Technology Update

Generating a Minimal Set of Templates for the Hippocampal Region in MR Neuroimages

R. Cataldo, MSc Phy, A. Agrusti, MSc Phy, G. De Nunzio, PhD, A. Carlà, MSc Phy, I. De Mitri, PhD, M. Favetta, PhD, M. Quarta, MSc Phy, L. Monno, MSc Phy, L. Rei, MSc Phy, E. Fiorina, MSc Phy, and the Alzheimer's Disease Neuroimaging Initiative*

From the Department of Materials Science, University of Salento, Lecce (RC, AA, GDN); Istituto Nazionale di Fisica Nucleare, Lecce (RC, AA, GDN, AC, IDM, MF, MQ); Department of Physics, University of Salento, Lecce (AC, IDM, MF); Department of Mathematics, University of Salento, Lecce (MQ); Department of Physics, University of Bari (LM); Istituto Nazionale di Fisica Nucleare, Bari (LM); Department of Physics, University of Genova, Genova (LR); Istituto Nazionale di Fisica Nucleare, Genova (LR); Department of Physics, University of Torino, Torino (EF); and Istituto Nazionale di Fisica Nucleare, Torino (EF).

ABSTRACT

OBJECTIVES

We detail a procedure for generating a set of templates for the hippocampal region in magnetic resonance (MR) images, representative of the clinical conditions of the population under investigation.

METHODS

The first step is robust standardization of the intensity scale of brain MR images, belonging to patients with different degrees of neuropathology (Alzheimer's disease). So similar tissues have similar intensities, even across images coming from different sources. After the automatic extraction of the hippocampal region from a large dataset of images, we address template generation, choosing by clusterization methods a small number of the extracted regions.

RESULTS

We assess that template generation is largely independent on the clusterization method and on the number and the clinical condition of the patients. The templates are chosen as the most representative images in a population. The estimation of the "minimum" number of templates for the hippocampal region can be proposed, using a metric based on the geometrical position of the extracted regions.

CONCLUSIONS

This study describes a simple and easily reproducible procedure to generate templates for the hippocampal region. It can be generalized and applied to other brain regions, which may be relevant for neuroimaging studies.

Keywords: Templates, MRI brain images, hippocampus, brain region, Alzheimer's Disease Neuroimaging Initiative.

Acceptance: Received May 8, 2011, and in revised form November 5, 2011. Accepted for publication November 20, 2011.

Correspondence: Address correspondence to Rosella Cataldo, MSc Phy, Department of Materials Science, University of Salento, Lecce (RC, AA, GDN); Istituto Nazionale di Fisica Nucleare, Lecce (RC, AA, GDN, AC, IDM, MF, MQ), E-mail: rosella.cataldo@unisalento.it.

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

J Neuroimaging 2012;XX:1–10.
DOI: 10.1111/j.1552-6569.2012.00713.x

Introduction

Magnetic resonance imaging (MRI) has been used in numerous in vivo anatomical studies of the brain,[1] especially for the hippocampus, and plays an important role in the diagnosis of temporal lobe epilepsy, or degenerative diseases such as Alzheimer's dementia, and in the evaluation of their course.[2] In MR images, a significant atrophy or a degeneration not only in the striatum, but also of other structures such as white matter (WM), cerebral cortex, amygdala, and hippocampus, contribute to identifying Huntington's chorea or cognitive impairment.[3,4]

Anatomical reference images (templates) are then becoming of vital importance for result comparison, and to allow better identification of structures. Such templates are primarily intended to serve as anatomical references for spatial normalization usually required before studying human anatomical or functional variability. Atlases derived from those templates, built with different modalities and most often characterized by specific structure labeling,[5,6] have been used with success in various computer-aided decision systems. They are usually built from a single acquisition of different control subjects reflecting the population targeted by the clinical study.[7,8]

Atlas-generation has to account for inter-subject variability of anatomy and function, in order to offer a powerful framework which facilitates comparison over time, between subjects, between groups of subjects and across sites.[9] This the significant goal. With this aim, various probabilistic and statistical approaches have been developed to establish anatomical reference images (templates), representative of the population under investigation.[10,11]

Studying and quantifying local anatomical differences or changes in a population, in a sense of characterizing anatomical differences between subjects and templates, is a very challenging and difficult task. Especially the problem of identifying differences in relative positions of brain structures and detecting local differences requires complex models. Linear or nonlinear forms of spatial normalization are used to register images from larger cohorts into a common stereotactic space, enabling region by region comparisons.[1,7-12,14-17]

In this work, we describe a rather simple procedure for generating a set of templates for the hippocampal region. It represents a refinement and a generalization of the Calvini et al[12] research. Against that background, many aspects are differently treated, especially regarding image standardization and the estimation of the "minimum" number of templates for the hippocampal region, and the effect of these different strategies are evaluated.

The complete procedure foresees different steps. First, we build up a robust method for standardizing the intensity scale of brain MR images. This way similar tissues have similar intensities, even across images coming from different sources. Then we automatically extract the hippocampal region from a large dataset of MR images, with great accuracy. Subsequently, template generation is addressed, by choosing a small number of the extracted regions, using clusterization methods.

Because morphological variability is captured by rigid transformations, a number of templates are necessary for group analysis. Thus, the number of images belonging to each considered template set, obtained in the clusterization step, are the means to account for variability.

Finally, we propose an estimation of the "minimum" number of templates for the chosen brain area, using a metric based on the geometrical position of the extracted hippocampal regions.

We also attempt a preliminary generalization test of the procedure, applying it to other brain areas, which may be relevant for future neuroimaging studies.

The procedure was developed within the "Medical Application on a Grid Infrastructure Connection" (MAGIC-5) group, an Italian collaboration related to Istituto Nazionale di Fisica Nucleare (INFN), involving many academic and clinical institutions in the field of computer-aided detection (CAD) software systems for the analysis of medical images,[11,13] such as MR or positron emission tomography (PET) images, as a support for the early diagnosis of neurological pathologies, especially the Alzheimer's disease (AD).[12]

Materials and Methods

Materials

Various datasets of brain T1-weighted MR images downloaded from the Alzheimer's Disease Neuroimaging Initiative (ADNI) website (http://www.loni.ucla.edu/ADNI/) were used to develop the procedure detailed in the next paragraphs.

Each image in the ADNI dataset has undergone specific image preprocessing correction steps.

These corrections include gradwarp, B1, and N3 corrections. Gradwarp is a system-specific correction of image geometry distortion due to gradient nonlinearity. B1 nonuniformity correction procedure employs the B1 calibration scans to correct the image intensity nonuniformity, that results when radiofrequency transmission is performed with a more uniform body coil, while reception is performed with a less uniform head coil. N3 is a histogram peak sharpening algorithm applied to all images, after gradwarp and after B1 corrections, for systems on which these two correction steps are performed. N3 processing reduces residual intensity nonuniformity.

