NeuroImage 203 (2019) 116206

Contents lists available at ScienceDirect

NeuroImage

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

Non-linear realignment improves hippocampus subfield segmentation reliability

Thomas B. Shaw [a,*], Steffen Bollmann [a], Nicole T. Atcheson [a], Lachlan T. Strike [b], Christine Guo [c], Katie L. McMahon [d], Jurgen Fripp [e], Margaret J. Wright [b], Olivier Salvado [e,f], Markus Barth [a,g]

[a] Centre for Advanced Imaging, The University of Queensland, Brisbane, Australia
[b] Queensland Brain Institute, The University of Queensland, Brisbane, Australia
[c] Mental Health Program, QIMR Berghofer Medical Research Institute, Brisbane, Australia
[d] School of Clinical Sciences and Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, Australia
[e] CSIRO Health and Biosecurity, Herston, Australia
[f] CSIRO Data61, Sydney, Australia
[g] School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, Australia

ARTICLE INFO

Keywords:
Hippocampus subfields
Realignment
Motion correction
Magnetic resonance imaging
Cornu ammonis
Segmentation

ABSTRACT

Participant movement can deleteriously affect MR image quality. Further, for the visualization and segmentation of small anatomical structures, there is a need to improve image quality, specifically signal-to-noise ratio (SNR) and contrast-to-noise ratio (CNR), by acquiring multiple anatomical scans consecutively. We aimed to ameliorate movement artefacts and increase SNR in a high-resolution turbo spin-echo (TSE) sequence acquired thrice using non-linear realignment in order to improve segmentation consistency of the hippocampus subfields. We assessed the method in 29 young healthy participants, 11 Motor Neuron Disease patients, and 11 age matched controls at 7T, and 24 healthy adolescents at 3T. Results show improved image segmentation of the hippocampus subfields when comparing template-based segmentations with individual segmentations with Dice overlaps $N = 75$; $ps < 0.001$ (Friedman's test) and higher sharpness $ps < 0.001$ in non-linearly realigned scans as compared to linearly, and arithmetically averaged scans.

1. Introduction

Imaging the subfields of the hippocampal formation with Magnetic Resonance Imaging (MRI) has garnered great interest in recent years. Characterising subfield tissue may elucidate susceptibility to – and provide more sensitive biomarkers for - neurodegenerative diseases including Alzheimer's disease (AD) (Balachandar et al., 2015; Boutet et al., 2014; Henry et al., 2011; Jacobsen et al., 2017; Kerchner et al., 2012; La Joie et al., 2013; Maruszak and Thuret, 2014; Pluta et al., 2012), as the subfields contain different cell types and functional properties (Duvernoy et al., 2013). Increasing spatial resolution for MRI through improved acquisition and/or post-processing techniques improves segmentation of various tissue types or anatomical structures (Thomas et al., 2008). Consequently, there is a need to improve signal-to-noise ratio (SNR) of scans by going to higher field strengths (ultra-high field [UHF] MRI, 7T and above) and by acquiring multiple repetitions of the same scan in a single session, which leads to longer acquisition times.

Previous UHF *in vivo* hippocampus subfield segmentation studies (for review, see Giuliano et al., 2017) have used dedicated T2-weighted Turbo-Spin Echo (TSE), Gradient Echo, or multi-echo sequences that, due to multiple refocusing pulses, exhibit differing contrast- and intensity characteristics for different tissue classes, and consequently the laminae of the hippocampus (Marques and Norris, 2018; Winterburn et al., 2013). With higher field strengths and better coils comes the possibility of increasing the number of repetitions due to shorter scan times per acquisition. Previously, we utilised three repetitions of a 4 min TSE sequence (12 min total acquisition time) that cover the whole hippocampus with a high in-plane resolution of $0.4\,mm^2 \times 0.4\,mm^2$ and a $0.8\,mm^2$ slice thickness. These acquisitions were combined to yield a minimum deformation average model for investigating hippocampus subfields (Jacobsen et al., 2017). The 4-min TSE sequence was designed to be repeated three times to both boost SNR and to be able to potentially discard one of the repetitions in case of participant movement. Unfortunately, finding the optimal realignment between these anatomical

* Corresponding author. Centre for Advanced Imaging, Building 57, Research Road, St Lucia, 4072, Australia.
E-mail address: t.shaw@uq.edu.au (T.B. Shaw).

https://doi.org/10.1016/j.neuroimage.2019.116206
Received 4 April 2019; Received in revised form 14 September 2019; Accepted 17 September 2019
Available online 17 September 2019

scans can be difficult due to gradient nonlinearities (Reuter et al., 2010) especially at ultra-high field. Reuter et al. (2010) used rigid registrations for realignment of anatomical T1-weighted (T1w) scans in an intermediate space as rigid parts of the brain and skull are assumed to go unchanged in the short time between scan repetitions (save for rigid location changes). At higher field strengths and resolutions, noise from movement interacts with gradient nonlinearities and may cause non-linear distortions in images unable to be corrected by rigid registrations alone.

Participant movement is more likely for long MRI scans and can deteriorate image quality (Kochunov et al., 2006), potentially leading to unusable scans due to movement artefacts, and subsequently increased costs. Further, patients with neurodegenerative disorders, elderly, young, and highly anxious participants, have a high propensity for movement in the scanner (Yushkevich et al., 2015). Previous work examining the benefits of motion correction in anatomical MRI have used either prospective or retrospective realignment techniques. Prospective techniques (Maclaren et al., 2013) mainly utilise navigators and motion tracking devices to correct for motion online. Retrospective realignment techniques may take magnitude data and attempt to estimate transformations between images to attain good registrations between them (e.g., Reuter et al., 2010), similar to fMRI realignment.

In human neuroimaging with a supine patient position, rotation commonly occurs at the posterior of the head, resulting in rotation-related blurring and movement artefacts progressing in severity anteriorly (Maclaren et al., 2013). Blurring and ghosting artefacts are somewhat prevalent in imaging the hippocampus *in vivo,* which has tightly packed laminae and multiple tissue contrast signals that blur together causing partial volume effects. This motion is inherent to the longer acquisition times necessary to capture the structure of the hippocampus (Marrakchi-Kacem et al., 2016). Mitigating movement artefacts in image post-processing is therefore an important challenge to overcome in neuroimaging.

