Medical Image Analysis 82 (2022) 102576

Contents lists available at ScienceDirect

Medical Image Analysis

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

Quantifiable brain atrophy synthesis for benchmarking of cortical thickness estimation methods

Filip Rusak [a,b,*], Rodrigo Santa Cruz [a], Léo Lebrat [a], Ondrej Hlinka [a], Jurgen Fripp [a], Elliot Smith [c], Clinton Fookes [b], Andrew P. Bradley [b], Pierrick Bourgeat [a], for the Alzheimer's Disease Neuroimaging Initiative [1]

[a] The Australian eHealth Research Centre, CSIRO Health and Biosecurity, 296 Herston Road, Herston, 4029, Australia
[b] Queensland University of Technology, 2 George St, Brisbane, 4000, Australia
[c] Maxwell Plus, 310 Edward St, Brisbane, 4000, Australia

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

Keywords:
Synthetic atrophy
Brain MRI
Generative adversarial networks

ABSTRACT

Cortical thickness (CTh) is routinely used to quantify grey matter atrophy as it is a significant biomarker in studying neurodegenerative and neurological conditions. Clinical studies commonly employ one of several available CTh estimation software tools to estimate CTh from brain MRI scans. In recent years, machine learning-based methods emerged as a faster alternative to the main-stream CTh estimation methods (e.g. FreeSurfer). Evaluation and comparison of CTh estimation methods often include various metrics and downstream tasks, but none fully covers the sensitivity to sub-voxel atrophy characteristic of neurodegeneration. In addition, current evaluation methods do not provide a framework for the intra-method region-wise evaluation of CTh estimation methods. Therefore, we propose a method for brain MRI synthesis capable of generating a range of sub-voxel atrophy levels (global and local) with quantifiable changes from the baseline scan. We further create a synthetic test set and evaluate four different CTh estimation methods: FreeSurfer (cross-sectional), FreeSurfer (longitudinal), DL+DiReCT and HerstonNet. DL+DiReCT showed superior sensitivity to sub-voxel atrophy over other methods in our testing framework. The obtained results indicate that our synthetic test set is suitable for benchmarking CTh estimation methods on both global and local scales as well as regional inter-and intra-method performance comparison.

1. Introduction

Quantitative analysis of brain Magnetic Resonance Imaging (MRI) may serve as a powerful tool for studying neurodegenerative and neurological disorders if the measurements are accurate and reliable (Rebsamen et al., 2020a). Accuracy and reliability are crucial when measuring biomarkers influenced by confounding factors such as cortical atrophy over time (Sharma et al., 2010). Some of the factors that impact cortical atrophy and its rate of change can be natural such as aging (Jernigan et al., 2001), while the others are related to a particular pathology (Whitwell et al., 2007) or lifestyle factors (Fein et al., 2002). The average yearly global atrophy rate in healthy people increases gradually from 0.2% in the age bracket 30–50 to 0.3%–0.5% in the age bracket 70–80 years old (Fox and Schott, 2004), while the average global atrophy rate in Alzheimer's disease (AD) patients is around 2.8% (Sluimer et al., 2008). Since the average cortical thickness (CTh) across regions ranges between 2.5 to 3 mm, with the local extremes between 1–4.5 mm, CTh estimation methods must be sensitive to sub-voxel change measurements from brain MRI scans (Hutton et al., 2008; Clarkson et al., 2011).

Several software packages that provide CTh estimation are currently available and used in clinical studies, such as FreeSurfer (Fischl, 2012), ANTs (Avants et al., 2009), FastSurfer (Henschel et al., 2020), DL+DiReCT (Rebsamen et al., 2020a) and HerstonNet (Santa Cruz et al., 2021). The software packages can be broadly clustered into three groups: surface, registration and machine learning-based methods. Nevertheless, the increased number of available CTh estimation methods boosts the variability between their results (Khanal et al., 2016a).

* Corresponding author at: The Australian eHealth Research Centre, CSIRO Health and Biosecurity, 296 Herston Road, Herston, 4029, Australia.
E-mail address: filip.rusak@csiro.au (F. Rusak).

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.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.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

https://doi.org/10.1016/j.media.2022.102576
Received 13 December 2021; Received in revised form 10 May 2022; Accepted 11 August 2022
Available online 24 August 2022

Since CTh is a well-established biomarker in the context of clinical studies focusing on neurodegenerative conditions, benchmarking of available CTh estimation methods is a fundamental problem (Sharma et al., 2010). Currently, it is impossible to perform such a comparison on a real MRI test set due to non-existing clinically applicable ground truth of absolute thickness (Sharma et al., 2010; Bernal et al., 2021). Even non-routinely available manual brain segmentation, which is time-consuming and requires highly-specialised anatomical expertise, may result in significant intra- and inter-rater variability (Bergouignan et al., 2009) and do not provide sub-voxel accuracy. Therefore, FreeSurfer measurements are typically used to generate 'silver-standard' ground truth (Rebsamen et al., 2020b), which implies an evaluation bias.

The evaluation of CTh estimation methods commonly includes group analysis (Henschel et al., 2020; Rebsamen et al., 2020a), test-retest reliability assessment, metrics like Intraclass Correlation Coefficient (ICC) (Henschel et al., 2020; Rebsamen et al., 2020b; Santa Cruz et al., 2021), Dice similarity coefficient (Henschel et al., 2020) (segmentation as a prerequisite for CTh estimation - FreeSurfer as ground truth), coefficient of determination (R^2) (Rebsamen et al., 2020b), average Hausdorff distance (Henschel et al., 2020) for segmentation boundaries quality evaluation, longitudinal annual cortical grey matter (GM) atrophy rates (Rebsamen et al., 2020a) for different methods and comparison of global mean thickness (Rebsamen et al., 2020a) for robustness. While the listed evaluation metrics cover a broad range of aspects relevant for CTh estimation methods evaluation, none especially evaluates each method's sensitivity to sub-voxel changes on global and local scales.