Each dataset is nonhomogeneous in terms of age, cardinality, and pathology of the subjects, with clinical conditions ranging from good health state (Normal) to probable dementia of AD type as well as with MCI. The minimental state examination (MMSE) test was used to estimate the severity of the cognitive impairment.

In particular, for a study including 190 patients, we have 65 Normal (27 women and 38 men) with average MMSE score 29.1 $\pm$ 1.1 and average age 81.9 $\pm$ 6.4; 63 MCI (30 women and 33 men) with average MMSE score 26.2 $\pm$ 2.3 and average age 70.3 $\pm$ 8.5; and 62 AD (28 women and 34 men) with average MMSE score 24.5 $\pm$ 2.7 and average age 75.7 $\pm$ 5.1. Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California – San Francisco. ADNI is the result of efforts of many coinvestigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the United States and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55 to 90, to participate in the research, approximately 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years and 200 people with early AD to be followed for 2 years. For up-to-date information, see www.adni-info.org.

Methods

Before detailing the procedure, some general considerations have to be addressed.

One of the most important application of an atlas is to identify the abnormality within patient population, in order to account for the maximum intersubject variability of anatomy and function, which facilitates comparison of anatomy and function over time, between subjects, between groups of subjects, and across sites.[9]

A critical step is to transform all individual images toward a common space consistently. In the literature pairwise registrations toward a chosen template, which can be either a subject in the population, or an averaged image obtained from multiple aligned subjects (such as ICBM152 template), are widely used for atlas construction.

However, a limitation associated with this approach is that the resulting atlas may be biased toward the predefined template.[15]

To overcome this limitation, Joshi et al,[16] for example, developed an unbiased groupwise registration approach,[16] in which the template (mean) image was gradually formed during the registration procedure. But this approach ignores the possible inhomogeneous distribution of data within the population, and uses only a single image to represent all the subjects.

On the other end, Wolz et al[14] consider the propagation of a relatively small number of atlases to a large and diverse set of MR brain images, exhibiting a significant amount of anatomical variability. Then a propagation framework is identified and labeled atlases can be propagated in a stepwise fashion, starting with the initial atlases, on the whole population. They assess that deformations between dissimilar images are broken down to several small deformations between comparatively similar images and registration errors are reduced.

We propose a different method, selecting a set of images to be used as suitably chosen templates[17] for a region of interest, here the hippocampal formation, preserving the ample morphological variability contained in the population of MR images.

Whereas in most of the aforementioned examples, the morphological variability is captured by nonrigid transformations, we make use of rigid transformations, accounting for the variability by the number of images belonging to each template set.

Our template generation starts with the extraction of the searched region from the images of a MRI dataset, by coregistering each image with a hippocampal reference image. This means that the extraction will be able to identify the searched region unambiguously if a reference image, hereafter called box, of well-defined initial coordinates, dimension, and shape is used. Some cautions have to be employed in choosing these parameters.

As regards the dimension of the reference image, we can consider that small boxes contain too little information. For example, if a brain region is represented by a box of $10 \times 10 \times 10$ pixels, coregistration and extraction can be performed, but it is questionable if the extracted region has the same anatomic meaning as the reference image. Therefore, the optimal extent for the box size is to be defined to properly distinguish it among the different structures. Neuroanatomical considerations are fundamental, because this extent has to take into account the intersubject variability of the anatomy that can be significant especially in template generation.

After various tests with boxes of different and increasing size on images belonging to patients with different degrees of neuropathology, we assessed that a box extension of $30 \times 70 \times 30$ mm^3 meets the requirements to identify a region containing the hippocampal formation unambiguously, taking into account intersubject variability

Hereafter, we detail the complete template-extraction procedure, composed of three main steps:

(1) Histogram standardization and spatial normalization to stereotactic space (ICBM152).
(2) "Exhaustive" extraction of the hippocampal regions from all the images.
(3) Template-set selection.

The open-source Insight Segmentation and Registration Toolkit (ITK, available at http://www.itk.org), Statistical Parametric Mapping (SPM8, available at http://www.fil.ion.ucl.ac.uk/spm), and MATLAB, a high-level technical-computing language (http://www.mathworks.com/products/matlab), are currently required as software platforms.

Methods: Histogram Standardization and Spatial Normalization

It is well known that the MRI intensity scale has no absolute, physical meaning: pixel gray levels in fact depend on the pulse sequence and other variable scanner and postprocessing parameters. It is evident that the lack of a standard image-intensity scale may cause difficulties in imaging analyses. The ability of a tissue classification method to automatically adapt to a MR image is especially important when the data is collected in multiple sites, or with several different MRI scanners.

Even if our sample dataset contains ADNI images that have undergone specific preprocessing image correction steps (eg, bias-field correction), the existence of images with quite different gray level histograms is immediately apparent. The extraction method described in this paper might be strongly affected by intervolume (of the same or different individuals) lack of intensity uniformity. Various approaches address the problem in the literature, for example, Nyúl and Udupa,[18] Nyúl et al,[19] Ge et al,[20] Hellier,[21] Weisenfeld and Warfield,[22] Schmidt,[23] Jäger and Hornegger,[24] and Leung et al.[25] A detailed review of some recent standardization algorithms is in Bergeest and Jäger.[26]

Here we build up a robust method for standardizing the intensity scale across multiple scans, in such a way that similar intensities have similar tissue meaning, even across images coming from different sources. In Calvini et al[12] this step was limitedly present, and only a histogram equalization was applied to the boxes after extraction. On the contrary, in the approach described in this paper, histogram standardization is considered essential to achieve general applicability and extraction quality.

Our standardization method is simple and easy to implement, quite light from the computational point of view, and produces reasonable results. It is mainly inspired by Nyúl and Udupa,[18] Nyúl et al,[19] and Ge et al[20] works, where gray level normalization is obtained by selecting for each image of a training set some histogram landmarks, averaging them to obtain a list of reference mean landmarks (to be used as a standard scale). Each training-set image is then standardized, by projecting its landmarks onto the standard ones, while the gray levels

between the landmarks are linearly interpolated. Thus a continuous, piecewise linear intensity mapping to a standard scale is achieved. When a new image is acquired, the transformation to the standard scale is used to standardize it.

In the original paper by Nyúl and Udupa[18] the landmarks were mode-based, that is, the local maxima of the histogram were used. In their subsequent work they chose a set of population percentiles instead, to make the method more robust and avoid incorrect standard scales. In fact, as pointed out by the authors, it might happen that a particular mode corresponded, in two images *A* and *B*, to different matters (eg, WM in image *A*, gray matter [GM] in image *B*). In this case, the mode should not be used as a landmark, because it would lead to tissue mixing, in fact different tissues would be projected to the same "standard" levels. The consequence of training with such landmarks would be to obtain a meaningless standard scale.