Previous work (Bollmann et al., 2017) has shown that fMRI realignment can be completed using an iterative averaging model (or template) to account for movement in participants and decrease partial volume effects. We aim to utilise a similar approach on anatomical scans with high resolution in order to reduce partial volume effects and increase SNR. Due to the small size of the hippocampus and its laminar structure, partial volume effects can lead to misclassification of subfields. To ameliorate movement artefacts, boost SNR and sharpness, and improve image segmentation reliability, we implemented a retrospective realignment technique that iteratively estimates non-linear transformations between the three TSE images to fit to an evolving model of anatomical consistency to attain improved reliability in segmentation results.

Labelling the subfields of the hippocampus with automatic segmentation strategies is now feasible with software including Automatic Segmentation of Hippocampal Subfields (ASHS; Yushkevich et al., 2015), and Freesurfer (Iglesias et al., 2015) all able to segment hippocampal subfields of conventional T1w and dedicated T2w TSE scans in multi-contrast approaches.

We aimed to explore the effectiveness of our technique in Young Healthy Participants (YHPs), patients with Motor Neuron Disease (MND) and their healthy aged-matched controls (HCs) at 7T, and an adolescent population at 3T by assessing registration consistency. Intuitively, image registration algorithms are more successful in images with more refined details, higher sharpness, and better SNR. Therefore, greater registration consistency will occur between participants if the registration procedure is more effective. If the technique is robust to movement, between-subject registrations will converge more readily due to increased boundary delineation and SNR increases. Thus, we hypothesise that each participant's non-linearly realigned (1: *Non-Linear*) registration procedure would produce higher segmentation consistency than a linear procedure (similar to fMRI realignment; 2: *Linear*) or a simple averaging procedure without a separate alignment (3: *Average*). This is based on the assumption that automatic segmentation algorithms rely on high SNR/CNR and good boundary delineation to function reliably. Finally, due to reduced movement artefacts and higher SNR after processing, we expect that non-linear realignment would produce the sharpest images compared to linear realignment or simple averaging.

2. Methods

Using a 7T whole-body research scanner (Siemens Healthcare, Erlangen, Germany), with maximum gradient strength of 70 mT/m and a slew rate of 200 mT/m/s and a 7T Tx/32 channel Rx head array (Nova Medical, Wilmington, MA, USA) we acquired a 2D TSE sequence (Siemens WIP tse_UHF_WIP729C. Variant: tse2d1_9, TR: 10300 ms, TE: 102 ms, FA: 132°, FoV: 220 mm, voxel size of $0.4 \times 0.4 \times 0.8\ mm^3$ Turbo factor of 9; iPAT (GRAPPA) factor 2); acquisition time (TA) 4 min 12 s, three times of a slab aligned orthogonally to the hippocampus. We scanned three 'groups' at 7T: 11 patients diagnosed with MND (age, M = 59.36, SD = 7.65), 11 age-matched control participants (HCs; age, M = 60.23, SD = 7.65), and 29 young healthy participants (YHPs; age, M = 26.31, SD = 0.66) for a total of 51 participants at 7T in order to test the robustness of the realignment on a wide range of participants in terms of age and disease states (reflected by different levels of neurodegeneration) and movement probability (patients vs. controls). An anatomical whole-brain T1w scan was acquired using a prototype MP2RAGE sequence (WIP 900; Marques et al., 2010; O'Brien et al., 2014) at 0.9 mm isotropic voxel size (TR/TE/TIs = 4300 ms/2.5 ms/840 ms, 2370 ms, TA: 6:54).

We also acquired images for 24 adolescent participants (age, M = 11.25, SD = 0.98) using a 3T whole-body scanner (PRISMA, Siemens Healthcare, Erlangen, Germany). We acquired a similar 2D TSE sequence (TR: 8460 ms, TE: 67 ms, FA: 150°, FoV: 192 mm, voxel size of $0.5 \times 0.5 \times 1.0\ mm^3$; interpolated in k-space to $0.25 \times 0.25 \times 1.0\ mm^3$ Turbo factor of 13; TA: 4:24) three times (totalling approximately 13 min) and a prototype MP2RAGE (WIP 900c, Variant VE11C [Marques et al., 2010; O'Brien et al., 2014];) with an isotropic resolution of $0.8\ mm^3$ (TR/TE/TIs = 4000 ms/2.9 ms/700 ms/2220 ms, TA: 6:04).

All TSEs were pre-processed by resampling to 0.3 mm isotropic, bias field corrected using N4 (ANTs Version: 2.2.0. dev116-gabc03; Tustison et al., 2010), skull stripped using the co-registered T1w as an initial mask after usage of ROBEX (Iglesias et al., 2011), and intensity normalised between two percent-critical thresholds using NiftiNorm (https://github.com/thomshaw92/nifti_normalise), a Nifti implementation of mincnorm from the medical imaging network common data toolkit (Vincent et al., 2016). To assess registration and segmentation consistency, we tested three different approaches (or 'methods'; Fig. 1): we performed no additional realignment and averaged the slabs arithmetically using ANTs AverageImage (1. *Average)*. We also registered TSEs linearly by: i) concatenating images in time using FSL (Jenkinson et al., 2012), ii) estimating affine registrations between each image using FSL's MCFLIRT (Jenkinson et al., 2002) and iii) averaging the image in the time dimension (2. *Linear*). Third, we performed non-linear registrations using ANTs' (Avants et al., 2010) SyN (Avants et al., 2008) registration (Fig. 1), with the three TSE scans being iteratively deformed using antsMultivariateTemplateConstruction2. sh into a minimum deformation average (MDA) template containing only anatomically consistent features (3. *Non-Linear*). This method involves averaging the TSE scans to an intermediate space and registering the individual TSEs to the intermediate space. The average of these transformations is applied to the intermediate template and the model is updated iteratively. We used the default settings and three iterations for the non-linear registrations. We deactivated the ANTs laplacian sharpness filter for fair comparison between realignment methods.

