z should get a higher weight with increasing deviation from the segmentation obtained through label fusion Π.

Here, we assume that if a label has a similar intensity distribution before and after the intensity-based refinement, the result obtained through the label fusion Π is reliable for this label, and thus Γ_i should be close to 1. In contrast, if for example the hippocampal label in Π erroneously contains ventricular CSF, the intensity distribution has a rather large standard deviation, because the label contains two tissue types. However, after intensity-based refinement the intensity distribution of the label in Z is rather sharp, because the mislabelling of CSF is corrected due to the intensity-based refinement. In this case – two or more intensity distributions within a label based on Π and Z – we aim to set $\Gamma_i \ll 1$. This means that the more the EM-refined segmentation deviates from the prior the more it contributes to the final segmentation estimate.

To model this behaviour, we choose Γ_i dependent on the most likely labels assigned to a certain voxel by the label fusion, $k_{\max,i} = \arg\max \pi_{ik}$, and the EM-refinement, $z_{\max,i} = \arg\max z_{ik}$. Specifically, we use the overlap of the normal distributions estimated on label $z_{\max,i}$ in both Z and Π, and for $k_{\max,i}$ accordingly. We thus calculate Γ_i as:

$$\Gamma_i = \int_{-\infty}^{+\infty} min\left(\mathcal{N}^{\Pi}_{k_{\max,i}}(y), \mathcal{N}^{Z}_{k_{\max,i}}(y)\right) dy \times \int_{-\infty}^{+\infty} min\left(\mathcal{N}^{\Pi}_{z_{\max,i}}(y), \mathcal{N}^{Z}_{z_{\max,i}}(y)\right) dy \tag{6}$$

This weighting approach is exemplified in Fig. 3. In this example, the common scenario is shown, in which joint label fusion labels a voxel as hippocampus, and the intensity-based refinement approach labels the same voxel as CSF. The two images on the right in Fig. 4 illustrate exemplary weights $(1 - \Gamma_i)$ for normal and abnormal images.

3. Experiments and results

The goal of this work was to devise a robust segmentation framework that can be employed to segment brain MRI with potentially highly abnormal brain configuration. Specifically we aimed to segment a database of traumatic brain injury patients and to extract biomarkers that can be correlated with clinical variables.

However, before applying the proposed methodology to clinical data in Section 3.2 we conducted quantitative experiments investigating our method's performance and characteristics on a well-studied benchmark dataset. For this dataset reference labels, which were manually annotated by experts, are available. This allowed us to calculate label overlaps, to perform a test–retest analysis and to compare our method to other state-of-the-art approaches in Section 3.1.

3.1. Quantitative evaluation on a benchmark dataset using manual labels

For evaluating MALP-EM, we used the dataset provided in the course of the "MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling" (Landman and Warfield, 2012) (cf. Section 2.1.2). The dataset consists of 35 T_1-weighted MR images with corresponding labels created manually by experts. As in the Grand

Table 3
Overview over all compared methods and their respective building blocks.

Method	Registration	Label fusion	Additional processing
MAPER	MAPER (cf. Section 2.2.2) (Heckemann et al., 2010)	Majority voting	None
MALP-JF	MAPER	JF (Wang et al., 2013)	None
MALP-EM$_{\Gamma_i}$	MAPER	JF	Proposed (cf. Sections 2.2.3, 2.2.4, 2.2.5, 2.2.6)
MALP-EM$_{\Gamma_i}$-BC	MAPER	JF	Proposed + bias correction (Wang et al., 2011)
PICSL-BC	ANTs (SyN) (Avants et al., 2008)	JF	Bias correction

Challenge, we divided the cohort into a training set of 15 subjects (10 female, 5 male, age: 23 ± 4.3 (mean ± SD) years, minimum age 19, maximum age 34) and a test set of 15 subjects (10 female, 5 male, age 45.7 ± 24.4, minimum age 18, maximum age 90). Including the 5 repeat scans, the test set consists of 20 images.

3.1.1. Label overlaps

In total we compared five different approaches:

- MAPER (Multi-Atlas Propagation with Enhanced Registration): Standard MAPER as proposed by Heckemann et al. (2010).
- MALP-JF (Multi-Atlas Label Propagation with Joint label Fusion): Segmentations obtained through joint label fusion using the implementation of Wang et al. (2013) with standard parameters.[2]
- MALP-EM$_{\Gamma_i}$ (Multi-Atlas Label Propagation with Expectation–Maximisation based refinement): MALP-JF followed by the proposed prior relaxation (Section 2.2.3), the EM-refinement (Section 2.2.4) and the spatially varying merging strategy (Section 2.2.6).
- MALP-EM$_{\Gamma_i}$-BC: MALP-EM$_{\Gamma_i}$ with additional learning-based segmentation bias correction[2] as proposed by Wang et al. (2011). We consider this as the setup yielding the highest accuracy on this benchmark dataset.
- PICSL-BC (PICSL research group - Bias Correction): best performing method in the Grand Challenge (Landman and Warfield, 2012) using SyN registration from the Advanced Normalization Tools (ANTs) and employs joint label fusion (Wang et al., 2013) and bias correction (Wang et al., 2011).

A further overview over the compared methods is provided in Table 3.

We segmented each of the 20 test images into 134 regions and calculated Dice overlaps (similarity indices, SI, Dice (1945)) with the available manual segmentations. SI values for the different segmentation methods are shown in Table 4. Individual SI values of non-cortical structures are shown in Appendix E.

We also compared our results to the best performing method in the Grand Challenge (Landman and Warfield, 2012) called PICSL-BC. PICSL-BC employs the joint label fusion presented in Wang et al. (2013) and a learning-based wrapper method presented in Wang et al. (2011) where segmentation bias with respect to the gold-standard segmentations is learned. Moreover, PICSL-BC has been evaluated on the same images (including the same split into training and test images) using the same 134 regions. No significant differences ($p > 0.01$) could be found between our proposed flexible MALP-EM$_{\Gamma_i}$ and PICSL-BC for the averaged similarity indices over all regions. Applying additional segmentation bias correction (Wang et al., 2011) to MALP-EM$_{\Gamma_i}$ significantly (Student's two-sided paired t-test, $p < 10^{-4}$) improved segmentation results. Since normal distribution cannot be assumed for similarity indices, we repeated the hypothesis testing using the non-parametric Wilcoxon signed-rank test. All differences shown in Table 4 remained significant at least at $p < 10^{-2}$.

Table 4
Similarity indices (SI) [%] averaged (unweighted) over 20 subjects for all 36 non-cortical regions, all 98 cortical regions and all 134 regions. The methods that were compared are MAPER using majority vote, MALP-JF using joint label fusion, MALP-EM$_{\Gamma_i}$ with a local merging strategy and optional segmentation bias correction (BC), and PICSL-BC. *,** = significantly different to the method in the column to the left. bold = significantly best results.

	MAPER	MALP-JF	MALP-EM$_{\Gamma_i}$	MALP-EM$_{\Gamma_i}$-BC	PICSL-BC
SI 36 non-cortical	82.0	82.7**	82.9	**83.4***	**83.8**
SI 98 cortical	72.4	73.2**	73.8**	**74.9****	73.9*
SI 134 regions	74.9	75.8**	76.3**	**77.2****	76.5*

* $p < 10^{-2}$.
** $p < 10^{-4}$.

3.1.2. Evaluation of the influence of the weighting factor Γ

As illustrated in Fig. 5, a weighting factor of $\Gamma = 0.8$ yields the best segmentation results on the training data set. This result shows that joint fusion yields accurate labels for the majority of voxels. For voxels with a high uncertainty (more than one label has a high non-zero probability) the EM-refined result should be considered as additional weighting. Using the more flexible model with a spatially varying Γ_i, we observed comparable overlaps to $\Gamma = 0.8$. This is encouraging, because a global and fixed Γ leads to a stricter model that is assumed to perform well on healthy, normal data while only a data-driven choice of Γ_i allows the flexibility to cope with highly abnormal images of TBI subjects.

3.1.3. Test–retest reliability

To investigate the consistency of segmentations calculated with the proposed method, we evaluated its test–retest reliability. We quantified this characteristic using the 5 subjects in our set for whom repeat images are available. We used the intraclass correlation coefficient (ICC; two-way random single measures, absolute agreement) following Shrout and Fleiss (1979). Specifically, we calculated the reproducibility of label volumes calculated on images of the same subject at different time points (scan interval less than 90 days). The assumption is that brains of healthy subjects do not change substantially within short periods. ICC is widely used to quantify test–retest reliability (Kempton et al., 2011; Nugent et al., 2013).

On the manual segmentations we calculated an average ICC of 0.80 ± 0.28 for non-cortical and 0.78 ± 0.25 for cortical regions. Using the proposed method MALP-EM$_{\Gamma_i}$ we obtained an average ICC of 0.97 ± 0.04 for non-cortical and 0.94 ± 0.08 for cortical regions. These results are slightly better than the average ICC obtained using joint label fusion only (non-cortical: 0.96 ± 0.06, cortical: 0.94 ± 0.09).

We further assessed the relative volume difference (Δ_{vol}) between a structure's volume at two time points, V_{t_1} and V_{t_2}, which we define as:

[2] Implementation publicly available at https://www.nitrc.org/projects/picsl_malf/ (Version 1.2, last accessed: 8 December 2014).

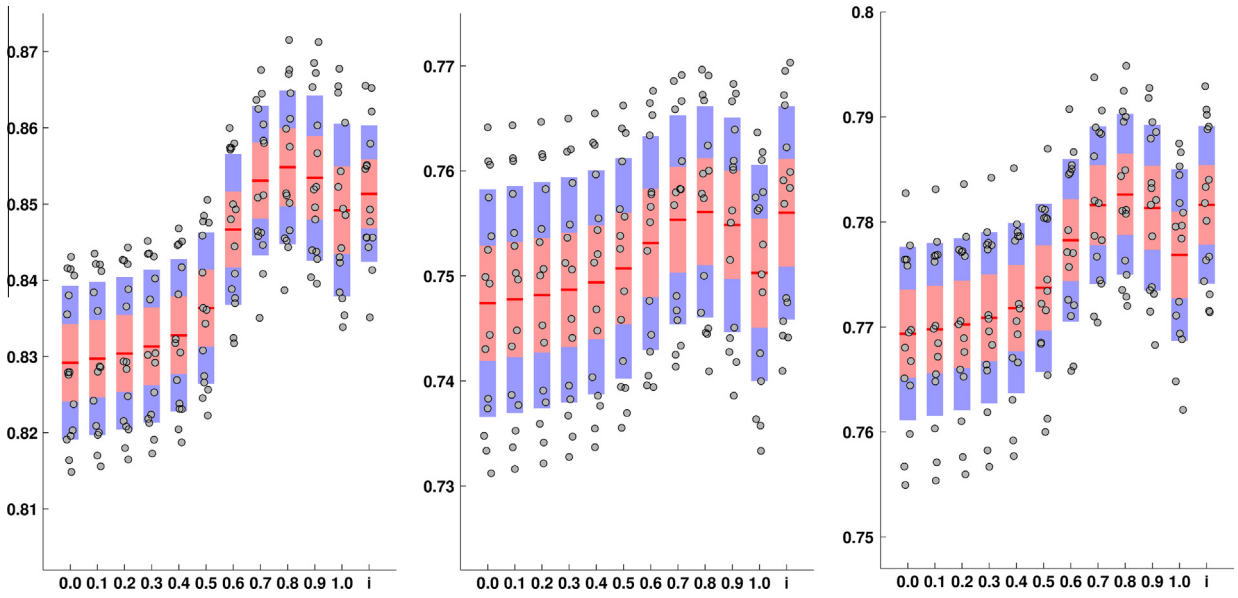

Fig. 5. MALP-EM$_\Gamma$ applied to the 15 training images with varying Γ. Here $\Gamma = 1.0$ is equivalent to exclusive joint label fusion and i denotes the spatially varying choice of Γ_i. Mean similarity indices (SI) with standard deviation and 95% confidence interval of 36 non-cortical structures (left), 98 cortical structures (middle) and all 134 regions (right).