In this paper, we address the problem of the missing ground truth in the context of CTh estimation evaluation by proposing a method for brain MRI synthesis with various quantifiable GM atrophy levels. Since clinically applicable ground truth of absolute thickness does not exist, we focus on clinically relevant thickness change measurements by synthesising baseline and follow-up brain MRIs and computing their CTh differences. We use our method to generate a synthetic dataset that consists of 20 subjects with 19 sub-voxel atrophy levels per subject. When synthesising the dataset, we uniformly introduce atrophy into synthetic brain MRIs across 34 brain regions (on both hemispheres). This approach enables per-region inter-and intra-method performance comparison on different atrophy levels. Moreover, such synthetic dataset is a great asset when determining the minimal level of atrophy a method can detect. Furthermore, we show that the proposed method can also generate synthetic brain MRIs with localised atrophy in an individual region. Synthesising atrophy in a single brain region enables more localised evaluation of CTh estimation methods. It also opens up the opportunity to model a more natural sequence of atrophy between synthetic time points (brain MRI scans of the same subject containing a certain level regional changes). Finally, we thoroughly evaluate the performance of three types of CTh methods (four methods) on our synthetic dataset: FreeSurfer cross-sectional (surface-based), FreeSurfer longitudinal (surface-based), DL+DiReCT (registration-based) and HerstonNet (machine learning-based). This is the first study that benchmarks surface, registration and machine-learning-based CTh estimation methods. The obtained results revealed valuable insights into the regional performance of CTh estimation methods under test with implications for future clinical studies.

In summary, alongside previously mentioned CTh estimation evaluation metrics, we show that our method and synthetic dataset largely contributes to building the more complete picture of local and global evaluation.

2. Related work

Several studies tackled the problem of missing ground truth for the evaluation of atrophy estimation methods by simulating atrophy (Smith et al., 2003; Camara et al., 2006; Karaçali and Davatzikos, 2006; Pieperhoff et al., 2008; Sharma et al., 2010; Khanal et al., 2016b; Larson and Oguz, 2021; Bernal et al., 2021). All listed methods, except (Bernal et al., 2021), compute a deformation field and then apply it to real brain MRIs to obtain simulated scans with a certain level of atrophy. These deformation-based methods can be further grouped into Jacobian based (Karaçali and Davatzikos, 2006; Pieperhoff et al., 2008; Sharma et al., 2010), biomechanical model-based (Smith et al., 2003; Camara et al., 2006; Khanal et al., 2016b) and morphological operation-based (Larson and Oguz, 2021).

Karaçali et al. in Karaçali and Davatzikos (2006) proposed a Jacobian-based method for brain tissue atrophy simulation in selected regions of brain MR scans. The authors estimated a topology-preserving deformation field to simulate a predefined set of volumetric changes, described in the form of a tissue loss statistical atlas, in a particular region. The deformation fields are estimated by minimising the sum of squared differences between the Jacobian determinants of the transformation and the desired Jacobian determinants which describe volumetric changes per voxel. To secure that estimated deformation fields preserve topology, Karaçali et al. also introduced a penalisation term that ensures positive corner Jacobians.

Pieperhoff et al., in Pieperhoff et al. (2008), proposed a method similar to Karaçali and Davatzikos (2006). The main difference between the two methods is that Pieperhoff et al. used Local Volume Ratio (LVR) instead of Jacobian determinants in their cost function. In this context, LVR describes the ratio between distorted voxel volume (source MRI) and voxel volume (target MRI). Pieperhoff et al. clarified that there is no fundamental difference between the LVR and the Jacobian determinant in the context of 3D deformation fields. The main advantage of LVR over Jacobian determinant approximations are smoother volume measures. Additionally, Pieperhoff et al. also added a regularisation term to secure smooth estimated transformation.

Sharma et al. proposed a method for brain tissue loss simulation that estimates deformation fields based on topology-preserving B-spline (Sharma et al., 2010). Instead of using the sum of squared differences between the Jacobian of the transformation and the desired level of atrophy, as in Karaçali and Davatzikos (2006), the authors considered the logarithm of the Jacobian where dilations and contractions impact the objective function similarly. Sharma et al. also introduced additional constraints to ensure the skull invariance by the estimated transformation. The skull invariance constraint is implemented by optimising B-spline parameters only and initialising other parameters to zero, which do not affect the skull.

Smith et al. proposed a phenomenological model for brain atrophy simulation on MR scans (Smith et al., 2003). The model simulates tissue volume reduction, ventricular cerebrospinal fluid (CSF) volume expansion and biomechanical tissue readjustment. The authors simulate atrophy by applying thermal loads to one or more brain tissue types to realise tissue expansion or contraction, while the Finite Element Method (FEM) is used for modelling mechanical readjustment. This finite element model does not model extra-cortical CSF, instead, it estimates CSF volume change derived from MRIs and calculated over the entire cranial cavity.

Camara et al. proposed a biomechanical-based method capable of simulating atrophy with a phenomenological model in various tissue compartments or neuroanatomical structures Camara et al. (2006). The authors used a fluid registration model to wrap a 3D mesh, created by the meshing of a labelled brain atlas, onto an image. To simulate atrophy, Camara et al. used the FEM solver and thermoelastic model of tissue deformation, where the atrophy rate progression is controlled by thermal coefficients, each corresponding to a separate tissue type. Brain structure segmentations are required to create FEM input and brain structures biomechanical readjustment is modelled with conventional physics-based techniques relying on biomechanical tissue properties. The proposed method also includes a skull invariance constraint.

Khanal et al. proposed a framework for brain atrophy simulation and prediction of realistic longitudinal AD follow-ups from a baseline brain MRI (Khanal et al., 2016b). The proposed framework includes three steps: atrophy generation, brain deformation and realistic MRI generation. The authors developed a biophysical brain deformation model that can realistically model various complex atrophy patterns by manipulating model parameters, even when the amount of prescribed atrophy does not change. The model reflects the shape deformation caused by neuronal death and brain atrophy. The prescribed atrophy is modelled by minimising the strain energy formulated as an instance of Saint Venant–Kirchhoff model. The atrophy is simulated by computing the deformation field from the amount of atrophy assigned to each voxel (volume change) while permitting the CSF to expand and compensate for freed volume caused by atrophy. The deformation field is computed by numerically solving a system of Partial Differential Equations (PDEs) with the Finite Difference Method (FDM). Finally, the computed deformation field is used to obtain simulated follow-up MRI by warping the baseline image.