Commonly, automatic segmentation precision and accuracy is compared to manual segmentations of the same subject. However, manual hippocampus subfield segmentation is time and labour intensive, taking up to 8 h initially, and 2 h after five months of training (Wisse

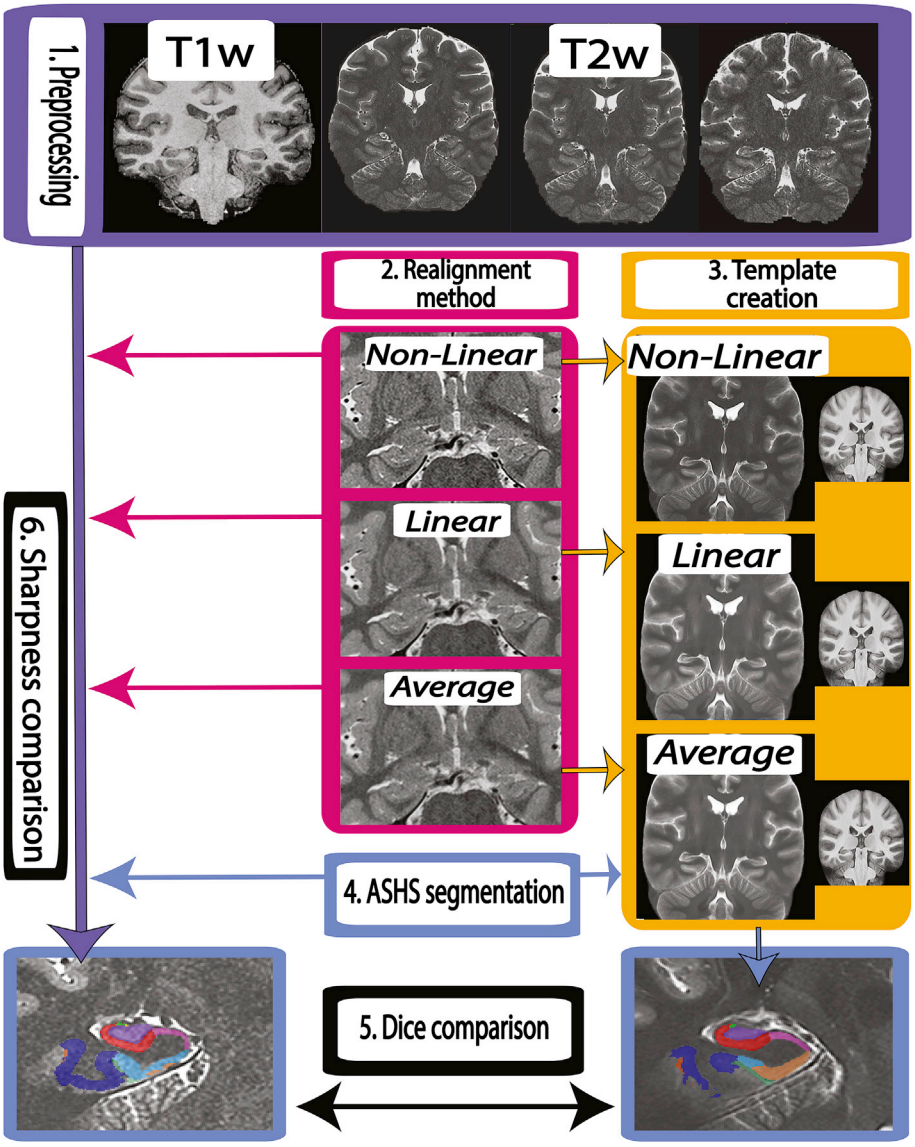

Fig. 1. Processing pipeline for assessing segmentation consistency between realignment conditions. Input images (purple) are preprocessed before the three TSE realignment methods are performed (in pink, Non-Linear, Linear, Average). Next, templates are constructed for each of the methods (yellow). Both the individuals and the templates are segmented using ASHS in a common space (blue) before Dice comparisons (black). The sharpness for individual images is also measured for each method (6, in black). All views show hippocampi for a coronal single-subject view. This processing pipeline was performed for each method (Non-Linear, Linear, and Average) and each group (YHPs, MND, HCs, 3T adolescents).

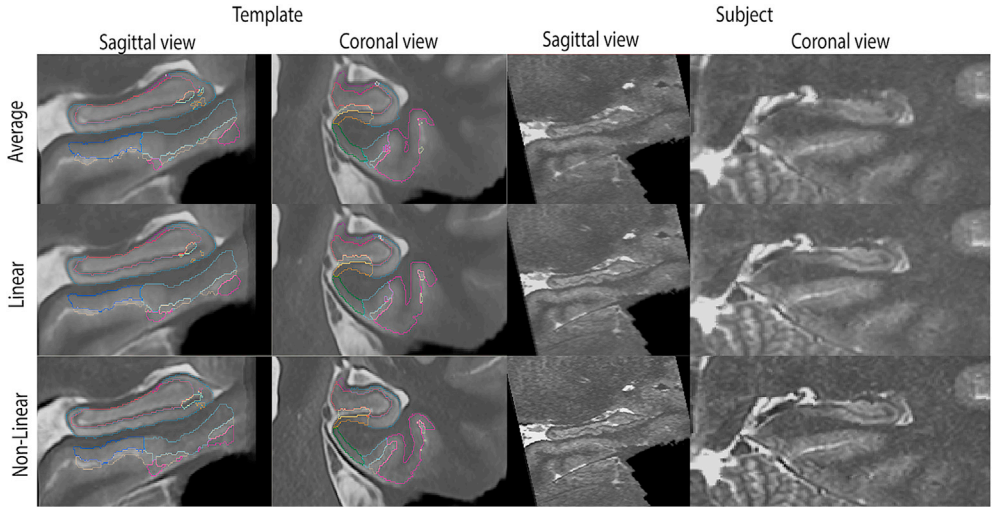

Fig. 2. Sagittal and coronal views of the alignment in the hippocampus after the three realignment procedures (Non-Linear, Linear, Average) for the method-templates (left, labelled) and a representative YHP subject that benefited from the non-linear realignment technique (right). Views capture the left hippocampus, and both hippocampi in a coronal single-subject view. Edges of the ASHS template segmentations have been superimposed as coloured lines.

et al., 2016) and is prone to inter- and intra-rater variability (Boccardi et al., 2011; Hsu et al., 2002; Mulder et al., 2014). Automatic segmentation algorithms have repeatedly been shown to accurately segment hippocampus subfields (Iglesias et al., 2015; Romero et al., 2017; Wisse et al., 2016; Yushkevich et al., 2015). Therefore, measuring the consistency and precision of segmentations within a group, rather than their accuracy, is a possible alternative for determining the effectiveness of preprocessing techniques.