$$\Delta_{\text{vol}} = 100\% \frac{|V_{t_1} - V_{t_2}|}{0.5(V_{t_1} + V_{t_2})} \tag{7}$$

On the manual segmentations we calculated an average Δ_{vol} of $8.3 \pm 7.5\%$ for non-cortical and $12.3 \pm 8.4\%$ for cortical regions. Using the proposed method MALP-EM$_{\Gamma_i}$ we obtained an average Δ_{vol} of $2.4 \pm 1.4\%$ for non-cortical and $4.1 \pm 2.3\%$ for cortical regions. These results are similar to the average Δ_{vol} obtained using joint label fusion only (non-cortical: $2.7 \pm 2.5\%$, cortical: $3.8 \pm 2.2\%$).

More extensive quantitative results can be found in Appendix E in Table E.9.

3.2. Segmentation and analysis of a traumatic brain injury (TBI) database

We used MALP-EM$_{\Gamma_i}$ (in the following referred to as MALP-EM) to automatically segment 125 MR brain scans of TBI subjects with potential pathology. This database is described in Section 2.1.1. For the segmentation we used all available atlas datasets, except the 5 repeat images, i.e., a total of 30.

All MR images were corrected for intensity inhomogeneities using the N4 algorithm (Tustison et al., 2010). The images were further brain extracted with a fully-automatic in-house brain extraction method called Extended Tissue Classification (ETC). In this method, first, an expectation–maximisation classifier based on Van Leemput et al. (1999) was applied for producing a coarse tissue segmentation, which is used as an initial brain mask. Thereafter, a deformable model-based approach combined with morphological operations was applied to tune the mask.

3.2.1. Quantitative evaluation on TBI datasets using expert validation scores

We devised a scoring protocol to semi-quantitatively assess the quality of automatically generated segmentations of TBI images. We selected four paired regions that frequently show morphological change in patients (hippocampus, thalamus, putamen, and occipital pole). We selected thalamus, putamen and occipital cortices because these structures are frequently implicated in TBI and its sequelae (Warner et al., 2010a; Strangman et al., 2010; Ramlackhansingh et al., 2011). We added the hippocampus because it is a challenging structure to segment (cf. Fig. 1). Consideration of the hippocampus is biologically justified, as it is typically involved in dementia, which in turn is a frequent long-term consequence of severe TBI. The protocol calls for the raters to assign a score on a six point scale (0, worst to 5, best). Three experienced raters (JCL, 1 year of clinical service; RAH, 12 years of clinical service; AH, 16 years of clinical service) developed the protocol by consensus, using 9 images with corresponding segmentations calculated with both MALP-EM and joint label fusion (MALP-JF). All raters had basic (JCL) or advanced (RAH, AH) training in neuroanatomy, neuropathology, radiology, and neuroimaging. The 9 images had been selected from the TBI database using an ad hoc approach that ensured that the sample was broadly representative (MC 2–6; 4 baseline and 5 follow up scans). The detailed protocol is provided as supplementary material to this manuscript.

Based on this protocol the three independent raters assessed 13 images using the tool `rview` from the Image Registration Toolkit (IRTK, https://github.com/BioMedIA/IRTK, last accessed: 8 December 2014). All raters were blind to the method (MALP-JF or MALP-EM). Results of both methods were presented in a balanced, randomised fashion. The raters viewed both methods' results of each subject back to back in order to break the tie if the two scores were equal. None were directly involved in the development of MALP-EM. The set of test scans did not intersect with the set of scans used for protocol development. Ten scans were randomly chosen with the constraint that five scans be of subjects with MC < 4 and five with MC $\geqslant 4$. In addition, a non-rater chose three further subjects to ensure scans with severe pathology were represented in the evaluation set.

The findings obtained through this expert validation confirm that MALP-EM is superior to joint label fusion in traumatic brain injury patients with severe pathology. The average expert scores are shown in Table 5. The distribution of the scores for the individual structures is shown in Fig. 6. Further, Fig. 7 shows the fraction of test images on which a method performs better than the other.

To further assess inter-rater variance we have calculated the intraclass correlation coefficient (ICC; two-way random single measures, absolute agreement, Shrout and Fleiss (1979)) between

Table 5
Mean (standard deviation) of the expert scores for the assessed segmentation quality of hippocampus, thalamus, putamen and occipital pole. Significant improvement is indicated. Intraclass correlation coefficients (ICC; two-way random single measures, absolute agreement) between all available raters.

	Hippocampus		Thalamus		Putamen		Occipital pole	
	MALP-JF	MALP-EM$_{\Gamma_i}$	MALP-JF	MALP-EM$_{\Gamma_i}$	MALP-JF	MALP-EM$_{\Gamma_i}$	MALP-JF	MALP-EM$_{\Gamma_i}$
Rater A	2.81(1.30)	3.23(1.48)	2.42(1.27)	3.12(1.63)*	2.81(2.25)	3.04(2.20)	3.42(1.39)	3.81(1.58)*
Rater B	3.00(1.13)	3.23(1.03)	2.88(1.07)	3.62(1.20)**	3.81(1.30)	3.96(1.37)	2.58(0.90)	3.08(0.93)*
Rater C	2.04(0.96)	2.65(0.89)*	2.00(0.69)	3.08(1.02)**	2.85(1.19)	3.19(1.27)*	–	–
All Raters	2.62(1.20)	3.04(1.18)**	2.44(1.09)	3.27(1.32)**	3.15(1.69)	3.40(1.69)*	3.00(1.24)	3.44(1.33)*
ICC	0.64		0.53		0.67		0.23	

* $p < 0.05$.
** $p < 10^{-4}$.

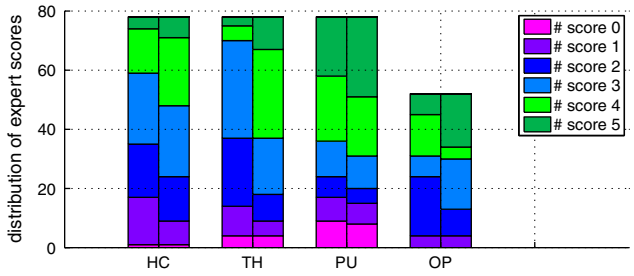

Fig. 6. Distribution of expert scores for investigated structures. Occipital pole was not rated by rater C. Comparison of MALP-JF (left bars) and MALP-EM (right bars).

the available raters. The calculated ICCs are summarised in Table 5. We observed moderate inter-rater agreement between all three raters for hippocampus, thalamus and putamen. The inter-rater agreement for the occipital pole, which was rated by two raters of rather different experience, is at a lower level. However, both raters agreed that MALP-EM yields significantly better results on this structure.

3.2.2. Qualitative confirmation of the robustness using MALP-EM on images with pathology

We considered segmentation quality sufficient if no major parts of the brain were missing due to a overly restrictive brain extraction, and the cortical grey matter/white matter and visually dominant non-cortical boundaries (e.g. ventricles/grey matter) were matched by label boundaries. Inclusion or exclusion of structures that are not present in the atlases (e.g. lesions or contusions) was not regarded as failure. In a small subset ($\approx 5-10\%$) of the processed images, we accepted local inaccuracies in the shape of brain extractions in the cortical region, e.g. Fig. 9, if most cortical and especially subcortical structures were segmented successfully.

After visual inspection we identified five segmentations of insufficient quality. One failure originated in misregistration due to significant intensity inhomogeneities that remained after the N4 bias correction. On this single subject we reapplied the bias correction using the generated brain mask to further reduce inhomogeneities. Registration and segmentation were subsequently successful; we therefore retained the image. For another three subjects the generated brain mask was of insufficient quality. One of the corresponding scans was acquired from a subject with a follow up image for which the brain extraction was fine. We thus used the brain mask of the follow up time point to extract the brain at the acute stage. The subsequent segmentation result was satisfactory. The remaining two subjects (cf. Fig. 8), which were highly abnormal and the scans had very poor quality, could not be processed. A single image was excluded due to consistency problems in the NIfTI format file after image conversion. We conclude that none of these failures were directly related to MALP-EM. Overall we were able to process 122 out of 125 available scans successfully.

On visual inspection, all segmentations of these 122 images were considered reasonable, allowing for pathology. Visual examples of calculated segmentation results are shown in Fig. 9. The image pair illustrates the advantages of MALP-EM over sole label fusion in images with substantial pathology. Fig. 9 also clearly reveals improved segmentation results obtained with MALP-EM at boundaries of anatomical regions with large intensity contrast. Improvements in both the hippocampal region and at the cortical grey matter/cerebrospinal fluid boundary are particularly striking.

Labels obtained with MALP-EM on images with little pathology (Fig. 10) are visually convincing for both non-cortical and cortical structures in most instances. When substantial pathology is present (Fig. 9), we observed some inaccuracies. A frequent problem is unlabelled cortical grey matter due to imperfections of the brain mask. Voxels excluded during brain extraction are not reconsidered during the segmentation process. A subject for which this problem is most striking is illustrated in Fig. 9. However, we still kept this subject for our analysis since even in subjects with significant pathology only a few cortical and no subcortical structures are affected by this problem. In regions showing severe deformations or atrophy, such as the hippocampal region in the subject shown in Fig. 9, the nonrigid atlas alignment may consistently fail. Due to the proposed prior relaxation step and spatially varying label combination we were able to relax this problem and improve

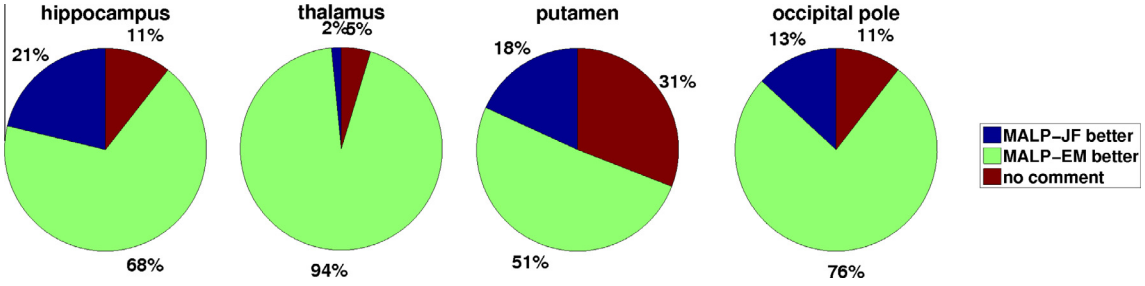

Fig. 7. Fraction of test images on which a method performs better than the other. Raters B and C were asked to rate which segmentation was better, even if scores were tied.