After coding and testing this standardization approach, we remarked that even when percentiles were used, the risk of tissue mixing was high. Figure 1 shows a typical case where tissue mixing, produced by inaccurate standardization based on percentiles, is evident, and would be inevitable. The histograms of two images acquired in different hospitals, and from different patients, are shown. The original histograms are the thick continuous lines on top of each plot. On these histograms a particular percentile (80%) was calculated as an example, and is marked as P on both images.

Following the approach by Nyúl et al,[19] this percentile could be chosen as a hypothetically stable landmark for standardization, without incurring in the tissue mixing involved in choosing maximum points as landmarks. We now show that this landmark potentially generates tissue mixing too. Consider the three gray level histograms (continuous lines) under the main original one (both images): from left to right, they correspond to cerebrospinal fluid (CSF), GM, and WM histograms, and were calculated by segmentation of the images into the three main brain tissues (the sum of the three histograms is shown with dotted lines).

It is evident that (neglecting background contribution) the aforementioned landmark corresponds to almost pure WM in the right image, and to a mix of about 50% WM and 50% GM in the left one. If many MRI scans were taken for the training database similar to the right one, the averaged 80% landmark would correspond to pure WM, and an image like the left one would undergo a "conversion" of part of its GM to WM. As a consequence, the choice of a percentile-based landmark does not ensure stability.

In our implementation of the algorithm, this drawback is limited by separately applying the standardization procedure, after segmentation by an atlas, to the three main cerebral tissue classes instead of the whole brain. Moreover, deciles are chosen as the histogram landmarks, so as to have a smoother map function. Some details follow, depicting the overall procedure and highlighting the novelty elements, whereas for a thorough examination of the benefits of this technique and a comparison with other approaches available in the literature, the reader is referred to a future specific paper.

A group of training images is chosen by taking a subset of the available database, checking that the corresponding histograms are as representative of population variability as possible.[18,19] The images are then coregistered and segmented into WM, GM, and CSF. Of course, a segmentation procedure based only on the pixel gray levels would not be satisfying, because of lack of standardization, for this reason we chose an atlas-driven procedure (Standard Unified Segmentation) implemented in SPM8.[27] This segmentation is based on a modified Gaussian Mixture Model, which has been extended to include spatial maps of prior belonging probability. It uses Bayes' rule to assign the probability for each voxel to belong to each tissue class, based on combining the likelihood for belonging to the tissue class and the prior probability.

In all of the reference papers, standardization was applied to the original image. In Leung et al,[25] mean GM, WM, and CSF gray values for the images to be standardized were determined by *k*-means segmentation, and used by linear regression as landmarks for the calculation of the intensity transformation. The latter is then applied to any brain tissue. On the contrary, in our approach three standardizing transformations are calculated and separately applied to the GM, WM, and CSF images, using a group of functions written in the MATLAB environment.

For this purpose, an analysis of probability maps for GM, WM, and CSF was performed. We observed in the original images the presence of partial volume voxels belonging to the edge of the brain neighborhood and that can be shared by two or three matters. To assign each voxel to one class of tissue and then get the masks on the three tissues, a threshold was applied in terms of probability maps.

For the voxels located on the edge of the brain a value of 1 was assigned, if their intensity value was greater than .5, otherwise a value of 0 was assigned. In the latter case, the voxel is presumably vacuum, or may represent bone or fat (removed during segmentation).

For mixed voxels, the voxel is assigned to the mask with the highest probability of membership. For example, a voxel with a probability value of belonging to GM, WM, and CSF of .7, .2, and .1, respectively, will be assigned to the gray matter mask.

In such a manner, a correspondence between the voxels of the original images and the corresponding three different types of tissues (GM, WM, and CSF) is obtained.

A large set of landmarks, composed by deciles, is chosen in the histograms of each of the three tissues, to get quite smooth descriptions of the three standardization functions. Once GM, WM, and CSF images are individually standardized, we combine them again to get the complete standardized images, whose voxel size is 1 mm × 1 mm × 1 mm. This way we can safely assume that the probability of mixing pixels of different tissues during intensity standardization is minimized. Some histograms before and after standardization are shown in Figures A and B. The standardized images, indexed with a cardinal number j, MR_j, $j = 1,..,n$, are then spatially normalized to stereotactic space (ICBM152) via a 12-parameter affine transformation[28] which coregisters the volumes so that all the hippocampi share similar position and orientation. The n MR standardized and coregistered images are now ready for the next step.

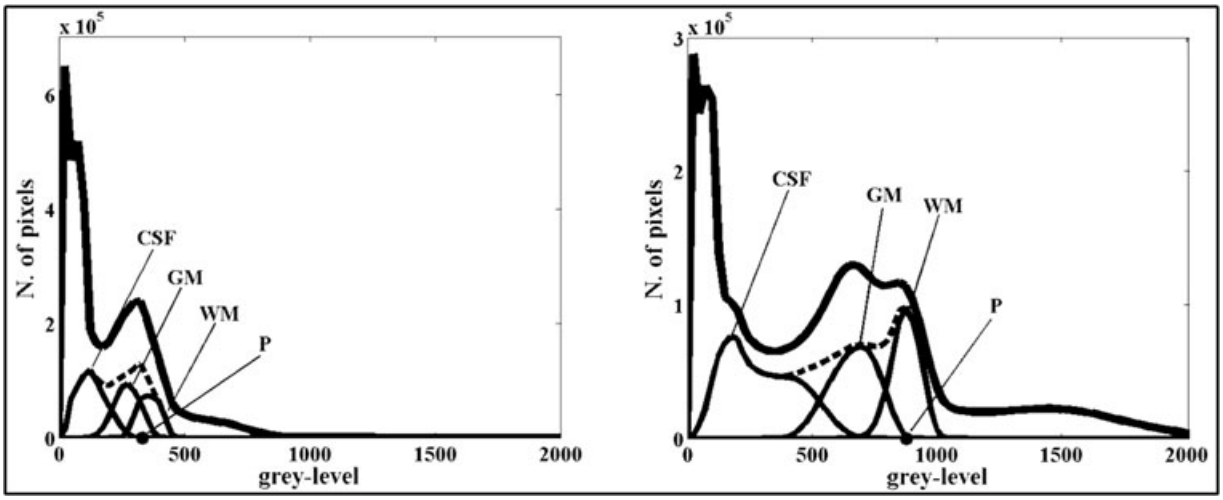

Fig 1. Tissue mixing produced by inaccurate standardization based on percentiles. The original histograms of two images acquired in different hospitals, from different patients, are the thick continuous lines on top of each plot. The dotted line represents the sum of the three histograms for CSF, GM, and WM, calculated by segmentation of the images into the three main brain tissues. On these histograms a particular percentile (80%) was calculated as an example, and is marked by a dot on the *x*-axis and labeled by P in both images.