Larson et al. proposed a registration-based method for the synthesis of longitudinal ground truth dataset, intended for validation of surface-based CTh estimation methods (Larson and Oguz, 2021). The authors obtained a FreeSurfer cortical parcellation, selected an anatomical structure, upsampled the selected mask (400%) and performed a set of binary mathematical operations (erosion of GM and dilatation of white matter - WM with equal erosion/dilatation kernel size) on the high-resolution mask to manipulate morphology. Once Larson et al. obtained the binary atrophied mask, they computed the deformation field by registering the initial to the atrophied mask. Finally, the authors warped the derived deformation field to the corresponding brain MRI and obtained an MRI with the same atrophy as introduced in binary masks. To compute the exact introduced changes, the authors obtained the surface of the baseline and a synthetic atrophied baseline by running the marching cubes algorithm and measuring their differences vertex by vertex. The proposed method can introduce atrophy in the range between [0.6, 2.6] mm.

Bernal et al. proposed a DL-based framework for longitudinal dataset generation that utilises T1-w brain MRI scans and corresponding segmentation probability maps to generate synthetic samples through probability map deformations Bernal et al. (2021). In particular, the authors trained a cascaded multi-path U-Net with multi-objective loss function on pairs of real T1-w brain MRI scans and corresponding segmentation probability maps of three tissues (GM, WM and CSF). The model was trained on patches ($32 \times 32 \times 32$ voxels). Once the authors trained a model, they altered the input segmentation probability maps by applying deformation fields, computed using FNIRT (Andersson et al., 2007) on subjects with high atrophy levels. The altered segmentation probability maps and the baseline MRI scan are then used as an input to the trained generative model, which creates an MRI scan that reflects introduced changes.

The main pitfall related to deformation field-induced atrophy methods (Smith et al., 2003; Camara et al., 2006; Karaçali and Davatzikos, 2006; Pieperhoff et al., 2008; Sharma et al., 2010; Khanal et al., 2016b; Larson and Oguz, 2021; Bernal et al., 2021) is the lack of control over the location and magnitude of introduced atrophy. While modelling localised atrophy is still possible with deformation-based methods, controlling the magnitude of the introduced atrophy in a sub-voxel range is not trivial. To localise the introduced CTh changes, Larson et al. utilised FreeSurfer-derived cortical parcellation maps, while the magnitude of CTh changes is controlled by the erosion kernel size, limiting the minimal CTh change to be 0.6 mm (Larson and Oguz, 2021). In contrast, we introduce atrophy on the mesh, which gives us control when modelling atrophy in the sub-voxel range (with atrophy levels smaller than 0.6 mm) and the ability to quantify the amount of atrophy introduced in a vertex-by-vertex fashion. Thus our method has the potential to greater control in both the extent and magnitude of the introduced atrophy.

While Larson et al. (Larson and Oguz, 2021) used meshes obtained by marching cubes to quantify the amount of atrophy introduced during the morphological operations on the binary segmentation maps, our approach was to deform FreeSurfer-derived meshes (vertex-by-vertex) to create Partial Volume (PV)-maps needed for MRI synthesis.

Additional benefit of per-vertex change introduction is the possibility of uniform atrophy simulation across mesh vertices. The benefit of introducing uniformly distributed atrophy is inter-and intra-method region-wise comparison, which is not possible with the above-listed methods.

Furthermore, methods that use automatic segmentations (Camara et al., 2006; Pieperhoff et al., 2008; Khanal et al., 2016b; Larson and Oguz, 2021) for anatomical changes simulation and deformation field computation implicitly assume no segmentation method bias, which could influence atrophy measurements between real and deformed MRIs. Despite using automatically derived meshes for PV-maps generation, we reduce segmentation method bias by synthesising baseline and all atrophied time point MRI scans from the same mesh. By doing so, we ensure a systematic bias, in a baseline MRI and corresponding time points, that mitigates the impact on CTh measurements.

Similar to our work, Bernal et al. in (Bernal et al., 2021) also generated synthetic brain MRIs from the representation of three tissues (GM,WM and CSF). Instead of probability maps, as proposed in (Bernal et al., 2021), for the representation of brain anatomy we used PV-maps. Probability maps are voxel-based probability estimations of a tissue class occurrence that may be obtained in several ways, e.g. by registering a substantial number of brains to a common space (Shi et al., 2010) or employing a segmentation network (Bernal et al., 2021). In contrast to probability maps, each voxel value in PV-maps represents the proportion of a particular tissue class in that voxel, computed on a single brain MRI scan for several tissue classes. While the probability of a tissue occurrence in probability maps is computed over a population (less accurate), the actual proportion of a tissue class in a voxel, represented with PV-maps, is derived from an individual brain MRI scan (more accurate). Therefore, probability maps are less reflective of actual tissue class proportion in a single voxel in comparison to PV-maps, which may stand for a possible source of noise in their model (Tohka, 2014).

Another difference between our work and the previously mentioned studies is the metric used for sensitivity to atrophy. All mentioned work except (Larson and Oguz, 2021) estimate whole-brain volume change as a unit of sensitivity to atrophy. In contrast to volume measures, limited by the defined ROIs, CTh is a more suitable metric for localised brain atrophy measurements as it enables vertex-based analysis (Burggren et al., 2008). While Larson et al. in their work focus on atrophy synthesis for accuracy validation of surface-based CTh estimation methods only (Larson and Oguz, 2021), we propose a method for evaluation of CTh estimation methods, either vertex or voxel based.

In summary, the main advantage of our method, in comparison to prior work, is that our method can synthesise novel follow-up scans with quantifiable variations in CTh, both globally (uniformly) and locally. The benefit of introducing uniformly distributed atrophy is inter-and intra-method region-wise comparison. Such comparison is essential to understand the per-region performance of a method under test. Further, the comparison results may ease the CTh estimation method selection for a given context (CTh measurements for semantic dementia vs AD). The benefit of introducing local atrophy, per-region or vertex-by-vertex, enables modelling disease progression (e.g. AD) over time.