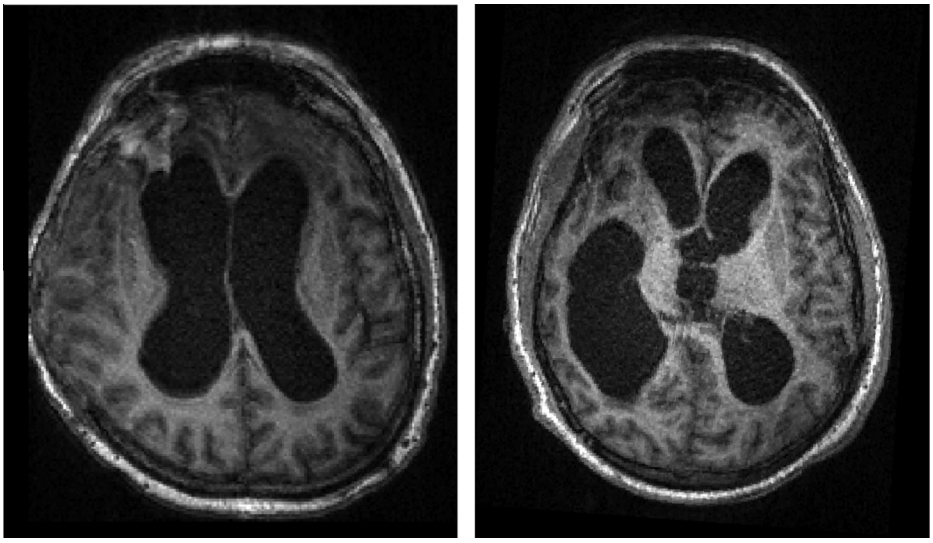

Fig. 8. Axial slices of the two subjects for which the brain extraction was not successful. left: male, age: 56, GOS = 3, MC = 5, follow up MRI taken two months after injury, right: male, age: 16, GOS = 3, MC = 5, follow up MRI taken seven months after injury.

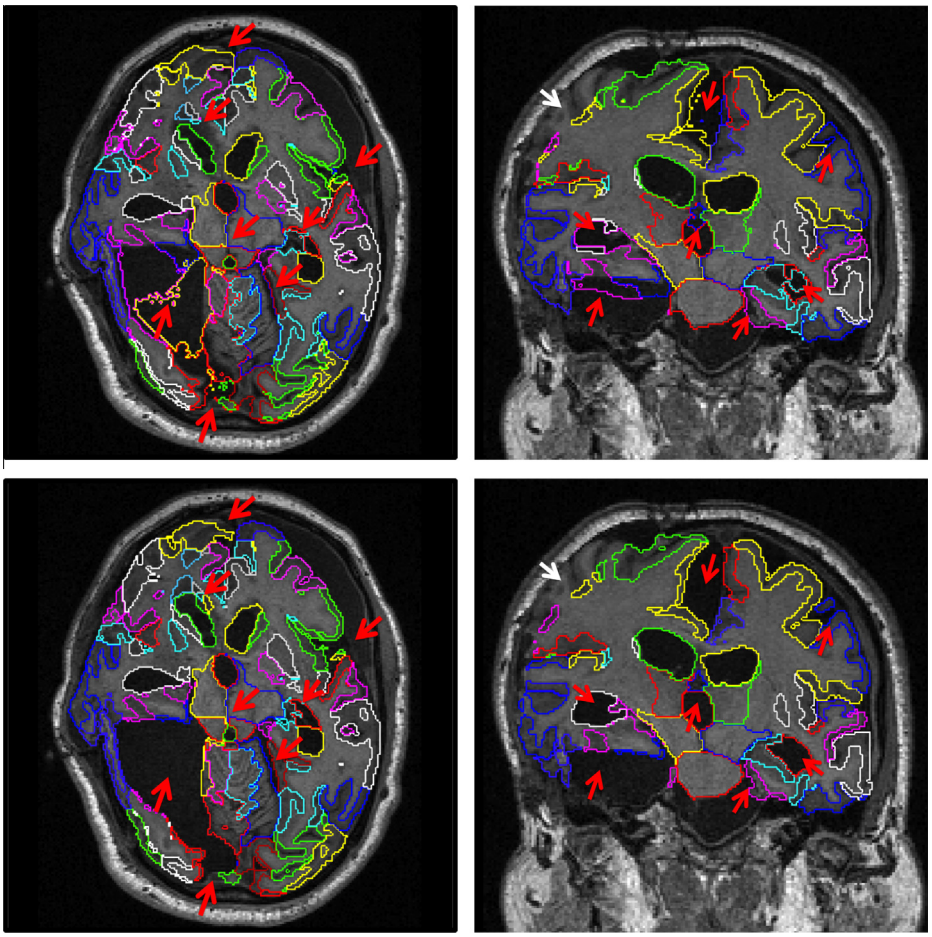

Fig. 9. Segmentations obtained with MALP-JF (top) and MALP-EM (bottom) of a subject with a markedly abnormal brain configuration on MRI (male, age: 45 years, GOS = 3, follow-up MRI, taken one month after injury, axial(left)/coronal(right) view). While the label fusion clearly fails to match various intensity boundaries, for example in the hippocampal region, MALP-EM is able to correct this problem to a large extent due to the strong intensity contrast CSF and grey matter structures. Red arrows highlight improvements obtained using MALP-EM over pure joint label fusion. White arrows highlight errors due to inconsistent brain extraction.

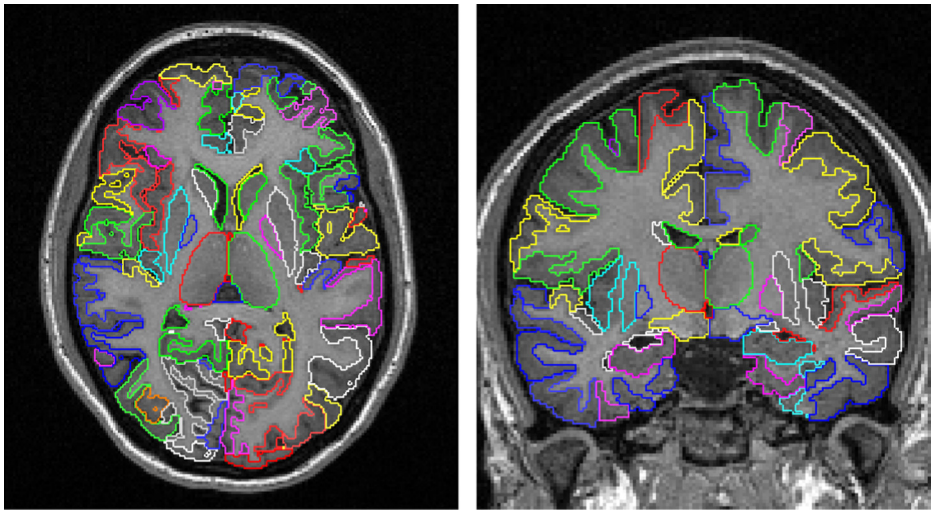

Fig. 10. Typical segmentation result obtained using MALP-EM$_{r_1}$ of a subject with a close-to-normal brain configuration (female, age: 17 years, GOS = 5, MC = 2, follow up MRI, taken nine months after injury, axial(left)/coronal(right) view).

upon non-intensity based approaches. We conclude from visual comparison that MALP-EM is superior to standard label fusion.

3.2.3. Separation of GOS and MC groups using absolute asymmetry indices

To assess the clinical usefulness of our method, we attempted a clinical classification of the available TBI subjects based on morphometric results. We used a classification of the brain images performed by an experienced clinician as a gold-standard reference. We assessed correlations with a widely used clinical scheme primarily devised for categorising admission X-ray CT images, applied to acute-phase MR images (MC) and one of the most common clinical outcome scores used in TBI (GOS). We focused on the particularly relevant differentiation between patients who would not be able to live an independent life (GOS < 4) and those with a more favourable outcome. The comparison with the GOS provided a means of estimating the prognostic value of automatically segmented acute-phase MR images. For the MC, we additionally dichotomised images into those without (MC $< 4 \triangleq$ DI I, DI II and DI III) or with (MC $\geqslant 4 \triangleq$ DI IV, EML, NEML) significant mass effect and midline shift. We employed MALP-EM for individual and independent cross-sectional experiments at the acute stage (60 subjects) and follow-up stage (62 subjects).

As classifier we used a linear discriminant analysis (LDA) implemented through the MATLAB function `classify`. For validation, we performed 1000 repetitions of a 10-fold cross validation. As feature we quantified structural asymmetry of the paired 63 structures (cf. Sections 2.1.2 and A). We employed an absolute asymmetry index (AAI) (Galaburda et al., 1987; Bonilha et al., 2014) based on a structure's volume (V) in the left and right hemisphere, defined as:

$$\mathrm{AAI} = 100\% \frac{|V_{\mathrm{left}} - V_{\mathrm{right}}|}{0.5(V_{\mathrm{left}} + V_{\mathrm{right}})} \tag{8}$$

Specifically, we used the sum of the absolute asymmetry indices of either all 14 non-cortical structures, all 49 cortical structures or all 63 structures.

The results for distinguishing two MC (MC < 4, MC $\geqslant 4$) and GOS (GOS < 4, GOS $\geqslant 4$) groups respectively, using either acute-phase or follow-up images, are summarised in Table 6. Since the clinical variables MC and GOS were missing for 3 (MC), respectively 5 (GOS) subjects, we reduced the number of subjects in each classification experiment accordingly. As groups were unbalanced we employed the balanced accuracy measure (Brodersen et al., 2010), the average of sensitivity and specificity, to report classification accuracy. Table 6 shows that our method yields 76.0% accuracy in distinguishing groups in the Marshall Classification system based on acute-phase images. The Marshall Classification system is not a linear scale as it takes both midline shift and the size of lesions into account (compare Appendix B). However, in the classification experiment we were able to discriminate between classes without (MC < 4) and with (MC $\geqslant 4$) significant midline shift or mass effect.

Furthermore, we were able to estimate from a single acute-phase MRI whether a TBI patient will be able to live an independent life (GOS $\geqslant 4$) or not with 64.7% accuracy. In comparison, when predicting outcome based on the MC score at baseline we calculated 59.3% accuracy. Based on the segmentations of non-cortical structures in the follow-up images, we achieved 66.8% accu-

Table 6
Classification results obtained separating MC < 4 vs. MC $\geqslant 4$ and GOS < 4 vs. GOS $\geqslant 4$ based on absolute asymmetry indices of either acute-phase (MC and GOS) or follow-up (GOS) MR images only. Results shown are averaged over 1000 cross validation runs. bold = best.