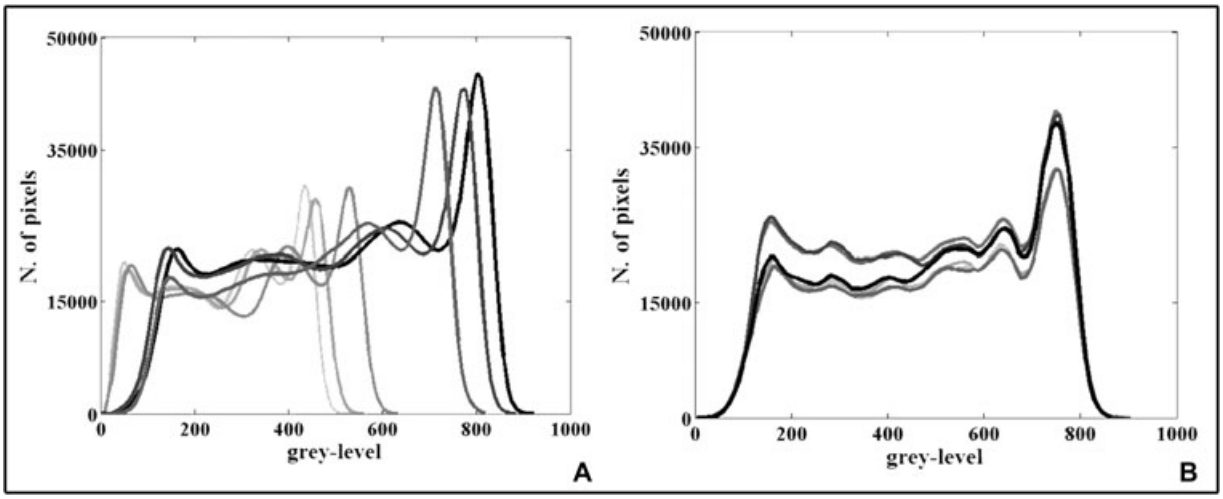

Fig 2. Histograms of some MR volumes (A) before and (B) after standardization. Only GM, WM, and CSF are considered, while other tissues (bone, fat, etc) and the dark background have been removed.

Methods: "Exhaustive" Extraction

This paragraph describes the extraction of the hippocampal regions from a MRI dataset, strictly following the description of such step given in Calvini et al,[12] with the aim to make this paper easier to read.

This procedure, henceforth called "exhaustive," is automatic, requires minimal manual intervention and can be employed both on the right and left hippocampal box (HB).

The ITK NormalizedCorrelationImageToImageMetric[29] module was used. It performs a coregistration, via a 6-parameter rigid transformation, that computes pixelwise cross-correlation and normalizes it by the square root of the autocorrelation of the "moving" (the *n* MR histogram-standardized and spatially normalized) images and the "fixed" image (a reference box of $30 \times 70 \times 30$ mm^3 in our case). The use of this module is limited to images obtained using the same imaging modality. This transformation returns the grabbed box and the registration parameters, that is, the Euler angles and the (x, y, z) coordinates of the lower leftmost anterior corner of the HB. The extraction is performed iteratively, so once extracted all the HBs from the whole dataset, for each iteration, we chose the next fixed image among the already extracted boxes, evaluating the correlation coefficient $C_{A,B}$.

$C_{A,B}$, for two assigned HBs, the fixed image with respect to which the extraction is performed, and the box extracted from the moving one, named A and B, respectively, each consisting of N pixels ($N = 63{,}000$ in our case), is

$$C_{A,B} = \frac{\sum_{i=1}^{N} (A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum_{i=1}^{N} (A_i - \bar{A})^2}\sqrt{\sum_{i=1}^{N} (B_i - \bar{B})^2}}, \qquad (1)$$

where A_i and B_i are the ith pixels of A and B, respectively, $\bar{A}$ and $\bar{B}$ are their average values.

The evaluation of similarity, that is, the success of the overlap between the moving and fixed images, is performed through the correlation coefficient. $C_{A,B}$ gives the best sensitivity because, as discussed earlier, the dimension and the initial coordinates of the fixed image are chosen in a manner that the search for the hippocampal formation requires the exploration of a small parameter space. Being all images aligned in the same stereotactic space, the level of spatial registration of similar anatomical structures is very high and $C_{A,B}$ is significant.

Furthermore, this sensitivity is significant also when introducing, for example, MR images with bad spatial normalization, obtaining low $C_{A,B}$. This means that the correlation coefficient value is able to capture the anatomical variations resulting from pathology across subjects, that is, between the fixed image and the boxes extracted from the whole moving image dataset, for each iteration.

In the first iteration, the fixed image (hereafter named HB_0) is an hippocampal formation manually segmented by a neurologist by accurately positioning the box boundaries, including also some neighboring regions (an example is available in the Supporting Information). The image from which HB_0 is extracted must also be preventively standardized and coregistered with the n images of the dataset.

After the coregistration of the n images with HB_0 and the calculation of C_{0j} (where j goes from 1 to n), a new HB is extracted from the image, for which C_{0j} was maximum. The latter box (called $HB_{k1,}$ if for example $j = k_1$ gives the maximum value) is inserted in the set of the extracted boxes $S = \{HB_{k1}\}$. This box takes the role of the new fixed image and its registration parameters are used to initialize the registration of the remaining $n - 1$ images in the following step.

Thus, in the second iteration, the population of the remaining $n - 1$ images is coregistered to HB_{k1}, giving C_{1j}. The box for which the maximum of $\{C_{0j}, C_{1j}\}$ is found, is chosen as fixed image for the next extraction of the HBs, suppose it is HB_{k2}: this box will be inserted in $S = \{HB_{k1}, HB_{k2}\}$.

The progressive extraction of all HBs stops once the whole set of n MR images has been processed and the n (right or left) HBs have been obtained. Set S is now full and contains all the extracted HBs, ex fixed images. Thus $n(n + 1)/2$ iterations are performed and except for the first extraction that requires a predefined hippocampal box HB_0 to start, the whole process is fully automatic and, as already pointed out, can be applied both to right and left hippocampi.

This procedure being based on an iterative process, its use is prohibitive when even a single image is added, because it does not permit a dynamic variation of the dataset, and the process has to be completely restarted. In principle, one could extract all the HBs in a dataset starting from a generic fixed image alone, without performing such a number of iterations, but a single fixed image (HB) that matches all anatomies in a dataset, even if constructed for example by group-averaging procedures, doesn't exist.[9] This iterative process assure instead that the extracted HBs really take into account the intersubject morphological variability, which in turn ensures that the extracted images can be useful in generating templates for that region of interest, allowing comparison over time and over datasets.