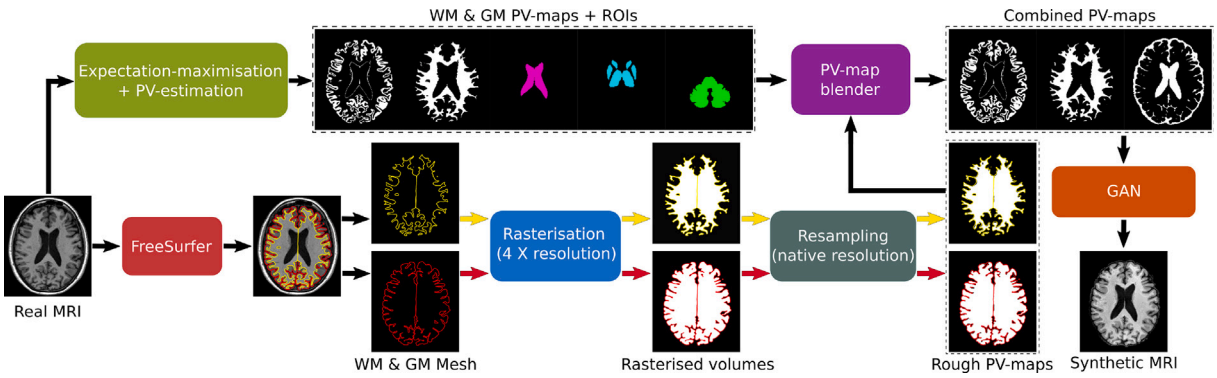

Fig. 1. Workflow: synthetic brain MRI generation from PV-maps of GM, WM and CSF derived from FreeSurfer surface meshes.

3. Methods

3.1. Data and pre-processing

In our experiments, we used a subset of 1000 baseline T1-weighted (T1-w) brain MRI scans from ADNI[2] (Jack Jr. et al., 2008; Weiner et al., 2017). The ADNI subset was selected randomly, with an equal proportion of three diagnosed pathologies among both genders. The subset includes 40.6% subjects diagnosed as Healthy Control (HC), 41.2% as Mild Cognitive Impairment (MCI) and 18.2% as AD. All 1000 scans in the ADNI subset were acquired with a 3T scanner as a part of ADNI1/2/3/GO studies. The downloaded scans included in ADNI1/2/Go were N3 corrected, while scans included in ADNI3 were raw (not preprocessed). The ADNI subset was randomly split into disjoint train, validation and test sets, with the 60:20:20 ratio, respectively. The population of ADNI subjects and the data split used in our previous work (Rusak et al., 2021) is consistent with the present work. The disjoint sets were stratified based on gender and pathology (HC, MCI and AD). The whole ADNI subset of 1000 brain MRI scans were further processed by performing bias field correction in the brain region of interest (ROI) (Van Leemput et al. (1999a), rigid registration to the MNI-space ($181 \times 217 \times 181$ voxels) and z-score intensity normalisation with the mean value computed from brain ROI.

3.2. PV-map generation

Every segmentation and surface reconstruction method comes with a bias and requires bias management. In the context of brain atrophy simulation, automatic segmentations are commonly used for deformation field computation (Camara et al., 2006; Pieperhoff et al., 2008; Khanal et al., 2016b; Larson and Oguz, 2021). While the computed deformation fields may be able to deform the corresponding MRI to reflect the desired atrophy, the deformation field still carries the segmentation/surface reconstruction bias. Therefore, to mitigate the method-specific surface reconstruction bias in the cortex region, we choose the surface-centric approach and synthesise brain MRIs to fit the initial surfaces. The synthetic brain MRIs are generated from surface-derived PV-maps. The PV-maps generation process is illustrated in Fig. 1.

We use FreeSurfer cross-sectional pipeline to derive the initial pial and WM surface meshes from a T1-w brain MRI scan. We then rasterise both meshes at a resolution four times higher than the resolution of an original brain MRI scan. Mesh rasterisation was performed using the Visualization Toolkit (VTK) (Geveci et al., 2012). The resulting high resolution GM and WM segmentation maps where resampled to the original resolution using average pooling to generate the PV-maps.

While FreeSurfer provides a surface representation of the cortex, it does not model anatomical structures such as ventricles, deep grey matter or cerebellum which are required for the construction of realistic PV-maps. The meshing in the surface area near the hippocampus, and amygdala are also typically prone to large errors and cannot be used.[3] Therefore, we derive auxiliary PV-maps of three tissue classes (GM, WM and CSF) by segmenting the initial MRIs using the expectation-maximisation (EM) algorithm (Van Leemput et al., 1999b) and estimating PV-maps as detailed in (Acosta et al., 2009). Acosta et al. estimated PV-maps, from tissue segmentation maps and bias field corrected MR images, in two stages. Firstly, they label each voxel as either a single-class (GM, WM, CSF) or multi-class (GM/WM or GM/CSF in the case of GM PV-map). The Potts model, as described in Shattuck et al. (2001), was used for voxels labelling, while Iterative Condition Modes (ICM) algorithm (Besag, 1986) was used for label solving. Secondly, once labelled, fractional content was computed for each voxel. Single-class voxels could be assigned the fractional content value of either 1 (voxel belongs to the class) or 0 (voxel does not belong to the class). The fractional content in multi-class voxels was computed as in Shattuck et al. (2001), ranging between [0, 1]. Once the auxiliary PV-maps are obtained, FreeSurfer parcellation maps are used to mask the ROIs, which correspond to ventricles, deep grey matter, cerebellum, hippocampus and amygdala. These regional PV-maps are then combined with the FreeSurfer-derived PV-maps. A WM PV-map is constructed by replacing the ROI in FreeSurfer-derived WM PV-map with a masked ROI in the auxiliary WM PV-map. A GM PV-map is created by subtracting the constructed WM PV-map from FreeSurfer-derived pial PV-map. A CSF PV-map is created by subtracting the constructed GM and WM PV-maps from the corresponding binarised brain MRI scan. The blending process of auxiliary PV-maps and FreeSurfer-derived (rough) PV-maps is visualised in Figure S1.