In the present study, quantitative metrics for examination of the most effective registration technique were chosen based in part on the work of Fonov and Collins (2018). It was determined that, much like in the original MP2RAGE paper (Marques et al., 2010), a useful metric for examining the effectiveness of a registration technique was through measuring segmentation performance and consistency (as qualitative analysis of motion correction is not sufficient).

Individuals from each group (YHPs, HCs, MNDs, 3T adolescents) for each method (*Non-Linear, Linear, Average*) were co-registered to a group-and-method-specific MDA template using ANTs. These templates were multivariate, having a T1w and T2w image component. Overall, this yielded twelve group-and-method templates (e.g., MND-*Linear* template, HC-*Non-Linear* template, YHP-*Average* template, etc.). Each template was constructed with identical parameters. As the template building process yielded a T1w and T2w image, the twelve group-and-method templates were independently segmented using ASHS (V2.0 with the Penn Memory Centre 3T ASHS Atlas; Yushkevich et al., 2015). Individuals' realigned TSE (for each method) and their corresponding MP2RAGE scan were labelled using ASHS, which requires both T1w and T2w inputs in native space then warped to their common group-method template space (Fig. 1).

Segmentation consistency was derived by examining Dice overlaps of segmentation labels between group-method template labels and individual volume labels (subject-to-template overlaps). We then compared Dice between *Non-Linear, Linear,* and *Average* using non-parametric statistics.

Sharpness of images was measured to describe brain structure delineation by calculating the median of the derivative of a Gaussian applied to the TSE at 1 mm FWHM as described in Fonov and Collins (2018). We provide all the code for this project at https://github.com/thomshaw92/NonLinRegImprovesSegAcc.

For validation, all TSEs were rated manually using adapted methods from Backhausen et al. (2016) and Jones and Marietta (2012), with the criteria of i) Image sharpness (including artefacts), ii) Ringing, iii) Contrast-to-Noise Ratio (CNR; subcortical structures), and 4) CNR (GM and WM). Two raters (NTA: senior research radiographer with 30 years medical imaging experience, and TBS: six years research experience in medical imaging) rated all participants on a scale from 1 to 3, with 1 = pass, 2 = check, and 3 = fail for quality assurance of scans and to assess motion artefacts before post-processing. A weighted average that favoured ratings of sharpness and subcortical CNR was used to summarise the findings. Kohen's Cappa estimates were calculated using SPSS Statistics Package Version 25 (SPSS Inc, Chicago, IL). Participants for the 3T study were chosen from a larger study based on movement artefacts determined by a third rater (LS, seven years research experience in medical imaging). All selected scans showed motion artefacts.

3. Results

To assess the performance of the methods at mitigating motion artefacts, manual ratings for quality assurance were first checked for inter-rater reliability. Cohen's K was run to determine the agreement between the two raters' judgements on participant movement. There was moderate agreement between the two raters' judgements $K = 0.536$. Individual criteria Cohen's K values were also in the moderate (0.55+) to good (0.73) range. We found that the MND and the older aged HCs performed worse in their motion rating scores than YHPs (Fig. 3). MND and HCs had more ringing and other artefacts, worse CNR in both subcortical structures and between WM and GM boundaries, and lower sharpness. We also observed 8 YHP with anatomical variations including hippocampal cysts and incomplete hippocampal inversions. We observed severe neurodegeneration in MND patients.

Fig. 4 shows a single-subject example of registration results for the TSEs. Sharper edges and more subfield information is observable in the *Non-Linear* method, followed by *Linear*, and very little information in *Average* for an example with motion between scans. When there is limited movement, small differences can be observed between the three methods.

Collapsing across groups, Dice overlap scores (Fig. 5) were found to be significantly different between all groups, (Friedman test $p < 0.001$). Significantly higher overlaps for *Non-Linear* with its template were found compared to both *Linear* and *Average,* and comparing *Linear* to *Average*, independently (N = 51; ** = $ps < 0.001$, Wilcoxon rank sum tests). When comparing within groups, MNDs, HCs, and YHPs showed the same pattern of results as when collapsing across groups (i.e., *Non-Linear* > *Linear* > *Average*). In the 3T group, Dice overlaps were higher for *Non-Linear* than *Average,* and *Linear* Dice overlaps were higher than *Average* (N = 24; $ps < 0.001$); while no significant difference between *Non-Linear* and *Linear* Dice overlaps were found ($p = 1.00$).

Interestingly, in a subset of participants (N = 1 for MND, and N = 8 for YHP) participants, *Linear* and/or *Average* registration techniques outperformed *Non-Linear* for Dice overlap. A case-by-case analysis of these participants was performed. In eight of these cases, there was an overall Dice failure (Dice = < 0.7 [Zou et al., 2004]) for template overlaps for all methods (i.e., *Linear* and *Average* included). We found five of the nine