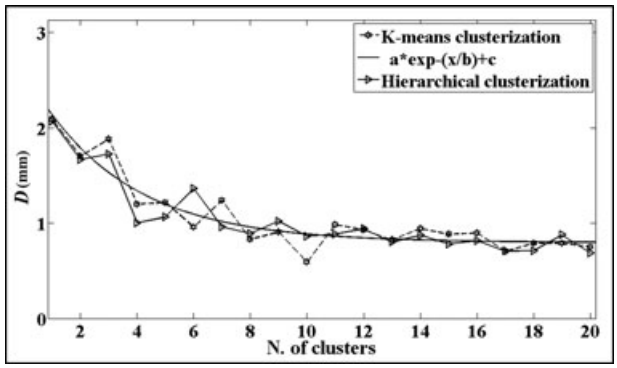

Fig 3. The pattern of the D parameter versus the number k of clusters, for two different clusterization methods (hierarchical and k-means), with templates extracted by images belonging to patients with different clinical conditions. The solid line represents the fitting curve for the mean values of the two datasets.

Of course the histogram standardization step is of crucial importance because, if the images were not based on a common standard gray level scale, the notion of metric given in Eq. (1) would lose sense and the extraction-procedure accuracy would be severely affected.

Methods: Template-Set Selection

The "exhaustive" procedure, explained earlier, shows how to extract a brain region (in particular, a hippocampal box) from a large dataset of volumetric MR images, iteratively building a set S of HBs from the image dataset.

Now we address template generation, by choosing a small number of the extracted regions, using clusterization methods.[30]

Our work does not fix a priori the number of hippocampal box templates (HBTs) to be selected for a target dataset. In our approach, each HB $\in S$ is treated as an object having a location in space, where the coordinates are given by the vector of gray levels of its N pixels ($30 \times 70 \times 30$ in our case) and whose distances from all other HBs are calculated. We use Euclidean and Correlation distances for the hierarchical and k-means clusterization methods, respectively, but we also assessed that other common distance measures such as Spearman or Mahalanobis did not significantly affect the results.

According to the above considerations, the procedure works applying clusterization methods to set S to obtain the desired S' set of templates, varying its cardinality k (starting from $k = 1$ upwards). The HBs closest to the centroids are chosen as the cluster representatives, and are inserted into a set S'_k as HBTs. This way we obtain a set of templates sets $\{S'_1, S'_2, S'_3, S'_4, S'_5, \ldots\}$, each element differing for cardinality and composition. For example, Figure 3 shows the templates selected after a k-means clusterization method for $k = 12$, from the dataset of 190 images belonging to subjects with different clinical conditions, and the fixed reference box HB_0 from which the "exhaustive" extraction started. As regards the clinical conditions, two boxes belong to AD, three boxes to MCI, and seven boxes to Normal subjects.

To assess the independence on the clusterization method, while choosing the cluster members on the same dataset of

HBTs, we considered the composition of each template set $\{S'_1, S'_2, S'_3, S'_4, S'_5, \ldots\}$ in terms of subjects with a particular clinical condition. In other words, we know how many AD, MCI, and Normal boxes are contained in each cluster, obtained by the hierarchical and *k*-means clusterization methods. Then an independent two-sample *t*-test is applied on two HBT datasets. Because the test statistic (t) does not fall into the rejection region, we can state that the null hypothesis is supported and thus that different clusterization methods give no appreciable difference between the means of the two samples. For example, $P(T \leq t)$ one-tail = .3 on the set of template sets $\{S'_1, S'_2, S'_3, S'_4, S'_5, \ldots\}$ extracted from 190 images belonging to subjects with different clinical conditions.

An interesting result is the independence of the template-set selection on the choice of the first fixed image (HB_0) from which the "exhaustive" procedure starts. Given the same dataset, we observed that the cluster members chosen after an "exhaustive" procedure started from an hippocampal formation (HB0) with severe atrophy, and those obtained starting from an HB0 with a minimal atrophy, are statistically compatible. We mean that the composition of each template set $\{S'_1, S'_2, S'_3, S'_4, S'_5, \ldots\}$, in terms of subjects with a particular clinical condition, is independent on the condition of the first hippocampal formation (HB_0).

At this point, we have a set of template sets, $\{S'_1, S'_2, S'_3, S'_4, S'_5, \ldots\}$, but we do not know yet which ones can be used as templates' set, able to represent the morphological variability of the whole population. A large *k* value, comparable with the number *n* of MRI, would reasonably give the complete description of the morphological variability, but this would be a trivial solution. On the contrary, a too small *k* value is expected to give poor accuracy.

Two questions arise, if a "natural" number *k* of templates exists, and, if not, how can we choose the "minimum" number *k* of templates. The problem of the existence of the "natural" number of clusters, representing the hippocampal boxes population is discussed in Esposito et al.[31] It is solved by analyzing the centroids distance distribution versus the number of clusters. As expected the resulting "natural" number of clusters is independent on the amount of boxes considered (if statistically consistent), and, for the hippocampal boxes, the centroids of the found clusters are the searched HBTs.

The problem of the selection of the "minimum" number *k* value for given set of images is very similar to a very known problem in unsupervised clustering techniques, that aim to identify a data-driven optimal number of clusters without a priori knowledge. For example, Wang and Zhang[32] conducted extensive comparisons of fuzzy cluster validity indices in conjunction with the Fuzzy *C*-Means clustering algorithm on a number of widely used datasets, with a simple analysis of the experimental results. They found that none of the abovementioned indices correctly recognizes optimal cluster numbers for all considered datasets, confirming the difficulties in cluster validation task.

For these reasons the next section will concern the assessment of quality of the template-based extraction procedure, and then the choice of a "minimum" *k* value, proposing an a posteriori validation.

Results and Discussions

Here we want to discuss the impact of the template-set selection upon template generation quality, carrying out a number of tests. For this purpose, we consider the possibility of an accurate extraction procedure (hereafter called "normal") based on a set S'_k of HBTs, having cardinality *k* far lower than *n*. The aim is to show how well a small subset of *S*, called ,S'_k, $S'_k \subset S$, is able to fit the task of extracting HBs from an arbitrary set of images, in particular any dataset not related to the one from which the elements of S'_k are selected, taking into account the morphological variability of the population.

We assume that the "exhaustive" extraction of *n* HBs represents our "gold standard," to which we shall compare the *n* boxes, obtained from the extraction performed by the selected HBTs ("normal" extraction).

The hippocampal boxes will now be extracted using as fixed images all the elements of each set S'_k, and coregistering them in turn with all the moving images of a dataset, using the same algorithm as in the "exhaustive" procedure. This way of extracting the HBs doesn't require a determined first fixed image to start, but the registration procedure of each *n* moving images is performed with respect to all the *k* HBTs. In terms of computation, the extraction cost is $n \times k$, instead of $n(n+1)/2$ as in the "exhaustive" procedure, and even if a number of images are added to the dataset the procedure doesn't need to be restarted.