	Based on acute-phase MRIs						Based on follow-up MRIs		
	MC < 4 (Negatives) vs. MC $\geqslant 4$ (Positives)			GOS < 4 (P) vs. GOS $\geqslant 4$ (N)			GOS < 4 (P) vs. GOS $\geqslant 4$(N)		
	All non-cortical	All cortical	All	All non-cortical	All cortical	All	All non-cortical	All cortical	All
Balanced accuracy (%)	60.8	**76.0**	72.5	**64.7**	61.5	61.8	**66.8**	59.3	62.6
Specificity (%)	82.0	**91.2**	88.2	**84.0**	81.8	78.5	**86.3**	76.0	83.1
Sensitivity (%)	39.6	**60.9**	56.7	**45.4**	44.4	44.4	**47.4**	42.5	42.1
Subjects per group	34 vs. 23			27 vs. 28			19 vs. 43		

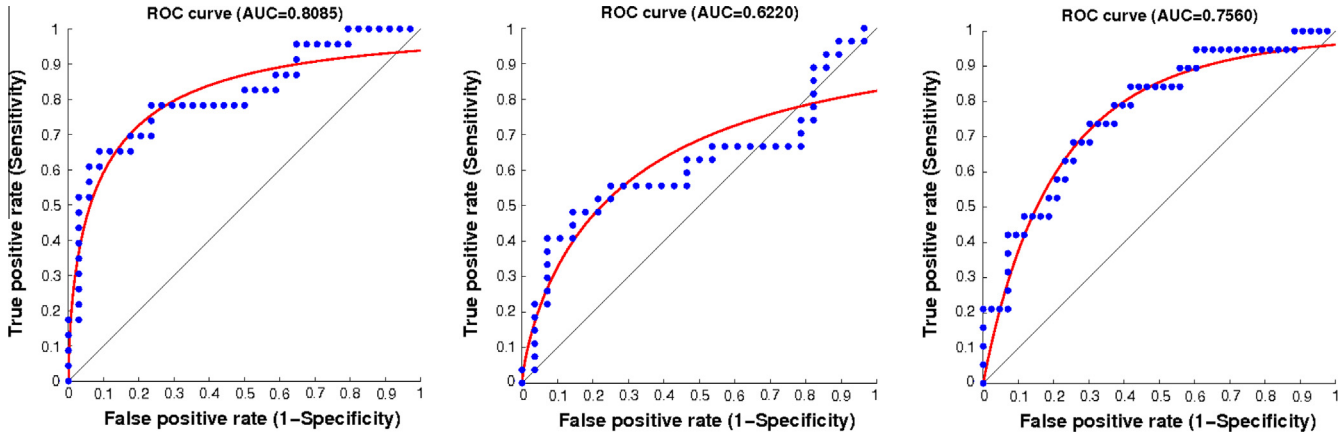

Fig. 11. Receiver operating characteristic curves for classifying subjects according to MC using the sum of cortical AAI (left), and according to GOS at baseline (middle) and follow up time point (right) using the accumulated non-cortical AAI.

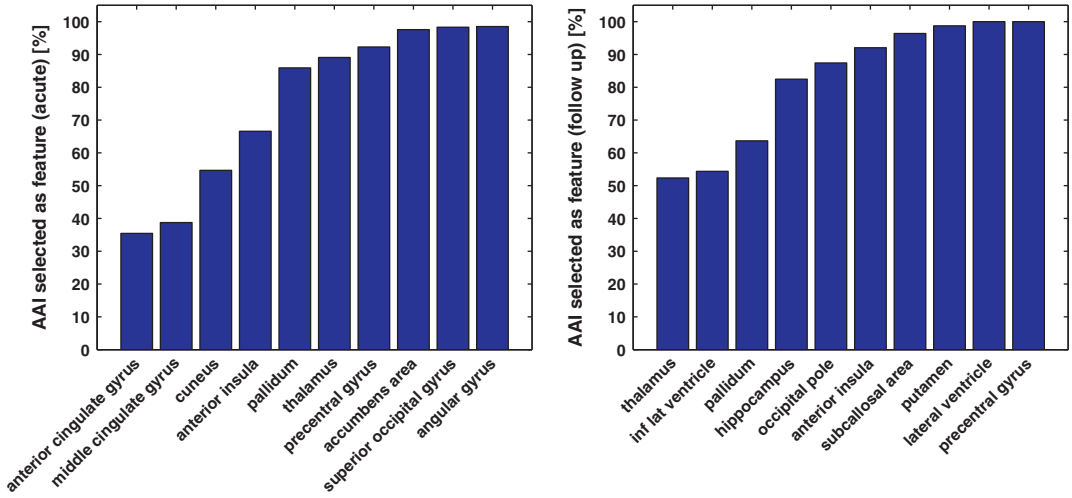

Fig. 12. Relevance of individual brain structures for GOS group separation. Sorted histogram of how often a structure's asymmetry index was one of the 10 most significant indices in the 10,000 (1000 rounds of 10-fold cross-validation) runs. The 10 structures that were picked most often based on acute-phase (left) or follow-up (right) MR images.

racy in GOS classification. The classification results are summarised in Table 6. The high specificity for MC classification shows that the presented method does very well in detecting normal appearing brains at the acute stage. The high specificity for GOS classification confirms that the presented approach is able to predict a favourable outcome of a TBI. These findings suggest that structural brain asymmetry could be a sufficient criterion to indicate an unfavourable disease outcome. On the other hand, symmetry seems to be a necessary criterion for favourable disease outcome. It is not, however, a sufficient criterion to rule out an unfavourable outcome. Receiver operating characteristic (ROC) curves for these classification experiments are shown in Fig. 11.

A detailed summary of results for individual non-cortical structures for both MC and GOS classification is provided in Appendix F, including *p*-values for group separation. These results suggest structural asymmetry of non-cortical brain structures does not correlate well with MC.

We calculated *p*-values for group separation using MALP-JF without the proposed processing. Unlike MALP-EM, this setup did not reveal any significant symmetry differences between GOS groups for the thalamus (at the acute time point) or for the caudate, hippocampus and inferior lateral ventricle (at the follow up time point). All the structures that show significant symmetry differences between groups of clinical variables in the MALP-JF setup are also found in the MALP-EM setup.

In an additional set of 1000 rounds of the 10-fold cross-validation, we determined in each run the *p*-value for the group separation on the training set using an unpaired two-sided Student's *t*-test for each of the 63 symmetry features (AAI). We calculated a histogram of the 10 most significant structures in each run. Fig. 12 shows the histogram for the GOS separation and thus the regions that are particularly correlated with the disease outcome. The plots show the 10 consistently most relevant structures for GOS group separation using acute-phase (left) or follow-up (right) MR images. This experiment reveals that asymmetry in subcortical structures is particularly correlated with poor patient recovery. Notably, asymmetry in the thalamus, pallidum, hippocampus, putamen and occipital pole was found to discriminate TBI patients with favourable from non-favourable outcome. Both thalamus and hippocampus are known to be involved in TBI disease progression (Bigler, 2001) and were found to have predictive value in previous studies based on MR imaging (Strangman et al., 2010; Warner

et al., 2010b; Warner et al., 2010a; Irimia et al., 2012). Our results also confirm the findings of Ramlackhansingh et al. (2011), where inflammation markers following a head trauma were significantly raised in the thalamus, putamen and occipital cortices.

4. Discussion and future work

In this work we introduced a framework called "Multi-Atlas Label Propagation with Expectation–Maximisation based refinement" (MALP-EM) for robust MR brain image segmentation. Building on state-of-the-art registration and label fusion techniques, we proposed to relax spatial priors obtained through multi-atlas label propagation and to combine segmentation results obtained with registration- and intensity-based approaches to exploit individual benefits. For prior relaxation we detect incorrect label priors at low intensity voxels or cisterns and redistribute corresponding probabilities to the most probable CSF-like structure. Here we assumed that low intensity voxels belong to cisterns that are filled with CSF. Potentially these low intensity voxels could also, especially in TBI patients, result from edema, hemorrhage, or direct injury. The employed atlas is built from healthy patients and our model does not allow for the detection of outliers or the classification of lesions, which is very challenging (Rao et al., 2014). Dependent on the disease, segmentation failures due to pathologies, such as contusions in TBI, could be addressed by an explicit lesion segmentation (Rao et al., 2014) or an outlier detection approach (Asman et al., 2013). This is, however, left for future work. In general, imaging features derived from automatic segmentations, such as structural volumes, need to be interpreted carefully, when pathologies are present.

We showed that MALP-EM significantly improves segmentation quality compared to non-intensity refined label fusion. Specifically, we observed significant improvements by combining results from joint label fusion and EM-refined fusion using a locally varying weighting factor Γ. This approach is similar to the weighting of different energy terms in the formulation presented by van der Lijn et al. (2008). Our formulation allows any combination of segmentation results calculated with independent models or unrelated labelling techniques. In the future it will be interesting to investigate how more sophisticated combination strategies and intensity models can further improve this approach.

Previously, the objective evaluation procedure of the "MICCAI Multi-Atlas-Segmentation Challenge 2012" (Landman and Warfield, 2012) has shown MALP-EM to be among the leading segmentation methods. While the implementation used in the challenge was preliminary and highly tuned, the implementation used for the present work is more generic and less dependent on parameter settings. Thanks to an improved registration and more sophisticated and general fusion strategy, we achieved significantly higher overlaps for both MAPER (overall SI: 74.9% vs. 74.1%) and MALP-EM (overall SI: 76.4% vs. 75.8%) than in the Grand Challenge. Additional application of a learning-based segmentation bias correction method (Wang et al., 2011) further improves our segmentation results (overall SI: 77.2%), yielding small but significant improvements over the best method (PICSL-BC; overall SI: 76.5%) in the Grand Challenge. We acknowledge that in the development of MALP-EM we benefitted from the experience of participating in the Grand Challenge, where the timeframe for algorithm development and tuning was tight. However, we did not use the testing set from the Grand Challenge to tune MALP-EM.

While the segmentation bias correction significantly improves label overlaps on the MICCAI Segmentation Challenge dataset, we did not employ this post processing technique to segment the TBI subjects. We reason that it is difficult to justify the application of a correction classifier that was trained exclusively on a homogeneous cohort of healthy subjects to a heterogeneous cohort of brain scans with severe pathology.

To evaluate MALP-EM on a TBI database, we specifically developed a protocol for the expert assessment of segmentation quality of the hippocampus, thalamus, putamen and occipital pole. The protocol is publicly available as supplementary material to this work. Using the protocol-based ratings of three independent experts, we showed on 13 subjects of the TBI cohort that the proposed modifications based on image intensities improve on pure label fusion.