3.3. Synthesising global quantifiable brain atrophy

To simulate quantifiable atrophy across all cortical regions, the pial surface is deformed towards the WM surface. Since each vertex on the pial mesh has a corresponding vertex on the WM mesh, we move pial vertices towards the corresponding WM vertices for a given atrophy level (Fig. 2). This approach results in uniform atrophy introduced

[2] Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public–private partnership, led by Principal Investigator Michael W. Weiner,MD. 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 AD. For up-to-date information, see www.adni-info.org.

[3] https://surfer.nmr.mgh.harvard.edu

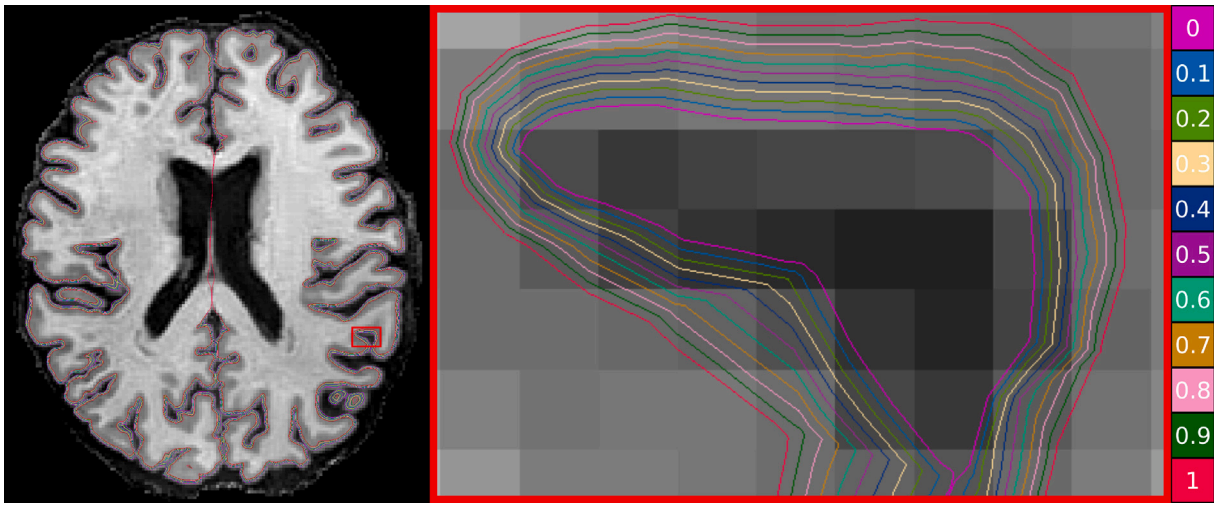

Fig. 2. Global Atrophy: baseline GM meshes and corresponding GM meshes at ten atrophy levels (in the range [0.1, 1] mm with 0.1 mm step between the time points) overlayed on top of a synthetic MRI. The meshes are colour-coded with the legend provided on the right.

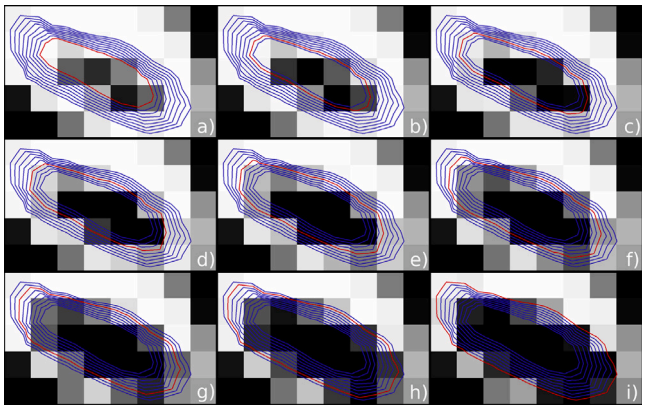

Fig. 3. Visualisation of mesh changes and their effect on mesh-derived PV-maps. The red contour stands for the active level of atrophy, while the blue contours stand for other introduced levels of atrophy. The introduced atrophy levels span across the [0.01, 1] mm range with 0.1 mm step between the atrophy levels (ten atrophy levels).

across all GM cortical regions. To prevent the deformed pial from intersecting with the WM surface mesh, we impose a threshold condition where the introduced atrophy level per-vertex cannot exceed the CTh value associated with the same vertex. In the case when the introduced atrophy exceeds the CTh value, we move the pial towards the WM mesh vertex by the CTh value. The above algorithm is formulated as follows:

$$v(P) = P + \varepsilon(A)\overrightarrow{PW} \tag{1}$$

$$\varepsilon(A) = \begin{cases} \frac{A}{\|\overrightarrow{PW}\|_2} \text{ if } A \leq \|\overrightarrow{PW}\|_2 \\ 1 \text{ otherwise.} \end{cases} \tag{2}$$

where $v(x)$ is a function that deforms pial surface vertex-by-vertex, P is an initial pial vertex and W is a fixed WM vertex. The $\varepsilon(A)$ stands for the conditional function that ensures the maximum level of atrophy not to be larger than the existing CTh at the particular vertex. The A denotes the amount of atrophy introduced on the baseline surface mesh, while $\overrightarrow{PW}$ stands for a vector crossing P and W, corresponding pial and WM vertices respectively. These surface deformations are directly reflected in the PV-maps as presented in Fig. 3, which are then used for MRI synthesis.

3.4. Synthesising local quantifiable brain atrophy

To demonstrate the capability of our method to synthesise quantifiable localised atrophy, we introduced local atrophy in the superior temporal gyrus, based on Desikan–Killiany atlas (Desikan et al., 2006), in 20 subjects. For each subject, ten different levels of atrophy between [0.1, 1] mm were generated, with 0.1 mm steps (Fig. 4). To ensure anatomically plausible deformed surfaces, we impose a smoothness constraint to the vertex displacement defined in Eq. (1). The subset of vertices that construct the boundary of a single anatomical structure (e.g. superior temporal gyrus) is fixed, as shown in Fig. 5 (a) with orange dots. Then, for each vertex on the mesh belonging to the anatomical structure, we computed the gradual attenuation coefficient as the Travelling Salesman Problem (TSP)-based distance (Gavish and Graves, 1978) to the fixed structure boundary and divided it by the total number of discrete attenuation levels. The displaced vertices of the anatomical structure are computed as the product of the introduced atrophy level and the attenuation coefficient. The resulting local atrophy approximation is presented in Fig. 5 (b) with colour-coded attenuation levels.