First we observe that it is not reasonable to evaluate template generation quality through the correlation coefficient between each box extracted by the "exhaustive" procedure (ie, $HB_j{}^x$) and the same box extracted with the "normal" one at given *k* (ie, $HB_j{}^k$). In fact, as expected due to the rationale of the procedure, the mean value of the correlation coefficient between all the ($HB_j{}^x$, $HB_j{}^k$) couples, for any *k*, gives a strong correlation between the two sets of boxes. With a *t*-test applied to the two sets of boxes we have a probability $<10^{-6}$ that the difference between the two means is caused by chance.

Therefore the correlation coefficient, although useful in the extraction process, is not sensible enough to discriminate between the boxes extracted by each S'_k.

To give this estimation, we propose a metric based on the geometrical position of the boxes. For each MR_j image, we take the corresponding $HB_j{}^x$ extracted by the "exhaustive" procedure and $HB_j{}^k$ extracted by the "normal" one (for each *k* value).

Let V_{j1} and V_{j2} be two opposite vertices of the "exhaustive" box $HB_j{}^x$, and V'_{j1} and V'_{j2} the corresponding vertices in the "normal" box, we calculate a parameter $D_j(k)$ as the mean of the Euclidean distances (*dist*) between V^x_{j1} and V^k_{j1}, and between V^x_{j2} and V^k_{j2}.

By averaging $D_j(k)$ on all the *n* boxes, we obtain a parameter that measures how much the HB set extracted by the "normal" method at a particular *k* value reproduces the "exhaustive" one:

$$D(k) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{2}\left[dist\left(V^x_{j1}, V^k_{j1}\right) + dist\left(V^x_{j2}, V^k_{j2}\right)\right]. \quad (2)$$

In principle, if we would use as HBTs all the boxes obtained by the "exhaustive" procedure and clusterized for $k = n$, this parameter would vanish, because any template would extract itself with the maximum accuracy, for this reason we

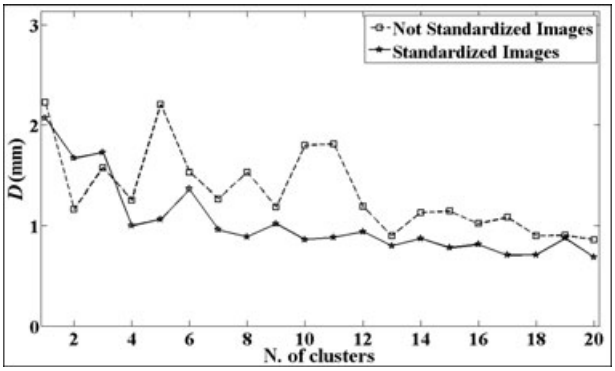

Fig 4. The pattern of the *D* parameter versus the number *k* of clusters, obtained by *k*-means clusterization method, when performed before or after the histogram standardization step.

expect a downward trend of $D(k)$ when the number of templates increases.

Figure 3 shows the D parameter versus the number k of clusters, for a dataset of 190 subjects with clinical conditions (AD) from Normal to MCI to AD, comparing two different clusterization methods, k-means and hierarchical, respectively. We observed that the accuracy error shown by the extraction process does not depend on the clusterization method, and both the plots confirm the decreasing trend of $D(k)$.

In Figure 4 the D parameter versus the number k of clusters is shown, for the same dataset of 190 subjects, obtained by the k-means clusterization method, with and without[14] the histogram standardization step. In general, D values are lower when standardization is applied and the trend of the curve exhibits smaller fluctuations, with a more stable behavior. These aspects affect the choice of a "minimum" k number of templates, as explained hereafter.

In Figure 5A a plot of the D parameter versus the number k of clusters is shown, for three different but homogeneous sets of MR images, belonging to subjects with analogous clinical conditions (ie, Normal or MCI or AD). Also, in these cases the $D(k)$ graphs show a similar decreasing trend as in Figure 3.

Now we want to discuss the possibility of estimating a "minimum" k number of templates, able to faithfully reproduce the extraction performed by the "exhaustive" procedure. In a sense, this set of k templates would be capable of describing the morphological variability of the whole image dataset.

As already pointed out, $D(k)$ has a downward trend, therefore we could estimate a "minimum" k value (hereafter called $k_{\min}$), where this function becomes stable and low enough, as a good compromise for describing population variability.

Two possible approaches to the choice of $k_{\min}$ can be addressed, involving absolute and relative thresholds.

An absolute threshold leads to an estimation of a suitable $k_{\min}$ as the value for which $y(k)$ becomes lower than an arbitrary but "small enough" threshold value θ, for example, 1 or .5 mm. This is a natural approach, and, for example, the choice of $\theta = 1$ mm applied in the case of the nonhomogeneous dataset (Fig 3) leads to $k_{\min} \approx 7$.

Extending this absolute θ value to the different $D(k)$'s in Figure 5 would lead to $k_{\min} \approx 4$, that is, a template set composed of only four templates. However, observing the Figure 5 it is evident that the $D(k)$ data is rapidly varying in the region around this $k_{\min}$ value. Therefore, this estimation cannot take into account that no stability is achieved in this k domain.

According to these observations, the option of choosing a relative approach was explored. It consists in an estimation obtained by fitting $D(k)$ with a parameterized exponential function, such as $y(k) = a\, e^{-k/b} + c$. This function is applied as fitting curve both to the mean values of the two sets of data in Figure 3 (solid line) and to the MCI data (Fig 5B).

The fitting lines seem to describe the experimental data quite well. Then, a suitable $k_{\min}$ can be chosen where y becomes stable enough and "small."

For example, we can decide to select $k_{\min}$ as the value for which the decreasing exponential reaches 10% of its amplitude, after subtracting the baseline, that is, $y(k_{\min}) = a/10+c$. This means that $k_{\min} = b\, ln\, 10$, that is, the "minimum" k, only depends (linearly) on the b parameter.

In all the cases we analyzed, we remarked that $D(k)$ fitting functions in homogeneous datasets are characterized by smaller a, b, and c values than nonhomogeneous ones, that is, $y(k)$ starts lower, goes down faster, and gets lower for high k values, as intuitively expected. Therefore, choosing $k_{\min}$ in a relative way assures us that the correlation between "homogeneous sample" and "small $k_{\min}$" is respected, reflecting the high intersubject anatomy similarity in homogeneous datasets.

From the above considerations, in the case of a relative threshold, $k_{\min} \approx 11$ can be considered the "minimum" number of templates able to describe the morphological variability of the hippocampal region in a sample of subjects with clinical conditions (AD) from Normal to MCI to AD (Fig 4). The "minimum" k for the three datasets of Figure 5 (solid line in Fig 5B concerns MCI data) is obtained for a lower value, about 7, as expected due to dataset homogeneity.