The proposed method is shown to be robust: 120 out of 125 TBI images were segmented successfully into 134 regions. After manual intervention on the preprocessing step, the successful record increased to 122/125. This is a solid basis for future research into image-based quantification of brain abnormality. Derived morphometric biomarkers, such as a structural asymmetry index, can serve as features for automatic classifiers, predicting how the image will be rated by an expert (MC) and prognosticating clinical outcome (GOS). Given that MC is assessed at the acute stage quantifying brain pathology (cf. Appendix B) it seems reasonable that structural asymmetry in acute MRIs correlates well with this score. In contrast to this, GOS is an outcome score assessed several months after the injury. It was expected that MRI features derived from follow up scans are more consistent with the outcome measure than features available at the acute stage. The segmentation setup for the TBI subjects inherently differs from the intra-atlas experiments (15 training, 20 test images) in that we used more atlases for the segmentation of the TBI subjects, the image source (scanner) differs from the atlas database, and, most importantly, we are segmenting subjects with potentially substantial pathology. Visual inspection confirmed the robustness and advantages of the proposed method under these new challenges (cf. Fig. 9). The generated segmentations of the TBI data set are a valuable resource for investigating further potential biomarkers for TBI disease progress.

This work also motivates further research and discussion about how meaningful or generalisable high Dice overlaps on a homogeneous cohort of healthy patients are. More informative similarity measures are desirable (Ledig et al., 2014). In many studies, e.g. Alzheimer's disease or TBI, the subjects of interest show high variability in both brain appearance and disease burden. More restrictive models might lead to a high labelling accuracy on subjects that are very similar to the atlas cohort. However, they are potentially too rigid to cope with images of subjects with significant pathology. Here more flexible formulations might be desirable, even if they are slightly less performant in intra-atlas cohort validations.

We conclude from our experiments that MALP-EM$_{\Gamma_i}$ yields a segmentation accuracy that is on healthy subjects comparable to other state-of-the-art methods while offering sufficient flexibility to cope with gross pathology. Most inaccuracies, as visually confirmed in the segmented TBI datasets, were due to minor problems in the brain extraction. This highlights the necessity of further improvement of fully automatic brain extraction tools, which is a very challenging task for brains in the presence of pathology.

5. Conclusions

We presented a fully automatic, highly robust and accurate segmentation framework called MALP-EM. This includes a new paradigm: We suggest the spatially weighted combination of probabilistic segmentation results obtained through different techniques into a common segmentation exploiting individual benefits. Extensive quantitative evaluation on a manually annotated atlas cohort of healthy subjects confirmed that MALP-EM significantly improves on existing label fusion techniques. Based on the ratings

Table B.7
Marshall Classification system based and modified from Marshall et al. (1991).

Marshall class	Description
1 Diffuse injury (DI) I	No visible intracranial pathological changes seen on CT
2 Diffuse injury II	Cisterns are present with midline shift of 0–5 mm and/or: Lesions densities present; No high or mixed density lesion >25 cm^3 may include bone fragments and foreign bodies[a]
3 Diffuse injury III (swelling)	cisterns compressed or absent with midline shift of 0–5 mm; no high Or mixed density lesion >25 cm^3
4 Diffuse injury IV (shift)	midline shift >5 mm; no high or mixed density lesion > 25 cm^3
5 Evacuated mass lesion (EML)	Any lesion surgically evacuated
6 Non-evacuated mass lesion (NEML)	High or mixed density lesion >25 cm^3; not surgically evacuated

[a] As may be the case in depressed skull fractures.

of three independent experts, MALP-EM is superior to joint label fusion for hippocampus, thalamus, putamen and occipital pole segmentation on TBI brain scans. We have demonstrated the benefits regarding robustness through intensity based refinement on 125 MR brain images of TBI subjects. Using MALP-EM we were able to segment 122 out of 125 available TBI brain images into 134 different anatomical regions. We observed correlations between asymmetry indices of paired structures and clinical variables, using acute-phase or follow-up MR images. We also observed and confirmed evidence that subcortical brain structures such as the thalamus, putamen and hippocampus have strong potential to predict the clinical outcome of individual TBI patients.

Acknowledgements

This work was partially funded under the 7th Framework Programme by the European Commission (http://cordis.europa.eu/ist/, TBIcare: http://www.tbicare.eu/, last accessed: 8 December 2014). The research was further supported by the National Institute for Health Research (NIHR) Biomedical Research Centre (BRC) based at Imperial College Healthcare NHS Trust and Imperial College London. AH is supported by the Department of Health via the NIHR comprehensive BRC award to Guy's & St Thomas' NHS Foundation Trust in partnership with King's College London and Kings College Hospital NHS Foundation Trust. This work was further supported by a Medical Research Council (UK) Program Grant (Acute brain injury: heterogeneity of mechanisms, therapeutic targets and outcome effects [G9439390 ID 65883]), the UK National Institute of Health Research Biomedical Research Centre at Cambridge, the Technology Platform funding provided by the UK Department of Health and an EPSRC Pathways to Impact award. VFJN is supported by a Health Foundation/Academy of Medical Sciences Clinician Scientist Fellowship. DKM is supported by an NIHR Senior Investigator Award. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. The funders had no role in study design, data collection and analyses, decision to publish, or preparation of the manuscript.

Appendix A. Structure names for the 63 calculated asymmetry indices

1: accumbens area, 2: amygdala, 3: caudate, 4: cerebellum exterior, 5: cerebellum white matter, 6: cerebral white matter, 7: hippocampus, 8: inf lat ventricle, 9: lateral ventricle, 10: pallidum, 11: putamen, 12: thalamus, 13: ventral DC, 14: forebrain, 15: ACgG anterior cingulate gyrus, 16: AIns anterior insula, 17: AOrG anterior orbital gyrus, 18: AnG angular gyrus, 19: Calc calcarine cortex, 20: CO central operculum, 21: Cun cuneus, 22: Ent entorhinal area, 23: FO frontal operculum, 24: FRP frontal pole, 25: FuG fusiform gyrus, 26: GRe gyrus rectus, 27: IOG inferior occipital gyrus, 28: ITG inferior temporal gyrus, 29: LiG lingual gyrus, 30: LOrG lateral orbital gyrus, 31: MCgG middle cingulate gyrus, 32: MFC medial frontal cortex, 33: MFG middle frontal gyrus, 34: MOG middle occipital gyrus, 35: MOrG medial orbital gyrus, 36: MPoG postcentral gyrus medial segment, 37: MPrG precentral gyrus medial segment, 38: MSFG superior frontal gyrus medial segment, 39: MTG middle temporal gyrus, 40: OCP occipital pole, 41: OFuG occipital fusiform gyrus, 42: OpIFG opercular part of the inferior frontal gyrus, 43: OrIFG orbital part of the inferior frontal gyrus, 44: PCgG posterior cingulate gyrus, 45: PCu precuneus, 46: PHG parahippocampal gyrus, 47: PIns posterior insula, 48: PO parietal operculum, 49: PoG postcentral gyrus, 50: POrG posterior orbital gyrus, 51: PP planum polare, 52: PrG precentral gyrus, 53: PT planum temporale, 54: SCA subcallosal area, 55: SFG superior frontal gyrus, 56: SMC supplementary motor cortex, 57: SMG supramarginal gyrus, 58: SOG superior occipital gyrus, 59: SPL superior parietal lobule, 60: STG superior temporal gyrus, 61: TMP temporal pole, 62: TrIFG triangular part of the inferior frontal gyrus, 63: TTG transverse temporal gyrus.

Appendix B. The Marshall Classification system

Table B.7.

Appendix C. The Glasgow Outcome Scale

Table C.8.

Appendix D. Expectation–maximisation optimisation

For the sake of readability and consistency with existing literature we have followed the notation used in Van Leemput et al. (1999), Cardoso et al. (2011), Ledig et al. (2012). Our implementation builds on the framework described in Ledig et al. (2012).

Table C.8
Glasgow Outcome Scale (GOS) based and modified from Jennett and Bond (1975). For dichotomised assessment, 1, 2 and 3 are often combined as "Unfavourable Outcomes", while 4 and 5 are combined as "Favourable Outcomes".

GOS	Description
1 (D)	*Dead*
2 (VS)	*Vegetative state*: no evidence of meaningful responsiveness
3 (SD)	*Severe disability*: conscious, but unable to live independently due to mental or physical disability
4 (MD)	*Moderate disability*: able to live independently, limited ability to return to work or school
5 (GR)	*Good recovery*: capacity to resume normal occupational and social activities, minor deficits possible

Given the parameters $\Phi = \{(\mu_1, \sigma_1), (\mu_2, \sigma_2), \ldots, (\mu_K, \sigma_K)\}$ of K structural classes the likelihood of observing, the log-transformed, intensity y_i at voxel i is given as:

$$f(y_i|\Phi) = \sum_k f(y_i|\mathbf{z_i} = \mathbf{e_k}, \Phi) f(\mathbf{z_i} = \mathbf{e_k}) \tag{D.1}$$

It is commonly assumed that the probability, $f(y_i|\mathbf{z_i} = \mathbf{e_k}, \Phi)$, of a voxel i to have intensity y_i, given that it belongs to class $k, (\mathbf{z_i} = \mathbf{e_k})$, is described by a normal distribution (Wells et al., 1996; Van Leemput et al., 1999; Zhang et al., 2001; Cardoso et al., 2011). We thus model $f(y_i|\mathbf{z_i} = \mathbf{e_k}, \Phi) = \mathcal{N}_k(y_i)$ where $\mathcal{N}_k$ denotes the Gaussian distribution with corresponding parameters (μ_k, σ_k). The prior probability $f(\mathbf{z_i} = \mathbf{e_k})$ that a voxel i belongs to structure k is given by the relaxed version, Π^R (cf. Section 2.2.3), of the probabilistic label estimates after multi-atlas label propagation.

Next to the spatial information provided by the prior estimates, Π^R, we further account for topological knowledge by incorporating a Markov Random Field (MRF). We thus expand $f(\mathbf{z_i} = \mathbf{e_k}) = \Pi^R_{ik}$ to

$$f(\mathbf{z_i} = \mathbf{e_k}|p^{(m)}_{\mathscr{N}_i}, G) = \frac{\pi^R_{ik} e^{-U_{MRF}(\mathbf{e_k}|p^{(m)}_{\mathscr{N}_i}, G)}}{\sum_{j=1}^K \pi^R_{ij} e^{-U_{MRF}(\mathbf{e_j}|p^{(m)}_{\mathscr{N}_i}, G)}} \tag{D.2}$$

We calculate the MRF energy function U_{MRF} based on the probabilistic label estimates in iteration m, p^m_{ik}, in the first-order neighbourhood of voxel $i, \mathscr{N}_i$, as:

$$U_{MRF}(\mathbf{e_k}|p^{(m)}_{\mathscr{N}_i}, G) = \tag{D.3}$$

$$\sum_{j=1}^K G_{kj} \left(\sum_{l \in \mathscr{N}^x_i} s_x p^{(m)}_{lj} + \sum_{l \in \mathscr{N}^y_i} s_y p^{(m)}_{lj} + \sum_{l \in \mathscr{N}^z_i} s_z p^{(m)}_{lj} \right) \tag{D.4}$$

Here, G denotes a $K \times K$ matrix defining the connectivity between class k and j and $s = \{\frac{1}{d_x}, \frac{1}{d_y}, \frac{1}{d_z}\}$ accounts for the anisotropic voxel spacing in world coordinates. We have defined G as:

$$G(k,j) = \begin{cases} 0, & \text{if } k = j \\ \beta, & \text{if structures } k \text{ and } j \text{ share a boundary} \\ \gamma, & \text{if structures } k \text{ and } j \text{ are distant} \end{cases} \tag{D.5}$$

Here β and γ, with $0 \leqslant \beta \leqslant \gamma$, are parameters describing the penalty for certain neighbourhood configurations. By assuming that voxels are statistically independent, the probability of observing an image $\mathbf{I_u}$, given that the parameters Φ are known, is given by $f(\mathbf{I_u}|\Phi) = \prod_i f(y_i|\Phi)$. We can now solve this model by interleaving the expectation of the class probabilities $p^{(m)}_{ik}$ and the maximisation of the model by updating the model parameters $\Phi^{(m)}$. We then assume that the label probabilities, $p^{(m+1)}_{ik}$, are known in iteration $(m+1)$ and update the model parameters as:

$$\mu_k^{(m+1)} = \frac{\sum_i p^{(m+1)}_{ik} y_i}{\sum_i p^{(m+1)}_{ik}}, \qquad \sigma_k^{(m+1)} = \sqrt{\frac{\sum_i p^{(m+1)}_{ik} (y_i - \mu_k^{(m+1)})^2}{\sum_i p^{(m+1)}_{ik}}} \tag{D.6}$$

Table E.9
Similarity indices [%] averaged over 20 test images. Intraclass correlation coefficients (ICC; two-way random single measures, absolute agreement, Shrout and Fleiss (1979)) and relative volume differences (Δ_{vol}) based on 5 subjects with available repeat scans for all 36 considered non-cortical structures.