To formalise, the local atrophy introduced into an anatomical structure $\mathcal{P}_k$ is defined as:

$$w(P) = \begin{cases} v(P)\ \gamma(P) \text{ if } P \in \mathcal{P}_k(P) \\ 0 \text{ otherwise.} \end{cases} \tag{3}$$

where $v(P)$ is the function in Eq. (1), and $\gamma(P)$ is an attenuation function which is proportional to the distance to the boundary of $\mathcal{P}_k$ defined as:

$$\gamma(P) = \min\left(\frac{d_{\bar{\mathcal{P}}_k}(P)}{d_{\max}}, 1\right). \tag{4}$$

The $d_{\bar{\mathcal{P}}_k}$ denotes the gradual attenuation coefficient and $d_{\max}$ the total number of discrete levels, which is in our case 16.

3.5. Generative adversarial network (GAN)

To synthesise realistic brain MRIs from surface-derived PV-maps of three tissue classes (GM, WM and CSF), we employed the high-frequency (HF)-GAN detailed in (Rusak et al., 2021). The HF-GAN is a conditional GAN, composed of a U-Net generator and two discriminators, PatchGAN- and ResNet-based, tailored to facilitate 3D image generation. The HF-GAN is trained sequentially, using one discriminator at a time. The training starts with a PatchGAN-based discriminator,

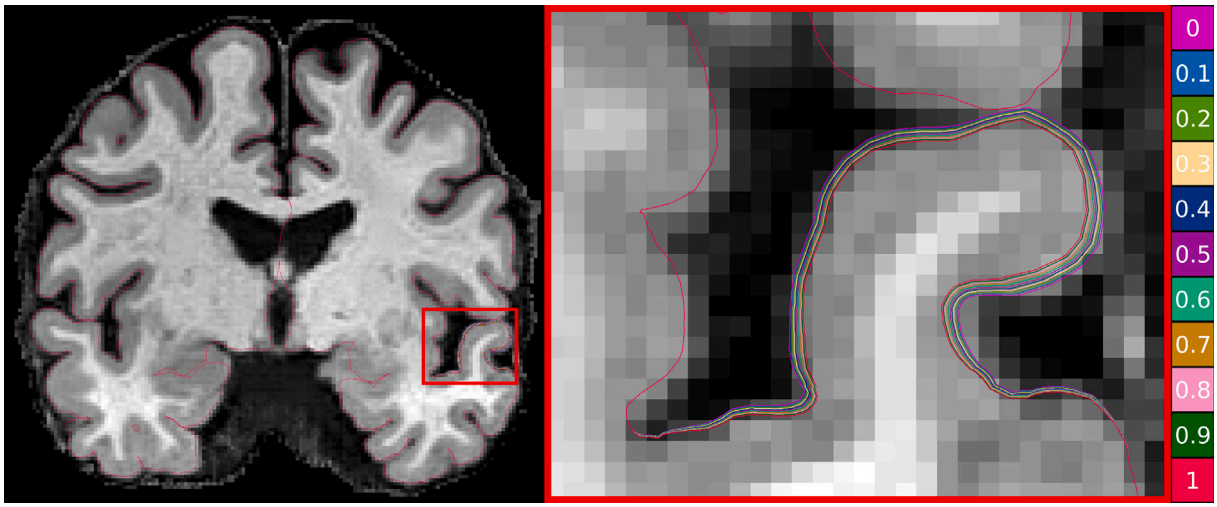

Fig. 4. Local Atrophy: baseline GM meshes and corresponding GM meshes with atrophied superior temporal gyrus at ten atrophy levels (in the range [0.1, 1] mm with 0.1 mm step between the time points) overlayed on top of a synthetic MRI. The meshes are colour-coded with the legend provided on the right.

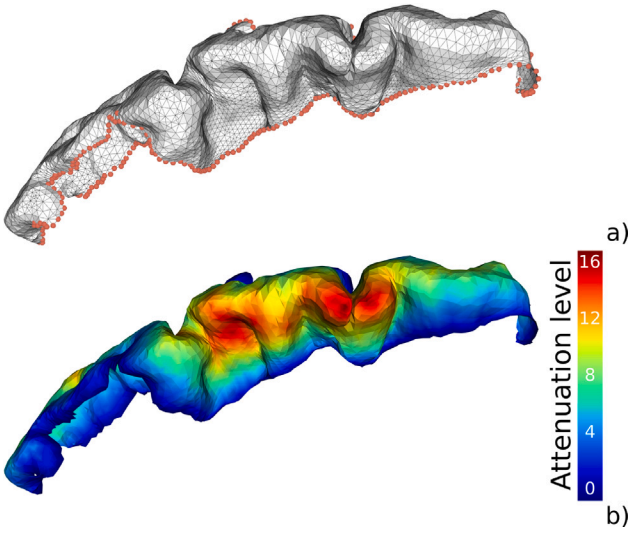

Fig. 5. The mesh of superior temporal gyrus with orange dots denoting fixed vertices that construct its boundary (a) and gradual attenuation coefficient mapped on the mesh with colour-coded graduality levels legend (b).

firstly proposed in (Rusak et al., 2020), which steers training towards learning the gross brain anatomy, while ignoring HF detail. The training continues with a ResNet-based discriminator, detailed in (Rusak et al., 2021), which focuses on learning HF-detail. When trained properly, HF-GAN can generate synthetic brain MRI scans with HF detail that resembles the corresponding real MRI scans (ground truth) better than synthetic brain MRI scans without HF detail, in terms of full-reference image quality metrics (IQMs) (Rusak et al., 2021).