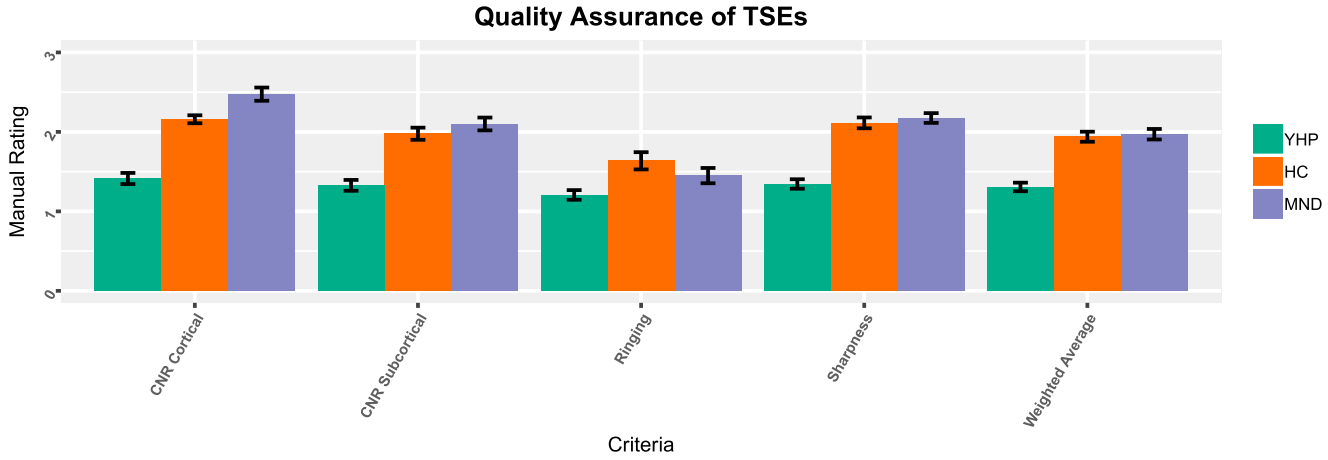

Fig. 3. Quality assurance of TSEs for Young Healthy Participants (YHPs), Healthy Controls (HCs), and Motor Neuron Disease patients (MND) patients. Average ratings (from 1 = pass to 3 = fail) with standard error of the mean are shown for the four assessed criteria and weighted average.

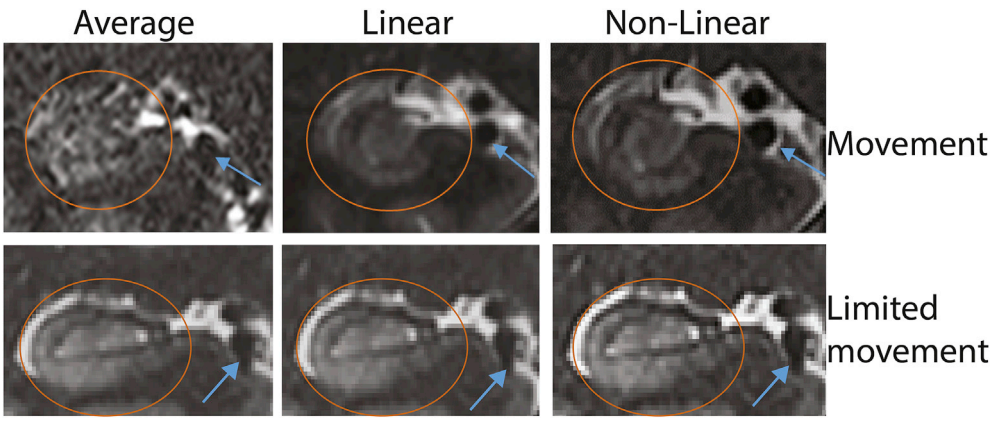

Fig. 4. Left to right: Average, Linear, and Non-Linear realignment examples from a participant with more movement (top) and less movement (bottom) with ellipses over the right hippocampus (coronal plane); blue arrows denote a common vessel.

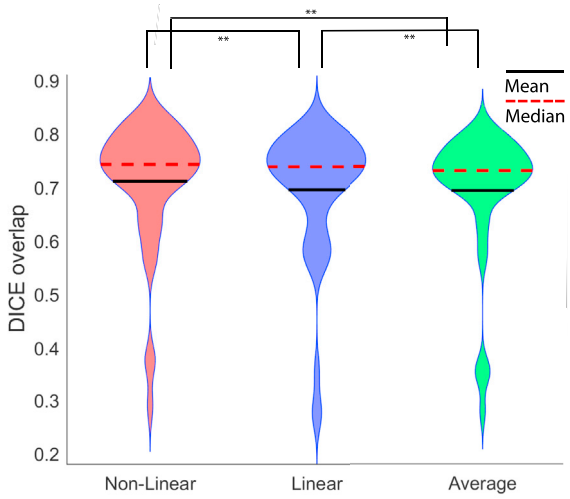

Fig. 5. Violin plot for the three realignment methods (Non-Linear, Linear, Average) of Dice overlap scores between individual subject TSEs and its respective method-template, collapsed over the three groups (MND, HC, YHP). ** = ps < 0.001, Wilcoxon rank sum tests).

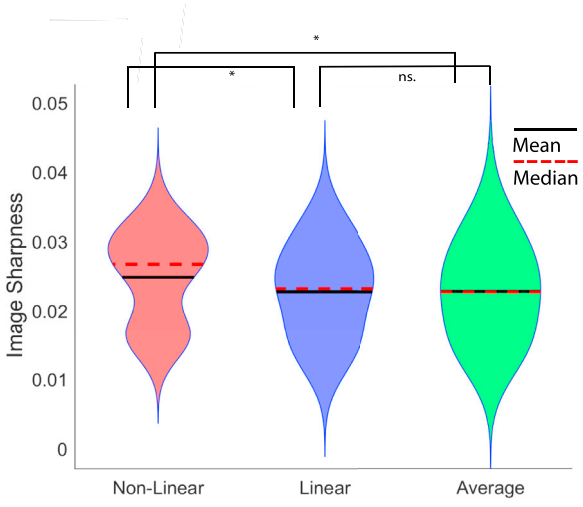

Fig. 6. Violin plot for the three realignment methods of sharpness scores of individual subject TSEs, collapsed over the three groups. * = ps < 0.05, Wilcoxon rank sum test.

participants contained either anatomical variability (e.g., incomplete hippocampal inversion or cysts) that caused mislabelling in all conditions, which suggests *Linear* or *Average* performed better by chance in these participants. In the four remaining under-performing *Non-Linear* participants, we found little or no movement (e.g., in YHPs). Most of the participants in the YHP group had no adverse motion artefacts before processing. Figs. 1 and 2 show examples of labelled method templates and corresponding labelled subject TSEs.