	SI				ICC/ Δ_{vol}		
	MAPER	MALP-JF	MALP-EM$_{\Gamma_i}$	MALP-EM$_{\Gamma_i}$ – BC	Manual (%)	MALP-JF (%)	MALP-EM$_{\Gamma_i}$ (%)
3rd ventricle	85.2	85.6	79.8**	85.8**	0.855/17.8 ± 11.4	0.994/3.6 ± 2.4	0.978/5.9 ± 3.3
4th ventricle	86.7	87.3*	87.1	87.1	0.991/3.0 ± 2.3	0.999/1.5 ± 0.7	0.998/1.6 ± 1.6
Accumbens area R	77.9	78.4	76.7	78.0	0.673/13.8 ± 10.5	0.961/3.6 ± 2.6	0.984/2.0 ± 2.6
Accumbens area L	77.3	77.3	77.6	78.7	0.755/11.2 ± 10.3	0.728/7.1 ± 6.9	0.867/5.9 ± 2.4
Amygdala R	80.2	80.6	79.4	81.4*	0.605/10.9 ± 14.0	0.893/3.6 ± 2.6	0.929/3.7 ± 1.7
Amygdala L	81.3	81.9	81.7	83.1*	0.897/6.0 ± 4.1	0.925/4.4 ± 3.2	0.982/2.5 ± 2.1
Brain stem	93.7	93.8*	93.7	94.0**	0.957/3.5 ± 1.8	0.993/1.1 ± 0.7	0.993/1.2 ± 0.7
Caudate R	87.0	87.5	86.8	87.9*	0.955/3.1 ± 1.8	0.992/1.2 ± 1.0	0.991/1.4 ± 0.9
Caudate L	87.0	87.8*	87.4	88.0	0.956/3.4 ± 2.4	0.998/0.6 ± 0.3	0.996/0.8 ± 0.5
Cerebellum exterior R	92.5	92.9**	93.4**	93.5	0.985/2.1 ± 1.4	0.993/1.2 ± 1.7	0.995/1.2 ± 1.4
Cerebellum exterior L	92.1	92.6**	93.1*	93.2	0.978/2.2 ± 2.2	0.989/1.6 ± 1.9	0.988/2.0 ± 1.5
Cerebellum white matter R	88.9	89.1	90.5**	89.9	0.978/3.8 ± 2.6	0.992/1.4 ± 1.7	0.995/1.3 ± 1.0
Cerebellum white matter L	89.0	89.2*	90.6**	90.2	0.884/5.3 ± 5.3	0.972/2.1 ± 2.1	0.975/2.2 ± 1.7
Cerebral white matter R	93.3	93.3	93.7*	94.0	0.952/3.1 ± 1.8	0.992/1.1 ± 1.2	0.993/1.0 ± 1.0
Cerebral white matter L	93.2	93.2	93.6*	93.9	0.969/2.2 ± 1.9	0.995/0.9 ± 0.9	0.996/0.8 ± 0.8
Cerebrospinal fluid	77.7	79.8	77.3*	81.0**	0.794/11.6 ± 9.9	0.859/4.7 ± 3.9	0.824/4.0 ± 6.1
Hippocampus R	85.1	86.3*	86.4	86.8	0.815/8.4 ± 7.3	0.997/0.9 ± 0.5	0.991/1.5 ± 1.0
Hippocampus L	85.2	86.5**	86.3	86.9*	0.870/7.6 ± 7.6	0.996/1.0 ± 0.5	0.996/0.9 ± 1.1
Inf lat ventricle R	55.6	63.2**	68.8**	70.9	0.794/20.4 ± 8.6	0.993/4.4 ± 3.3	0.992/3.8 ± 2.8
Inf lat ventricle L	55.5	62.2**	67.1**	67.2	0.983/12.9 ± 7.6	0.991/5.9 ± 4.7	0.996/4.1 ± 3.8
Lateral ventricle R	91.9	92.5*	93.0	93.0	0.999/9.4 ± 8.3	1.000/2.6 ± 1.6	1.000/3.1 ± 2.1
Lateral ventricle L	92.3	93.0*	93.5	93.3	0.999/9.1 ± 7.0	1.000/2.8 ± 1.6	1.000/2.4 ± 1.7
Pallidum R	86.6	87.5**	87.6	87.6	0.468/8.8 ± 1.4	0.948/2.4 ± 1.4	0.937/2.8 ± 1.7
Pallidum L	85.1	86.7**	86.6	86.7	0.640/4.7 ± 3.3	0.917/2.2 ± 1.7	0.914/2.1 ± 2.4
Putamen R	91.0	91.3	91.1	90.8	0.961/2.9 ± 1.5	0.995/1.0 ± 0.6	0.990/1.4 ± 1.0
Putamen L	90.8	91.1	91.1	91.0	0.978/2.0 ± 1.8	0.980/1.4 ± 1.8	0.972/1.6 ± 2.2
Thalamus proper R	91.7	92.1*	91.4*	92.0**	0.950/2.4 ± 1.6	0.990/1.0 ± 0.9	0.985/1.4 ± 0.9
Thalamus proper L	91.9	92.1	91.5*	91.9*	0.876/3.5 ± 3.0	0.988/0.9 ± 0.9	0.978/1.4 ± 1.0
Ventral DC R	88.6	88.8*	88.1*	88.9*	0.835/4.6 ± 2.9	0.996/1.0 ± 0.3	0.988/1.8 ± 0.8
Ventral DC L	88.7	88.7	88.1**	88.9**	0.862/4.8 ± 2.8	0.994/1.0 ± 0.7	0.985/1.6 ± 1.4
Optic chiasm	52.0	49.1	53.9	43.5	−0.355/40.5 ± 38.4	0.801/14.4 ± 10.3	0.947/5.9 ± 3.5
Cerebellar vermal lobules I–V	81.9	82.3	82.7	83.2	0.470/11.5 ± 11.1	0.993/2.3 ± 1.3	0.990/2.6 ± 1.8
Cerebellar vermal lobules VI–VII	77.0	77.8	78.1	79.7**	0.867/4.9 ± 2.2	0.902/3.1 ± 1.9	0.859/3.5 ± 2.8
Cerebellar vermal lobules VIII–X	87.0	87.5	87.6	87.7	0.962/5.0 ± 3.8	0.995/1.6 ± 1.1	0.998/1.2 ± 0.8
Basal forebrain R	43.9	44.5	44.6	47.2	0.722/12.2 ± 6.9	0.880/3.7 ± 2.8	0.945/2.1 ± 2.0
Basal forebrain L	45.4	45.0	45.0	46.8	0.040/18.8 ± 7.5	0.926/3.8 ± 3.4	0.887/3.5 ± 1.8

* Significantly different SI compared to column to the left indicated at $p < 10^{-2}$.
** Significantly different SI compared to column to the left indicated at $p < 10^{-4}$.

Table F.10
Classification results (10-fold cross-validation, 1000 runs) obtained separating Marshall Classification (MC) < 4 vs. MC ⩾ 4 based on absolute asymmetry indices (AAI) of acute-phase MR images. Significant group differences indicated by + ($p < 0.05$) and ++ ($p < 0.01$).

Structure	Classification of MC < 4 (Negatives) vs. MC ⩾ 4 (Positives) (acute scans)					
	Balanced ACC	SPEC	SENS	mean AAI (SD) for MC < 4 and MC ⩾ 4	*p*-value	Significance
All non-cortical	0.608	0.820	0.396	175.9 (88.4)	308.9 (231.1)	3.50e−03	++
All cortical	0.760	0.912	0.609	759.3 (151.6)	1072.2 (354.8)	2.74e−05	++
All	0.725	0.882	0.567	935.3 (200.4)	1381.1 (556.7)	7.25e−05	++
Accumbens area	0.628	0.799	0.458	16.4 (13.6)	32.7 (35.8)	1.90e−02	+
Amygdala	0.506	0.775	0.237	15.2 (16.0)	24.7 (44.2)	2.55e−01	o
Caudate	0.506	0.699	0.312	6.9 (7.0)	12.2 (17.8)	1.18e−01	o
Cerebellum exterior	0.660	0.793	0.526	4.2 (4.2)	8.7 (7.1)	3.67e−03	++
Cerebellum white matter	0.499	0.604	0.393	10.9 (8.4)	15.3 (13.5)	1.37e−01	o
Cerebral white matter	0.484	0.618	0.351	2.7 (1.7)	4.6 (7.0)	1.22e−01	o
Hippocampus	0.370	0.656	0.085	8.5 (7.0)	12.3 (26.5)	4.21e−01	o
Inf lat ventricle	0.685	0.814	0.555	24.0 (19.1)	51.9 (43.7)	1.74e−03	++
Lateral ventricle	0.555	0.746	0.364	24.9 (18.7)	47.4 (48.7)	1.78e−02	+
Pallidum	0.596	0.911	0.281	9.5 (15.3)	22.0 (31.7)	5.26e−02	o
Putamen	0.572	0.782	0.362	11.9 (22.0)	16.3 (23.1)	4.63e−01	o
Thalamus	0.550	0.794	0.307	5.6 (8.5)	10.3 (14.5)	1.22e−01	o
Ventral DC	0.386	0.391	0.380	8.3 (7.8)	8.2 (5.5)	9.88e−01	o
Forebrain	0.549	0.751	0.348	27.1 (24.6)	42.2 (50.3)	1.37e−01	o

Table F.11
Classification results (10-fold cross-validation, 1000 runs) obtained separating Glasgow Outcome Score (GOS) < 4 vs. GOS ⩾ 4 based on absolute asymmetry indices (AAI) of acute-phase MR images. Significant group differences indicated by + ($p < 0.05$) and ++ ($p < 0.01$).