Comparing the $k_{\min}$ values obtained by the two different approaches, we can conclude that

(1) both approaches give lower values for homogeneous datasets with respect to nonhomogeneous ones; and
(2) in our opinion, the choice of a relative threshold gives more reliable results.

Another observation can be made. Due to variability in shape and/or asymmetry between left and right hippocampus, is it possible that $D(k)$ can be different in the two cases, and hence may yield to different $k_{\min}$ values for the same population. In point of fact we assessed for all the datasets we investigated, that the choice of a "minimum" k with a relative threshold gives no significant difference for left and right hippocampi.

With the aim of testing the generality of our procedure, we wanted to perform template-set selection on other regions, rather well characterized in the brain, properly choosing the initial coordinates and dimensions of the fixed image. A first investigation on a second brain zone ($40 \times 40 \times 40$ pixel large), near the amygdala on a set of 100 new ADNI images not included in the dataset from which the hippocampal extraction was performed, shows a downward trend of $D(k)$ versus the number of templates, which is promising and suggests that the procedure may be generalized to other brain regions, as long as

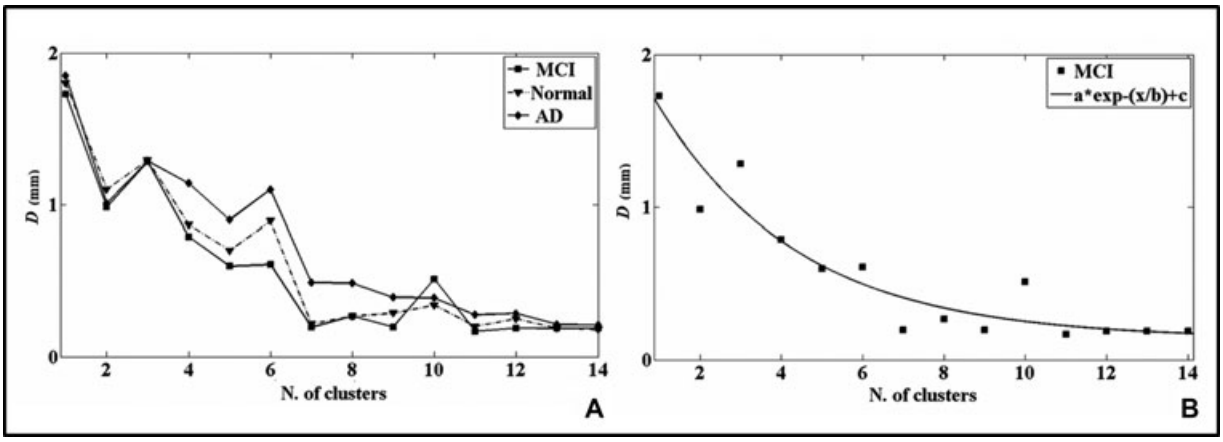

Fig 5. (A) The pattern of the *D* parameter versus the number *k* of clusters, for datasets of images belonging to patients with homogeneous clinical conditions (Alzheimer's disease). (B) Fitting curve on the *D* parameter for a dataset of images belonging to MCI patients.

the chosen regions are well distinguishable from surrounding tissue.

As regards the effectiveness of the proposed procedure, Calvini et al[12] evaluated the efficiency of the "normal" extraction by HBTs generated, as described, on images belonging to patients with different clinical conditions. They showed that the use of the HBTs significantly reflects on the capture of differences between subgroups of interest with different stages of cognitive impairment, with comparable discriminating capability between MCI converters and controls and between AD patients and controls. We repeated these tests, investigating the influence of histogram standardization on the ability of the Δ-box projection to discriminate between the different classes of subjects. The same simple classification scheme as in Calvini et al[12] was adopted, and ROC curves were drawn for each classification task (eg, control patients vs AD, AD vs MCI, etc). Our calculations showed that the introduction of preliminary gray level standardization really makes the classification task more accurate, with AUC values that increased about two-three percentage points, when the images were acquired with the same equipment. Comparing with the work in Calvini et al, we also used datasets that were particularly nonhomogeneous in terms of gray level scale, and the increase was even more evident. In one case, for example, the discriminating power of the Δ-box projection increased from AUC value equals to .78 without histogram standardization and to .86 with standardization. ROC curve examples, controls versus AD, for a dataset with strong gray level scale differences (with and without histogram standardization) are shown in the Supporting Information.

Conclusions

In conclusion, we detailed a procedure for generating a set of templates for the hippocampal region, considering various MRI datasets, different for cardinality and (homogeneous/nonhomogeneous) conditions of the population.

In general, the images are not based on a common standard gray level scale and the extraction-procedure accuracy would be severely affected, so robust standardization is applied. Then an "exhaustive" extraction of the hippocampal region is performed, starting from a fixed hippocampal box (HB_0). At the beginning, the early extractions exhaust the set of the HBs which are very similar to the defined HB_0. Then, the procedure continues extracting HBs that are progressively different from the first ones, but diversity creeps into the growing HB database very slowly, thanks to the relevant size of the population of the available MR images. Thus, the orientation and position of the essential geometrical features of the searched region are preserved during the whole process of HB extraction.

Then the extracted regions are clusterized to choose the most representative images in a large population, proposing an estimation of the "minimum" number of templates that fulfill our purpose.

We assess that this "minimum" number of templates is largely independent on the clusterization method and on the number of the MR images, if statistically consistent with respect to the clinical conditions of the patients.

From the above, the best strategy, to be used when nonhomogeneous populations are considered, strictly depends on the features and characteristics we want to emphasize better.

We stress that

(1) the extraction procedure is severely affected when the images do not come from the same source (hospital/scanner) and do not undergo proper histogram standardization. On the contrary, when this step is applied, it is easier to propose a "minimum" number of templates. Moreover, the discriminating power of indicators based on the gray values of the extracted boxes is increased;
(2) the "exhaustive" extraction of a given region is mandatory and to be performed on a large number of images only once;
(3) a "minimum" number of templates can be estimated for the hippocampal region;
(4) we chose the templates not as the mean images of a dataset, but as a group of the most representative ones; and
(5) the number of templates able to describe the population is lower for patients with homogeneous clinical conditions than with mixed degrees of neuropathology (eg, AD).

The selected template set can be used for the extraction/evaluation of that region on various and different MRI datasets.[33]

The procedure is simple to use and can be considered a promising approach in atlas generation. Because the first results are encouraging, future work will be to extensively generalize and apply the procedure to other brain regions, which may be relevant for future neuroimaging studies.