Sharpness (Fig. 6) was significantly higher for *Non-Linear* compared to *Linear* and *Averaging* ($p < 0.001$, Friedman test). Significant differences were found between the three registration techniques, with *Non-Linear* performing better than *Linear,* and *Averaging*, independently (* = $ps < 0.05$). There was no significant difference between *Average* and *Linear* realignment ($p = 0.982$, Wilcoxon rank sum test).

When comparing within groups, it was found that MND and HC groups individually did not show significant differences in sharpness scores between any of the three methods. The YHP group alone showed the same (significant) pattern of sharpness results as when collapsing across groups (i.e., *Non-Linear* > *Linear* > *Average*). For the 3T group, we found that *Non-Linear* again produced higher sharpness than *Linear* and *Average* ($ps < 0.05$, Wilcoxon rank sum test), while there was no significant difference between *Linear* and *Average* ($p = 0.63$).

4. Discussion

We investigated the influence of different realignment strategies for repeated high-resolution images of the hippocampus and found that *Non-Linear* realignment out-performs *Linear* and *Average* registrations for improving segmentation consistency for the hippocampus on a group level. This improves the robustness of brain image segmentation and mitigates the effects of motion artefacts in high-resolution anatomical MRI. We found that non-linear registration assists in registration consistency between individuals and a representative group average compared to other realignment techniques. We hypothesised that greater registration consistency would occur between participants if the registration procedure is more effective and the between-subject registrations would consequently converge more readily due to increased boundary delineation and higher sharpness. The results largely reflect that *Non-Linear* realignment is a suitable technique on the whole for increasing image sharpness, which leads to better segmentation consistency. We suggest that participant motion during high-resolution hippocampus acquisitions can be mitigated by non-linear realignment.

Studying individual participants, *Linear* and *Average* registration techniques occasionally out-perform *Non-Linear registration* in participants with little or no movement, but only in YHPs. From examining these cases, it is apparent that non-linear realignment may have only modest effects when there is no movement, image artefacts, or CNR

issues. It was found that in most of the cases where *Non-Linear* was an under-performer in segmentation consistency, *Average* and *Linear* techniques also failed in terms of Dice overlap (Dice < 0.7). These cases are visible in Fig. 5 and fall in the lower range of Dice scores. In the remaining cases where only modest benefits of non-linear realignment are observed, we propose that compounding interpolation errors during the realignment process and regularisation errors are responsible for the poor performance in these participants. Multiple non-linear registrations to an intermediate space require multiple interpolation steps in the *Non-Linear* method. It is possible that participants with minimal movement were adversely affected by (the largely unnecessary) non-linear registrations.

In most participants where *Linear* or *Average* out-performed *Non-Linear*, we note the anatomical variability of five participants that influenced the segmentation of the participants. We found three incomplete hippocampal inversions, and two participants with large hippocampal cysts, which negatively impacted the segmentation consistency in all three methods. These aberrations in anatomy were not reflected in the overall templates and therefore Dice overlaps were negatively affected by aberrations in the labelling. Additionally, due to the small sizes of the subfields, Dice coefficients are impacted by many other factors, including the segmentation strategy and the robustness of the registrations. The comparison between the individual and the template segmentation is also highly variable, given the Dice score reflects the segmentation consistency, the registration quality, and the size of the subfield. We additionally resample the segmentations into the space of the template, resulting in interpolation artefacts, and therefore affecting the Dice coefficients for some participants.

In participants with severe motion artefacts, non-linear realignment showed its greatest utility, with much of the information lost to partial volume effects or motion artefacts being reclaimed through the technique. Older aged and diseased populations may also show decreased contrast (Wisse et al., 2014). Our results show consistent and improved automatic segmentation can be achieved in older aged and diseased populations using non-linear registrations. MND and HCs, who had more motion artefacts, showed the most improvement for segmentation consistency with non-linear realignment, strengthening this case. However, we found no significant differences in terms of sharpness in the MND group between methods. This suggests difficulties resolving SNR and contrast decreases in older-aged and diseased groups, and future work is required to ameliorate contrast loss in these populations.

Non-linear registrations (especially ANTs SyN) incur a high computational cost. At higher resolution, these computational costs may be prohibitive. The symmetrical condition of the registration does not necessarily need to be fulfilled for the realignment to be successful. Therefore, other non-linear registration platforms including Greedy (Xie et al., 2018) or VolGenModel (Janke and Ullmann, 2015) should be explored in future works.

Measuring segmentation consistency is not a direct measurement of any realignment technique's effectiveness, and the overlap resulting from registrations may not provide a true representation of segmentation *accuracy*. Here, we report on segmentation consistency, though consistency may have been more reliably derived through manual segmentations. Indeed, to describe *accuracy*, a comparison between manual segmentations and automatic segmentations would be required. Our measure of image sharpness offers convergent validity to the measure of image segmentation, as sharpness was found to be highest in the *Non-Linear* condition. Image sharpness largely reflects the smoothness of anatomical boundaries, with higher sharpness denoting better delineation between anatomical structures. Hippocampus segmentation relies on the distinction between anatomical landmarks such as *Cornu Ammonis* 1 and Dentate Gyrus (as separated by the *stratum radiatum lacunosum moleculare*). These features can clearly be seen, e.g., in Figs. 2 and 4, and we are therefore confident our measure of segmentation consistency is meaningful.

We conclude that hippocampus image segmentation can benefit from non-linear registrations when participant motion between scans is an issue. It is proposed that due to the high computational cost, non-linear realignment be used judiciously in participants with high inter-scan motion, as determined by metrics such as sharpness or qualitative assessment.