We further trained the pre-trained HF-GAN model on a training set consisting of 596 pairs (which corresponds to 60% of ADNI subset) of brain MRIs and PV-maps derived from FreeSurfer meshes. We resumed the training with a learning rate of 3.96×10^{-5} and linearly decreased it every epoch to reach zero over 200 epochs. Our stopping criteria was defined as generator loss plateaus for at least ten epochs with fluctuations less than 0.01. Examples of synthetic MRIs with several levels of atrophy are illustrated in Fig. 6.

3.6. Synthetic dataset

We created a synthetic dataset that consists of 400 brain MRI scans. For the generation of the synthetic dataset, we randomly selected 20 HC baseline subjects from the test set (ADNI dataset), ten females and ten males. We selected HC instead of MCI or AD subjects for synthetic dataset generation to minimise the chances that the introduced atrophy (up to 1 mm) could exceed the minimum CTh of each baseline scan. The mean age of ADNI subjects selected for the generation of the synthetic dataset is 70.6 years, with subject age ranging in the [59.1, 78.3] age interval. Each of the ADNI baseline subjects was synthesised with no introduced atrophy to obtain synthetic baseline subjects. Additional 19 different atrophy levels were synthesised per subject. The atrophy levels fall into two ranges, [0.01, 0.1) mm, with the atrophy step rate of 0.01 mm (nine levels) and [0.1, 1] mm, with the atrophy step rate of 0.1 mm (ten levels). Our synthetic dataset is publicly available[4] on the CSIRO data access portal.

3.7. Cortical thickness estimation methods under test

We evaluate four different CTh estimation methods on our synthetic dataset:

FreeSurfer Cross-sectional pipeline is the standard processing stream of the surface-based medical image analysis suite (Fischl, 2012). The pipeline consists of several steps for GM and WM surfaces reconstruction and mapping of morphometric measurements on the reconstructed surface. FreeSurfer estimates CTh as an average of the minimal distance between a point on the GM surface mesh to the WM surface mesh and the minimal distance between the corresponding point on the WM surface mesh to the GM surface (Fischl and Dale, 2000; Han et al., 2006). In our experiments we used software version 6.0.1.

FreeSurfer Longitudinal (2 TPs) pipeline aims to reduce inter-subject variability and increase re-scanning measurements reproducibility which are usually affected by hardware-related factors. The FreeSurfer longitudinal stream employs inverse consistent registration to create an unbiased inter-subject template (Reuter et al., 2010). The longitudinal pipeline uses the inter-subject template to initialise processing steps such as skull stripping, Talairach transforms, atlas registration, spherical surface maps and parcellations with the template information to increase the reliability and statistical power (Reuter et al., 2012). In the longitudinal pipeline, CTh measurements

[4] https://doi.org/10.25919/4ycc-fc11

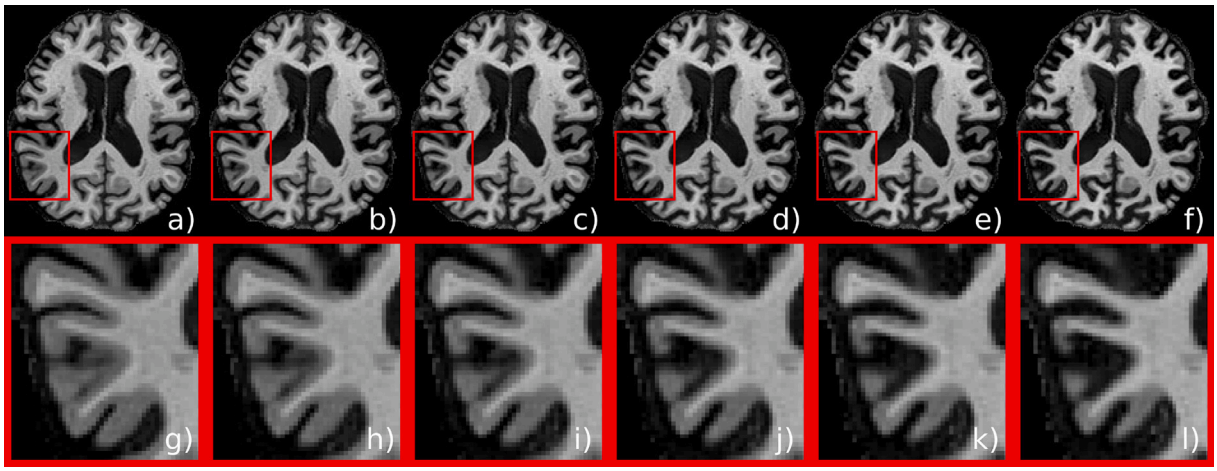

Fig. 6. Synthetic brain MRIs with uniformly introduced atrophy at different levels: (a) 0 mm, (b) 0.2 mm, (c) 0.4 mm, (d) 0.6 mm, (e) 0.8 mm and (f) 1 mm. The corresponding zoomed areas are shown in (g), (h), (i), (j), (k)and (l), respectively.

are performed in the same way as in the cross-sectional pipeline. In our experiments we used the longitudinal pipeline from software version 6.0.1. Each pair of baseline-n^{th} atrophy level, two time points (2TPs), were computed independently of each other.

DL+DiReCT[5] combines deep learning (DL)-based neuroanatomy segmentations and diffeomorphic registration-based CTh (DiReCT) measurements to achieve accurate and reliable CTh estimations quickly. The authors employed a DL model called DeepSCAN (McKinley et al., 2021) to derive GM and WM segmentation as well as parcellation and DiReCT (Das et al., 2009) to obtain CTh from T1-w brain MRI (Rebsamen et al., 2020a). The CTh is defined as a distance measure between corresponding CSF/GM and GM/WM interfaces, where continuous one-to-one mapping is ensured by diffeomorphic registration (Das et al., 2009; Rebsamen et al., 2020a). The authors used scans from Adolescent Brain Cognitive Development (ABCD) (Casey et al., 2018), IXI,[6] ADNI (((Jack Jr. et al., 2008)) and a publicly unavailable dataset to train DeepSCAN with labels derived by FreeSurfer 6.0.