Structure	Classification of GOS < 4 (Positives) vs. GOS ⩾ 4 (Negatives) (acute scans)					
	Balanced ACC	SPEC	SENS	Mean AAI (SD) for GOS < 4 and GOS ⩾ 4	*p*-value	Significance
All non-cortical	0.647	0.840	0.454	272.8 (213.8)	183.2 (121.4)	6.02e−02	o
All cortical	0.631	0.818	0.444	954.2 (372.6)	821.7 (198.6)	1.04e−01	o
All	0.615	0.785	0.444	1227.0 (556.6)	1004.8 (290.4)	6.77e−02	o
Accumbens area	0.597	0.754	0.440	29.7 (33.9)	15.0 (11.0)	3.36e−02	+
Amygdala	0.589	0.843	0.335	25.4 (41.4)	13.4 (14.8)	1.56e−01	o
Caudate	0.547	0.719	0.376	9.9 (13.3)	7.9 (12.4)	5.64e−01	o
Cerebellum exterior	0.600	0.465	0.734	5.2 (5.1)	7.7 (6.6)	1.27e−01	o
Cerebellum white matter	0.439	0.519	0.360	13.6 (12.6)	12.8 (9.6)	7.78e−01	o
Cerebral white matter	0.472	0.593	0.351	3.8 (6.3)	3.1 (2.5)	5.51e−01	o
Hippocampus	0.510	0.757	0.263	12.6 (23.6)	7.8 (9.4)	3.20e−01	o
Inf lat ventricle	0.426	0.508	0.343	35.0 (34.8)	33.1 (33.9)	8.34e−01	o
Lateral ventricle	0.564	0.729	0.398	39.1 (40.6)	24.3 (30.1)	1.28e−01	o
Pallidum	0.578	0.895	0.261	21.4 (29.8)	8.7 (15.5)	5.21e−02	o
Putamen	0.605	0.820	0.391	18.6 (28.3)	9.8 (14.6)	1.48e−01	o
Thalamus	0.582	0.821	0.342	10.8 (15.1)	4.6 (5.5)	4.61e−02	+
Ventral DC	0.374	0.352	0.395	8.4 (5.3)	8.4 (8.4)	9.99e−01	o
Forebrain	0.521	0.709	0.333	39.0 (48.6)	26.6 (24.5)	2.36e−01	o

Table F.12
Classification results (10-fold cross-validation, 1000 runs) obtained separating Glasgow Outcome Score (GOS) < 4 vs. GOS ⩾ 4 based on absolute asymmetry indices (AAI) of follow-up MR images. Significant group differences indicated by + ($p < 0.05$) and ++ ($p < 0.01$).

Structure	Classification of GOS < 4 (Positives) vs. GOS ⩾ 4 (Negatives) (follow-up scans)					
	Balanced ACC	SPEC	SENS	mean AAI (SD) for GOS < 4 and GOS ⩾ 4	*p*-value	Significance
All non-cortical	0.668	0.863	0.474	340.0 (219.8)	187.5 (101.1)	3.84e−04	++
All cortical	0.593	0.760	0.425	950.9 (357.7)	792.2 (166.8)	1.97e−02	+
All	0.626	0.831	0.421	1290.9 (522.9)	979.6 (227.2)	1.70e−03	++
Accumbens area	0.648	0.928	0.369	54.4 (58.6)	25.2 (40.4)	2.66e−02	+
Amygdala	0.572	0.743	0.401	20.5 (19.4)	12.3 (15.0)	7.54e−02	o
Caudate	0.557	0.692	0.421	12.9 (13.9)	7.5 (6.9)	4.54e−02	+
Cerebellum exterior	0.572	0.706	0.438	6.7 (4.4)	4.5 (4.9)	1.07e−01	o
Cerebellum white matter	0.599	0.657	0.541	10.2 (7.4)	6.7 (4.6)	2.62e−02	+
Cerebral white matter	0.517	0.797	0.238	7.7 (15.3)	4.9 (9.3)	3.75e−01	o
Hippocampus	0.578	0.774	0.383	16.4 (16.1)	8.4 (6.9)	7.84e−03	++
Inf lat ventricle	0.574	0.679	0.468	61.7 (48.1)	38.2 (25.1)	1.42e−02	+
Lateral ventricle	0.644	0.814	0.474	33.5 (31.1)	14.9 (10.3)	8.10e−04	++
Pallidum	0.574	0.727	0.421	22.9 (23.1)	11.6 (11.5)	1.20e−02	+
Putamen	0.598	0.877	0.319	32.2 (51.2)	8.0 (10.2)	4.06e−03	++
Thalamus	0.590	0.850	0.331	24.2 (24.1)	10.8 (16.4)	1.33e−02	+
Ventral DC	0.367	0.409	0.325	8.0 (4.2)	7.9 (5.0)	9.25e−01	o
Forebrain	0.478	0.531	0.424	28.7 (24.9)	26.5 (21.3)	7.25e−01	o

Given the updated model parameters Φ we can then estimate the class probabilities in the next iteration as:

$$p_{ik}^{(m+1)} = \frac{f(y_i|\mathbf{z_i} = \mathbf{e_k}, \Phi^{(m)}) f(\mathbf{z_i} = \mathbf{e_k}|p_{\mathscr{G}_i}^{(m)}, G)}{\sum_{j=1}^{K} f(y_i|\mathbf{z_i} = \mathbf{e_j}, \Phi^{(m)}) f(\mathbf{z_i} = \mathbf{e_j}|p_{\mathscr{G}_i}^{(m)}, G)} \quad \text{(D.7)}$$

Usually the model converges after a few iterations. In our experiments, we have performed ten iterations to better control the runtime of the algorithm. The parameters for describing the MRF were set to $\beta = 1.0$ and $\gamma = 1.5$. The background was modelled with an explicit class.

Appendix E. Individual non-cortical similarity indices and intraclass correlation coefficients

Table E.9.

Appendix F. Classification results obtained based on absolute asymmetry indices of individual structures

Tables F.10, F.11, F.12.

Appendix G. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.media.2014.12.003.

References

Aljabar, P., Heckemann, R.A., Hammers, A., Hajnal, J.V., Rueckert, D., 2009. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. NeuroImage 46, 726–738.

Andersen, S.M., Rapcsak, S.Z., Beeson, P.M., 2010. Cost function masking during normalization of brains with focal lesions: still a necessity?. NeuroImage 53, 78–84.

Artaechevarria, X., Munoz Barrutia, A., Ortiz, C.d.S., 2009. Combination strategies in multi-atlas image segmentation: application to brain MR data. IEEE Trans. Med. Imag. 28, 1266–1277.

Asman, A.J., Landman, B.A., 2011. Robust statistical label fusion through consensus level, labeler accuracy, and truth estimation (COLLATE). IEEE Trans. Med. Imag. 30, 1779–1794.

Asman, A.J., Landman, B.A., 2013. Non-local statistical label fusion for multi-atlas segmentation. Med. Image Anal. 17, 194–208.

Asman, A.J., Chambless, L.B., Thompson, R.C., Landman, B.A., 2013. Out-of-atlas likelihood estimation using multi-atlas segmentation. Med. Phys. 40, 043702-01–043702-10.

Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. 12, 26–41.

Bauer, S., Wiest, R., Nolte, L.P., Reyes, M., 2013. A survey of MRI-based medical image analysis for brain tumor studies. Phys. Med. Biol. 58, R97–R129.

Bendlin, B.B., Ries, M.L., Lazar, M., Alexander, A.L., Dempsey, R.J., Rowley, H.A., Sherman, J.E., Johnson, S.C., 2008. Longitudinal changes in patients with traumatic brain injury assessed with diffusion-tensor and volumetric imaging. NeuroImage 42, 503–514.

Bigler, E.D., 2001. Quantitative magnetic resonance imaging in traumatic brain injury. J. Head Trauma Rehab. 16, 117–134.

Bonilha, L., Nesland, T., Rorden, C., Fridriksson, J., 2014. Asymmetry of the structural brain connectome in healthy older adults. Front. Psychiat. 4, 186.

Brett, M., Leff, A.P., Rorden, C., Ashburner, J., 2001. Spatial normalization of brain images with focal lesions using cost function masking. NeuroImage 14, 486–500.

Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M., 2010. The balanced accuracy and its posterior distribution. In: 20th International Conference on Pattern Recognition (ICPR), 2010, pp. 3121–3124..

Cabezas, M., Oliver, A., Lladó, X., Freixenet, J., Bach Cuadra, M., 2011. A review of atlas-based segmentation for magnetic resonance brain images. Comput. Methods Prog. Biomed. 104, e158–e177.

Cardoso, M.J., Clarkson, M.J., Ridgway, G.R., Modat, M., Fox, N.C., Ourselin, S., 2011. LoAd: a locally adaptive cortical segmentation algorithm. NeuroImage 56, 1386–1397.

Cardoso, M.J., Leung, K., Modat, M., Keihaninejad, S., Cash, D., Barnes, J., Fox, N.C., Ourselin, S., 2013a. STEPS: similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation. Med. Image Anal. 17, 671–684.

Cardoso, M.J., Melbourne, A., Kendall, G.S., Modat, M., Robertson, N.J., Marlow, N., Ourselin, S., 2013b. AdaPT: an adaptive preterm segmentation algorithm for neonatal brain MRI. NeuroImage 65, 97–108.

Cardoso, M.J., Modat, M., Ourselin, S., 2013c. BrianGraph: tissue segmentation using the geodesic information flows framework. In: Proceedings of MICCAI Challenge Workshop on Segmentation: Algorithms, Theory and Applications (SATA), pp. 23–32..

Chitphakdithai, N., Duncan, J.S., 2010. Non-rigid registration with missing correspondences in preoperative and postresection brain images. In: Medical Image Computing and Computer-Assisted Intervention MICCAI 2010. Lecture Notes in Computer Science, vol. 6361, pp. 367–374..

Coupé, P., Manjón, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L., 2011. Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. NeuroImage 54, 940–954.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26, 297–302.

Ding, K., Marquez de la Plata, C., Wang, J.Y., Mumphrey, M., Moore, C., Harper, C., Madden, C.J., McColl, R., Whittemore, A., Devous, M.D., Diaz-Arrastia, R., 2008. Cerebral atrophy after traumatic white matter injury: correlation with acute neuroimaging and outcome. J. Neurotrauma 25, 1433–1440.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341–355.

Galaburda, A.M., Corsiglia, J., Rosen, G.D., Sherman, G.F., 1987. Planum temporale asymmetry, reappraisal since geschwind and levitsky. Neuropsychologia 25, 853–868.

Hammers, A., Allom, R., Koepp, M., Free, S.L., Myers, R., Lemieux, L., Mitchell, T.N., Brooks, D.J., Duncan, J.S., 2003. Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. Hum. Brain Mapp. 19, 224–247.

Heckemann, R.A., Hajnal, J.V., Aljabar, P., Rueckert, D., Hammers, A., 2006. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage 33, 115–126.