Acknowledgements

The authors acknowledge the facilities and scientific and technical assistance of the National Imaging Facility, a National Collaborative Research Infrastructure Strategy (NCRIS) capability, at the Centre for Advanced Imaging, The University of Queensland. MB acknowledges funding from Australian Research Council Future Fellowship grant FT140100865. This research was undertaken with the assistance of resources and services from the Queensland Cyber Infrastructure Foundation (QCIF). The authors gratefully acknowledge Aiman Al Najjer and Saskia Bollmann for acquiring data. The data acquisition was partly funded by a study from the Cooperative Research Centre for Mental Health. Data for the 24 adolescent participants was provided by the Queensland Adolescent Twin Brain project, funded by the NHMRC (APP1078756).

References

Avants, 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 (1), 26 41. https://doi.org/10.1016/j.media.2007.06.004.

Avants, Brian, Tustison, N., Song, G., 2010. Advanced Normalization Tools (ANTS), 35.

Backhausen, L.L., Herting, M.M., Buse, J., Roessner, V., Smolka, M.N., Vetter, N.C., 2016. Quality control of structural MRI images applied using Free Surfer—a hands-on workflow to rate motion artifacts. Front. Neurosci. 10. https://doi.org/10.3389/fnins.2016.00558.

Balachandar, R., John, J.P., Saini, J., Kumar, K.J., Joshi, H., Sadanand, S., et al., 2015. A study of structural and functional connectivity in early Alzheimer's disease using rest fMRI and diffusion tensor imaging: structural and functional connectivity in early Alzheimer's disease. Int. J. Geriatr. Psychiatry 30 (5), 497–504. https://doi.org/10.1002/gps.4168.

Boccardi, M., Ganzola, R., Bocchetta, M., Pievani, M., Redolfi, A., Bartzokis, G., et al., 2011. Survey of protocols for the manual segmentation of the Hippocampus: preparatory steps towards a joint EADC-ADNI harmonized protocol. J. Alzheimer's Dis.: JAD 26 (0 3). https://doi.org/10.3233/JAD-2011-0004.

Bollmann, S., Bollmann, S., Puckett, A., Janke, A.L., Barth, M., 2017. Non-linear realignment using minimum deformation averaging for single-subject fMRI at ultra-high field. Proc. Intl. Soc. Mag. Reson. Med. 25.

Boutet, C., Chupin, M., Lehéricy, S., Marrakchi-Kacem, L., Epelbaum, S., Poupon, C., et al., 2014. Detection of volume loss in hippocampal layers in Alzheimer's disease using 7 T MRI: a feasibility study. Neuroimage: Clin. 5, 341–348. https://doi.org/10.1016/j.nicl.2014.07.011.

Duvernoy, H.M., Cattin, F., Risold, P.-Y., Vannson, J.L., Gaudron, M., 2013. The Human hippocampus: Functional Anatomy, Vascularization and Serial Sections with MRI, fourth ed. Springer, Heidelberg ; New York.

Fonov, V., Collins, D.L., 2018. Comparison of different methods for average anatomical templates creation: do we really gain anything from a diffeomorphic framework? https://doi.org/10.1101/277087.

Giuliano, A., Donatelli, G., Cosottini, M., Tosetti, M., Retico, A., Fantacci, M.E., 2017. Hippocampal subfields at ultra high field MRI: an overview of segmentation and measurement methods. Hippocampus 27 (5), 481–494. https://doi.org/10.1002/hipo.22717.

Henry, T.R., Chupin, M., Lehéricy, S., Strupp, J.P., Sikora, M.A., Sha, Z.Y., et al., 2011. Hippocampal sclerosis in temporal lobe epilepsy: findings at 7 T. Radiology 261 (1), 199–209. https://doi.org/10.1148/radiol.11101651.

Hsu, Y.-Y., Schuff, N., Du, A.-T., Mark, K., Zhu, X., Hardin, D., Weiner, M.W., 2002. Comparison of automated and manual MRI volumetry of hippocampus in normal aging and dementia. J. Magn. Reson. Imaging 16 (3), 305–310. https://doi.org/10.1002/jmri.10163.

Iglesias, J.E., Augustinack, J.C., Nguyen, K., Player, C.M., Player, A., Wright, M., et al., 2015. A computational atlas of the hippocampal formation using ex vivo, ultra-high resolution MRI: application to adaptive segmentation of in vivo MRI. Neuroimage 115, 117–137. https://doi.org/10.1016/j.neuroimage.2015.04.042.

Iglesias, J.E., Liu, C.-Y., Thompson, P.M., Tu, Z., 2011. Robust brain extraction across datasets and comparison with publicly available methods. IEEE Trans. Med. Imaging 30 (9), 1617–1634. https://doi.org/10.1109/TMI.2011.2138152.

Jacobsen, N., Munk, J.B., Plocharski, M., Ostergaard, L.R., Marstaller, L., Reutens, D., et al., 2017. Building a high-resolution in vivo minimum deformation average model of the human hippocampus. BioRxiv, 160176. https://doi.org/10.1101/160176.

Janke, A.L., Ullmann, J.F.P., 2015. Robust methods to create ex vivo minimum deformation atlases for brain mapping. Methods 73, 18–26. https://doi.org/10.1016/j.ymeth.2015.01.005.

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17 (2), 825–841. https://doi.org/10.1006/nimg.2002.1132.

Jenkinson, M., Beckmann, C.F., Behrens, T.E.J., Woolrich, M.W., Smith, S.M., 2012. FSL. NeuroImage 62 (2), 782–790. https://doi.org/10.1016/j.neuroimage.2011.09.015.

Jones, K., Marietta, J., 2012. Quality Analysis of raw MRI scans using BRAINSImageEval. Retrieved March 6, 2019, from. https://slideplayer.com/slide/3280070/.

Kerchner, G.A., Deutsch, G.K., Zeineh, M., Dougherty, R.F., Saranathan, M., Rutt, B.K., 2012. Hippocampal CA1 apical neuropil atrophy and memory performance in Alzheimer's disease. Neuroimage 63 (1), 194–202. https://doi.org/10.1016/j.neuroimage.2012.06.048.

Kochunov, P., Lancaster, J.L., Glahn, D.C., Purdy, D., Laird, A.R., Gao, F., Fox, P., 2006. Retrospective motion correction protocol for high-resolution anatomical MRI. Hum. Brain Mapp. 27 (12), 957–962. https://doi.org/10.1002/hbm.20235.