We warmly thank our anonymous reviewers for their pertinent comments and useful suggestions. This work is supported by the Italian "INFN" and the Italian "Ministero dell'Istruzione, dell'Università e della Ricerca" (MIUR). Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129, K01 AG030514, and the Dana Foundation. This work was undertaken at UCLH/UCL with funding from the Department of Health's National Institute of Health Research Centres funding scheme, the Medical Research Council and Alzheimer's Research UK. The Dementia Research Centre is an Alzheimer's Research UK Co-ordinating Centre and has also received equipment funded by the Alzheimer's Research UK.

References

1. Lalys F, Haegelen C, Ferre JC, et al. Construction and assessment of a 3-T MRI brain template. *NeuroImage* 2010;49:345-354.
2. Henneman WJP, Sluimer JD, Barnes J, et al. Hippocampal atrophy rates in Alzheimer disease: added value over whole brain volume measures. *Neurology* 2009;72:999-1007.
3. Rosas H, Feigin A, Hersch S. Using advances in neuroimaging to detect understand and monitor disease progression in Huntington's disease. *NeuroRX* 2004;1(2):263-272.
4. Kandel E, Schwartz J, Jessell T. Principles of neural science. Norwalk, CT: Appleton & Lange, 2000:1414.
5. Talairach J, Tournoux P. *Co-planar Stereotaxic Atlas of the Human Brain: 3-Dimensional Proportional System: an Approach to Medical Cerebral Imaging*. Stuttgart: Thieme Medical Publishers, 1988:122.
6. Ono M, Kubik S, Abernathey CD. *Atlas of the Cerebral Sulci*. Stuttgart; New York: Thieme Medical Publishers, 1990:218.
7. Seghers D, D'Agostino E, Maes F, et al. Construction of a brain template from MR images using state-of-the-art registration and segmentation techniques. *Lect Notes Comput Sci* 2004;3216:696-703.
8. Lee JS, Lee DS, Kim J, et al. Development of Korean standard brain templates. *J Korean Med Sci* 2005;20:461-483.
9. Baloch S, Davatzikos C. Morphological appearance manifolds in computational anatomy: groupwise registration and morphological analysis. *NeuroImage* 2009;45:S73-S85.
10. Shattuck D, Mirza M, Adisetiyo V, et al. Construction of a 3D probabilistic atlas of human cortical structures. *NeuroImage* 2008;39(3):1064-1080.
11. Mazziotta JC, Toga AW, Evans AC, et al. A four-dimensional probabilistic atlas of the human brain. *J Am Med Inform Assoc* 2001;8(5):401-430.
12. Calvini P, Chincarini A, Gemme G, et al. Automatic analysis of medial temporal lobe atrophy from structural MRIs for the early assessment of Alzheimer disease. *Med Phys* 2009;36:3737-3747.
13. Golosio B, Masala GL, Piccioli A, et al. A novel multi-threshold method for nodule detection in lung CT. *Med Phys* 2009;36:3607-3618.
14. Wolz R, Aljabar P, Hajnal J, et al. LEAP: learning embeddings for atlas propagation. *NeuroImage* 2010;49:1316-1325.
15. Chen Y, Shen D, Zhu H, et al. *Hierarchical Unbiased Group-wise Registration for Atlas Construction and Population Comparison*. Florida: SPIE Medical Imaging, 2009.
16. Joshi S, Davis B, Jomier M, et al. Unbiased diffeomorphic atlas construction for computational anatomy. *NeuroImage* 2004;23:S151-S160.
17. Aljabar P, Heckemann RA, Hammers A, et al. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. *NeuroImage* 2009;46:726-738.
18. Nyul LG, Udupa JK. On standardizing the MR image intensity scale. *Magn Reson Med* 1999;42:1072-1081.
19. Nyúl LG, Udupa JK, Zhang X. New variants of a method of MRI scale standardization. *IEEE Trans Med Imag* 2000;19(2):143-150.
20. Ge Y, Udupa JK, Nyúl LG, et al. Numerical tissue characterization in MS via standardization of the MR image intensity scale. *J Magn Reson Imag* 2000;12(5):715-721.
21. Hellier P. Consistent intensity correction of MR images. *Proceedings of the IEEE International Conference on Image Processing*. Barcelona, Spain, 2003: Vol 1, 1109-1112.
22. Weisenfeld N, Warfield S. Normalization of joint image-intensity statistics in MRI using the Kullback-Leibler divergence. *Proceedings of the IEEE International Symposium on Biomedical Imaging*. Arlington, VA, 2004:101-104.
23. Schmidt M. A method for standardizing MR intensities between slices and volumes. Technical Report. Edmonton, AB: University of Alberta, 2005:TR05-14.
24. Jäger F, Hornegger J. Nonrigid registration of joint histograms for intensity standardization magnetic resonance imaging. *IEEE Trans Med Imag* 2009;28(1):137-150.
25. Leung KK, Clarkson MJ, Bartlett JW, et al. Robust atrophy rate measurement in Alzheimer's disease using multi-site serial MRI: tissue-specific intensity normalization and parameter selection. *NeuroImage* 2010;50:516-523.
26. Bergeest JP, Jäger F. A comparison of five methods for signal intensity standardization in MRI. In Tolxdorff T, Braun J, Deserno TM, et al, eds. *Bildverarbeitung fur die Medizin*, in series: *Informatik aktuell*. Heidelberg: Springer Berlin, 2008: 36-40.
27. Ashburner J, Friston KJ. Unified segmentation. *Neuroimage* 2005;26:839-851.
28. Evans A, Kamber M, Collins D, et al. An MRI-based probabilistic atlas of neuroanatomy. In: Shorvon S, Fish D, Andermann F, et al, eds. *Magnetic Resonance Scanning and Epilepsy*, in series: *NATO ASI Series A Life Sciences*, Vol 264. New York: Plenum Press, 1994:263-274.
29. Ibanez L, Schroeder W, Ng L, Cates J. *The ITK Software Guide*. 2nd ed. Updated for ITK version 2.4, 2005 (Kitware Inc). ISBN: 1-930934-15-7.
30. Seber G. *Multivariate Observations*. New York: J Wiley and Sons, 1984.

31. Esposito M, Bosco P, Rei L, et al. Volumetric analysis on MRI and PET images for the early diagnosis of Alzheimer's disease. *Nuovo Cimento C* 2011;34(1):175-185.
32. Wang W, Zhang Y. On fuzzy cluster validity indices. *Fuzzy Sets Syst* 2007;158:2095-2117.
33. Rohlfing T, Brandt R, Menzel R, et al. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. *NeuroImage* 2004;21(4):1428-1442.

Supporting Information

Additional supporting information may be found in the online version of this article:

Figure S1. Left: the initial hippocampal box (HB_0) from which an "exhaustive" extraction started Right: Templates selected after clusterization for k = 12.

Figure S2. ROC curve examples, controls vs AD, for a dataset with strong gray level scale differences. The introduction of preliminary gray level standardization (black) makes the classification task more accurate than without standardization (blue).

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.