Heckemann, R.A., Keihaninejad, S., Aljabar, P., Rueckert, D., Hajnal, J.V., Hammers, A., 2010. Improving intersubject image registration using tissue-class information benefits robustness and accuracy of multi-atlas based anatomical segmentation. NeuroImage 51, 221–227.

Heckemann, R.A., Keihaninejad, S., Aljabar, P., Gray, K.R., Nielsen, C., Rueckert, D., Hajnal, J.V., Hammers, A.The AD Neuroimaging Initiative, 2011. Automatic morphometry in Alzheimer's disease and mild cognitive impairment. NeuroImage 56, 2024–2037.

Irimia, A., Chambers, M.C., Alger, J.R., Filippou, M., Prastawa, M.W., Wang, B., Hovda, D.A., Gerig, G., Toga, A.W., Kikinis, R., Vespa, P.M., Van Horn, J.D., 2011. Comparison of acute and chronic traumatic brain injury using semi-automatic multimodal segmentation of MR volumes. J. Neurotrauma 28, 2287–2306.

Irimia, A., Wang, B., Aylward, S.R., Prastawa, M.W., Pace, D.F., Gerig, G., Hovda, D.A., Kikinis, R., Vespa, P.M., Van Horn, J.D., 2012. Neuroimaging of structural pathology and connectomics in traumatic brain injury: toward personalized outcome prediction. NeuroImage: Clin. 1, 1–17.

Jennett, B., Bond, M., 1975. Assessment of outcome after severe brain damage: a practical scale. The Lancet 306, 480–484.

Kempton, M.J., Underwood, T.S., Brunton, S., Stylios, F., Schmechtig, A., Ettinger, U., Smith, M.S., Lovestone, S., Crum, W.R., Frangou, S., Williams, S.C.R., Simmons, A., 2011. A comprehensive testing protocol for MRI neuroanatomical segmentation techniques: evaluation of a novel lateral ventricle segmentation method. NeuroImage 58, 1051–1059.

Landman, B.A., Warfield, S.K., 2012. MICCAI 2012 Workshop on Multi-Atlas Labeling. <https://masi.vuse.vanderbilt.edu/workshop2012/images/c/c8/miccai_2012_workshop_v2.pdf>..

Landman, B.A., Asman, A.J., Scoggins, A.G., Bogovic, J.A., Xing, F., Prince, J.L., 2012. Robust statistical fusion of image labels. IEEE Trans. Med. Imag. 31, 512–522.

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

Ledig, C., Shi, W., Bai, W., Rueckert, D., 2014. Patch-based evaluation of image segmentation. In: Proceedings of CVPR, pp. 3065–3072..

Liu, X., Niethammer, M., Kwitt, R., McCormick, M., Aylward, S., 2014. Low-rank to the rescue atlas-based analyses in the presence of pathologies. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013, Lecture Notes in Computer Science, vol. 8675, pp. 97–104..

Lötjönen, J.M., Wolz, R., Koikkalainen, J.R., Thurfjell, L., Waldemar, G., Soininen, H., Rueckert, D., 2010. Fast and robust multi-atlas segmentation of brain magnetic resonance images. NeuroImage 49, 2352–2365.

Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2007. Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosci. 19, 1498–1507.

Marshall, L.F., Bowers Marshall, S., Klauber, M.R., van Berkum Clark, M., Eisenberg, H.M., Jane, J.A., Luerssen, T.G., Marmarou, A., Foulkes, M.A., 1991. A new classification of head injury based on computerized tomography. J. Neurosurg. 75, S14–S20.

Meythaler, J.M., Peduzzi, J.D., Eleftheriou, E., Novack, T.A., 2001. Current concepts: diffuse axonal injury-associated traumatic brain injury. Arch. Phys. Med. Rehab. 82, 1461–1471.

Modat, M., Ridgway, G.R., Taylor, Z.A., Lehmann, M., Barnes, J., Hawkes, D.J., Fox, N.C., Ourselin, S., 2010. Fast free-form deformation using graphics processing units. Comput. Methods Prog. Biomed. 98, 278–284.

Niethammer, M., Hart, G.L., Pace, D.F., Vespa, P.M., Irimia, A., Van Horn, J.D., Aylward, S.R., 2011. Geometric metamorphosis. In: Medical Image Computing and Computer-Assisted Intervention MICCAI 2011, Lecture Notes in Computer Science, vol. 6892, pp. 639–646..

Nugent, A.C., Luckenbaugh, D.A., Wood, S.E., Bogers, W., Zarate, C.A., Drevets, W.C., 2013. Automated subcortical segmentation using FIRST: test–retest reliability, interscanner reliability, and comparison to manual segmentation. Hum. Brain Mapp. 34, 2313–2329.

Periaswamy, S., Farid, H., 2006. Medical image registration with partial data. Med. Image Anal. 10, 452–464.

Ramlackhansingh, A.F., Brooks, D.J., Greenwood, R.J., Bose, S.K., Turkheimer, F.E., Kinnunen, K.M., Gentleman, S., Heckemann, R.A., Gunanayagam, K., Gelosa, G., Sharp, D.J., 2011. Inflammation after trauma: microglial activation and traumatic brain injury. Ann. Neurol. 70, 374–383.

Rao, A., Ledig, C., Newcombe, V., Menon, D., Rueckert, D., 2014. Contusion segmentation from subjects with traumatic brain injury: a random forest framework. In: Proceedings of ISBI 2014, pp. 333–336..

Ribbens, A., Hermans, J., Maes, F., Vandermeulen, D., Suetens, P., 2014. Unsupervised segmentation, clustering, and groupwise registration of heterogeneous populations of brain MR images. IEEE Trans. Med. Imag. 33, 201–224.

Rohlfing, T., Brandt, R., Menzel, R., Maurer Jr., C.R., 2004a. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. NeuroImage 21, 1428–1442.

Rohlfing, T., Russakoff, D.B., Maurer, C.R., 2004b. Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. IEEE Trans. Med. Imag. 23, 983–994.

Rousseau, F., Habas, P.A., Studholme, C., 2011. A supervised patch-based approach for human brain labeling. IEEE Trans. Med. Imag. 30, 1852–1862.

Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L.G., Leach, M.O., Hawkes, D.J., 1999. Nonrigid registration using free-form deformations: application to breast MR images. IEEE Trans. Med. Imag. 18, 712–721.

Sabuncu, M.R., Yeo, B.T.T., Van Leemput, K., Fischl, B., Golland, P., 2010. A generative model for image segmentation based on label fusion. IEEE Trans. Med. Imag. 29, 1714–1729.

Shenton, M.E., Hamoda, H.M., Schneiderman, J.S., Bouix, S., Pasternak, O., Rathi, Y., Vu, M.A., Purohit, M.P., Helmer, K., Koerte, I., Lin, A.P., Westin, C.F., Kikinis, R., Kubicki, M., Stern, R.A., Zafonte, R., 2012. A review of magnetic resonance imaging and diffusion tensor imaging findings in mild traumatic brain injury. Brain Imag. Behav. 6, 137–192.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86, 420–428.

Stefanescu, R., Commowick, O., Malandain, G., Bondiau, P.Y., Ayache, N., Pennec, X., 2004. Non-rigid atlas to subject registration with pathologies for conformal brain radiotherapy. In: Medical Image Computing and Computer-Assisted Intervention MICCAI 2004, Lecture Notes in Computer Science, vol. 3216, pp. 704–711..

Strangman, G.E., O'Neil-Pirozzi, T.M., Supelana, C., Goldstein, R., Katz, D.I., Glenn, M.B., 2010. Regional brain morphometry predicts memory rehabilitation outcome after traumatic brain injury. Front. Hum. Neurosci. 4, 182.

Teasdale, G., Jennett, B., 1974. Assessment of coma and impaired consciousness: a practical scale. The Lancet 2, 81–84.

Tong, T., Wolz, R., Coupé, P., Hajnal, J.V., Rueckert, D., 2013. Segmentation of MR images via discriminative dictionary learning and sparse coding: application to hippocampus labeling. NeuroImage 76, 11–23.

Tustison, N., Avants, B., Cook, P., Zheng, Y., Egan, A., Yushkevich, P., Gee, J., 2010. N4ITK: improved N3 bias correction. IEEE Trans. Med. Imag. 29, 1310–1320.

van der Lijn, F., den Heijer, T., Breteler, M.M.B., Niessen, W.J., 2008. Hippocampus segmentation in MR images using atlas registration, voxel classification, and graph cuts. NeuroImage 43, 708–720.

Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 1999. Automated model-based tissue classification of MR images of the brain. IEEE Trans. Med. Imag. 18, 897–908.

Wang, H., Das, S.R., Suh, J.W., Altinay, M., Pluta, J., Craige, C., Avants, B.B., Yushkevich, P.A., 2011. A learning-based wrapper method to correct systematic errors in automatic image segmentation: consistently improved performance in hippocampus, cortex and brain. NeuroImage 55, 968–985.

Wang, H., Suh, J.W., Das, S.R., Pluta, J., Craige, C., Yushkevich, P.A., 2013. Multi-atlas segmentation with joint label fusion. IEEE Trans. Pattern Anal. Mach. Intell. 35, 611–623.

Wang, L., Shi, F., Li, G., Gao, Y., Lin, W., Gilmore, J.H., Shen, D., 2014. Segmentation of neonatal brain MR images using patch-driven level sets. NeuroImage 84, 141–158.

Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imag. 23, 903–921.

Warner, M.A., de la Plata, C.M., Spence, J., Wang, J.Y., Harper, C., Moore, C., Devous, M., Diaz-Arrastia, R., 2010a. Assessing spatial relationships between axonal integrity, regional brain volumes, and neuropsychological outcomes after traumatic axonal injury. J. Neurotrauma 27, 2121–2130.

Warner, M.A., Youn, T.S., Davis, T., Chandra, A., de la Plata, M.C., Moore, C., Harper, C., Madden, C.J., Spence, J., McColl, R., Devous, M., King, R.D., Diaz-Arrastia, R., 2010b. Regionally selective atrophy after traumatic axonal injury. Arch. Neurol. 67, 1336–1344.

Wells, W.M.I., Grimson, W.E.L., Kikinis, R., Jolesz, F.A., 1996. Adaptive segmentation of MRI data. IEEE Trans. Med. Imag. 15, 429–442.

Wolz, R., Aljabar, P., Hajnal, J.V., Hammers, A., Rueckert, D., 2010a. LEAP: learning embeddings for atlas propagation. NeuroImage 49, 1316–1325.

Wolz, R., Heckemann, R.A., Aljabar, P., Hajnal, J.V., Hammers, A., Lötjönen, J., Rueckert, D., 2010b. Measurement of hippocampal atrophy using 4D graph-cut segmentation: application to ADNI. NeuroImage 52, 109–118.

Zacharaki, E.I., Hogea, C.S., Shen, D., Biros, G., Davatzikos, C., 2009. Non-diffeomorphic registration of brain tumor images by simulating tissue loss and tumor growth. NeuroImage 46, 762–774.

Zhang, J., 1992. The mean field theory in EM procedures for Markov random fields. IEEE Trans. Signal Process. 40, 2570–2583.

Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation maximization algorithm. IEEE Trans. Med. Imag. 20, 45–57.