La Joie, R., Perrotin, A., de La Sayette, V., Egret, S., Doeuvre, L., Belliard, S., et al., 2013. Hippocampal subfield volumetry in mild cognitive impairment, Alzheimer's disease and semantic dementia. Neuroimage: Clinical 3, 155–162. https://doi.org/10.1016/j.nicl.2013.08.007.

Maclaren, J., Herbst, M., Speck, O., Zaitsev, M., 2013. Prospective motion correction in brain imaging: a review. Magn. Reson. Med. 69 (3), 621–636. https://doi.org/10.1002/mrm.24314.

Marques, J.P., Kober, T., Krueger, G., van der Zwaag, W., Van de Moortele, P.-F., Gruetter, R., 2010. MP2RAGE, a self bias-field corrected sequence for improved segmentation and T1-mapping at high field. Neuroimage 49 (2), 1271–1281. https://doi.org/10.1016/j.neuroimage.2009.10.002.

Marques, J.P., Norris, D.G., 2018. How to choose the right MR sequence for your research question at 7T and above? Neuroimage 168, 119–140. https://doi.org/10.1016/j.neuroimage.2017.04.044.

Marrakchi-Kacem, L., Vignaud, A., Sein, J., Germain, J., Henry, T.R., Poupon, C., et al., 2016. Robust imaging of hippocampal inner structure at 7T: in vivo acquisition protocol and methodological choices. Magn. Reson. Mater. Phys. Biol. Med. 29 (3), 475–489. https://doi.org/10.1007/s10334-016-0552-5.

Maruszak, A., Thuret, S., 2014. Why looking at the whole hippocampus is not enough—a critical role for anteroposterior axis, subfield and activation analyses to enhance predictive value of hippocampal changes for Alzheimer's disease diagnosis. Front. Cell. Neurosci. 8. https://doi.org/10.3389/fncel.2014.00095.

Mulder, E.R., de Jong, R.A., Knol, D.L., van Schijndel, R.A., Cover, K.S., Visser, P.J., et al., 2014. Hippocampal volume change measurement: quantitative assessment of the reproducibility of expert manual outlining and the automated methods FreeSurfer and FIRST. Neuroimage 92, 169–181. https://doi.org/10.1016/j.neuroimage.2014.01.058.

O'Brien, K.R., Kober, T., Hagmann, P., Maeder, P., Marques, J., Lazeyras, F., et al., 2014. Robust T1-weighted structural brain imaging and morphometry at 7T using MP2RAGE. PLoS One 9 (6), e99676. https://doi.org/10.1371/journal.pone.0099676.

Pluta, J., Yushkevich, P., Das, S., Wolk, D., 2012. In vivo analysis of hippocampal subfield atrophy in mild cognitive impairment via semi-automatic segmentation of T2-weighted MRI. J. Alzheimer's Dis. 31 (1), 85–99. https://doi.org/10.3233/JAD-2012-111931.

Reuter, M., Rosas, H.D., Fischl, B., 2010. Highly accurate inverse consistent registration: a robust approach. Neuroimage 53 (4), 1181–1196. https://doi.org/10.1016/j.neuroimage.2010.07.020.

Romero, J.E., Coupé, P., Manjón, J.V., 2017. HIPS: a new hippocampus subfield segmentation method. Neuroimage 163 (Suppl. C), 286–295. https://doi.org/10.1016/j.neuroimage.2017.09.049.

Thomas, B.P., Welch, E.B., Niederhauser, B.D., Whetsell, W.O., Anderson, A.W., Gore, J.C., et al., 2008. High-resolution 7T MRI of the human hippocampus in vivo. J. Magn. Reson. Imaging: JMRI 28 (5), 1266–1272. https://doi.org/10.1002/jmri.21576.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Yuanjie, Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging 29 (6), 1310–1320. https://doi.org/10.1109/TMI.2010.2046908.

Vincent, R.D., Neelin, P., Khalili-Mahani, N., Janke, A.L., Fonov, V.S., Robbins, S.M., et al., 2016. MINC 2.0: a flexible format for multi-modal images. Front. Neuroinf. 10. https://doi.org/10.3389/fninf.2016.00035.

Winterburn, J.L., Pruessner, J.C., Chavez, S., Schira, M.M., Lobaugh, N.J., Voineskos, A.N., Chakravarty, M.M., 2013. A novel in vivo atlas of human hippocampal subfields using high-resolution 3 T magnetic resonance imaging. Neuroimage 74, 254–265. https://doi.org/10.1016/j.neuroimage.2013.02.003.

Wisse, L.E.M., Biessels, G.J., Geerlings, M.I., 2014. A critical appraisal of the hippocampal subfield segmentation package in FreeSurfer. Front. Aging Neurosci. 6. https://doi.org/10.3389/fnagi.2014.00261.

Wisse, L., Kuijf, H.J., Honingh, A.M., Wang, H., Pluta, J.B., Das, S.R., et al., 2016. Automated hippocampal subfield segmentation at 7 tesla MRI. AJNR. Am. J. Neuroradiol. 37 (6), 1050–1057. https://doi.org/10.3174/ajnr.A4659.

Xie, L., Shinohara, R.T., Ittyerah, R., Kuijf, H.J., Pluta, J.B., Blom, K., et al., 2018. Automated multi-atlas segmentation of hippocampal and extrahippocampal subregions in Alzheimer's disease at 3T and 7T: what atlas composition works best? J. Alzheimer's Dis. 63 (1), 217–225. https://doi.org/10.3233/JAD-170932.

Yushkevich, P.A., Pluta, J.B., Wang, H., Xie, L., Ding, S.-L., Gertje, E.C., et al., 2015. Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment. Hum. Brain Mapp. 36 (1), 258–287. https://doi.org/10.1002/hbm.22627.

Zou, K.H., Warfield, S.K., Bharatha, A., Tempany, C.M.C., Kaus, M.R., Haker, S.J., et al., 2004. Statistical validation of image segmentation quality based on a spatial overlap index. Acad. Radiol. 11 (2), 178–189. https://doi.org/10.1016/S1076-6332(03)00671-8.