HerstonNet is a neural network regression model for brain morphometry (Santa Cruz et al., 2021). HerstonNet estimates volume, CTh and curvature directly from T1-w brain MRI scans. The ResNet-derived architecture selected for the implementation of HerstonNet deepens the model which improves the estimation performance, with limited execution overhead. HerstonNet was trained on a subset of ADNI (Jack Jr. et al., 2008) and the Australian Imaging, Biomarkers and Lifestyle (AIBL) (Rowe et al., 2010) dataset where morphometric measurements were computed with FreeSurfer 6.0. pipeline (Fischl et al., 1999). Since HerstonNet is trained on a dataset paired with FreeSurfer morphometric measurements, HerstonNet implicitly computes CTh using the FreeSurfer definition.

4. Experiments

4.1. Synthetic image quality evaluation

To evaluate the quality of synthetic scans, we derived PV-maps of three tissue classes (GM, WM and CSF) from the test set (206 subjects) as described in Section 3.2. Then, we used the derived PV-maps to generate corresponding synthetic brain MRI scans using the GAN model described in Section 3.5. Once we obtained the synthetic MRIs that correspond to the real MRIs in the test set, we evaluated the quality of synthesised MRIs in brain ROI only by computing three full-reference IQMs: Structural Similarity Index Measure (SSIM), Normalised Root Mean Square Error (NRMSE) and Peak Signal-to-Noise Ratio (PSNR).

$$MSE = \frac{1}{mnk}\sum_{x=1}^{m}\sum_{y=1}^{n}\sum_{z=1}^{k}\left(I_R(x,y,z) - I_S(x,y,z)\right)^2 \tag{5}$$

$$NRMSE = \frac{\sqrt{MSE}}{\sigma_{I_R}} \tag{6}$$

$$PSNR = \frac{MAX_{I_R}^2}{MSE} \tag{7}$$

$$SSIM = \frac{(2\mu_{I_R}\mu_{I_S} + c_1)(2\sigma_{I_R I_S} + c_2)}{(\mu_{I_R}^2 + \mu_{I_S}^2 + c_1)(\sigma_{I_R}^2 + \sigma_{I_S}^2 + c_2)} \tag{8}$$

where I_R is the real image of dimension (m,n,k), I_S is the synthetic image, σ_{I_R} is the standard deviation computed across all real samples in the test set, MAX_{I_R} is the real image maximum value in the test set, μ_{I_R} is the mean value of real image, μ_{I_S} is the mean value of synthetic image, $\sigma_{I_R}^2$ is the variance of I_R, $\sigma_{I_S}^2$ is the variance of I_S and $\sigma_{I_R I_S}$ is the co-variance of I_R and I_S. Variables $c_1 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$ stabilise the division with weak denominator, where L stands for a dynamic range, $k_1 = 0.01$ and $k_2 = 0.03$.

Further, we obtained PV-maps (GM, WM and CSF) from both real and synthetic MRIs by first segmenting the images using an EM-algorithm (Van Leemput et al., 1999b) followed by PV-maps estimation algorithm described in (Acosta et al., 2009). We compared the PV-maps derived from real MRIs to the PV-maps derived from synthetic MRIs by computing NRMSE between corresponding PV-maps for each class (GM, WM and CSF). While in this experiment, we utilised the auxiliary PV-maps, derived from real and corresponding synthetic MRI scans to further evaluate the image quality, in the following experiments, brain structures extracted from auxiliary PV-maps (Fig. 1) are not taken into account by the methods under test.

4.2. Evaluation of introduced atrophy level

The difference between synthetic MRIs with and without introduced atrophy was used to qualitatively check that the synthetic MRIs reflect the introduced atrophy. First, we generated synthetic MRIs with nine different levels of uniform cortical atrophy in the range [0.01, 0.1) mm with the atrophy step rate of 0.01 mm. Then, we subtracted synthetic time point scans (atrophy > 0 mm) from the synthesised baseline scan (atrophy = 0 mm). Visualising the difference between baseline and synthetic time points at various atrophy levels helps us in the evaluation of GAN's ability to preserve introduced atrophy.

[5] Model taken from: https://github.com/SCAN-NRAD/DL-DiReCT
[6] http://brain-development.org/ixi-dataset

Table 1
Image quality assessment by full-reference IQM and segmentation/PV-estimation error of three tissue classes. The character ↑ denotes that the higher metric values indicate better results, while the character ↓ denotes that lower metric values indicate better results.

SSIM ↑	NRMSE ↓	PSNR ↑	NRMSE ↓		
			GM	WM	CSF
0.9451 ± 0.0146	0.0133 ± 0.0955	36.82 ± 1.6286	0.0111 ± 0.0054	0.0073 ± 0.005	0.0043 ± 0.0029

4.3. Cortical thickness estimation methods benchmark

To evaluate the performance of the four methods on our synthetic dataset (20 subjects), we measured sensitivity to atrophy changes across 19 atrophy levels per subject, nine in the range of [0.01, 0.1) mm with the step of 0.01 mm and ten in the range of [0.1, 1] mm with the step of 0.1 mm. For each method under test, we measured the average detected CTh per atrophy level across all brain regions and plotted the results in relation to the introduced atrophy (19 atrophy levels in the [0.01, 1] mm range). Additionally, we plotted ten atrophy levels in the range of [0.01, 0.1] mm separately to better represent the methods' ability to recover very small atrophy changes.

We also evaluated each method's performance region-wise, where we consider 34 ROIs per hemisphere defined in Desikan–Killiany atlas (Desikan et al., 2006). For every region we computed R^2 across 19 atrophy levels, nine in the range of [0.01, 0.1) mm with the step of 0.01 mm and ten in the range of [0.1, 1] mm with the step of 0.1 mm. The resulting R^2 values were mapped on a template surface brain using its Desikan–Killiany parcellation. The R^2 was computed according to the following definition:

$$y_{i,j} = BL_i - TP_{i,j} \tag{9}$$

$$\bar{y} = \frac{1}{s \times a} \sum_{i=1}^{s} \sum_{j=1}^{a} y_{i,j} \tag{10}$$

$$SS_{res} = \sum_{i} \sum_{j} (y_{i,j} - A_j)^2 \tag{11}$$

$$SS_{tot} = \sum_{i} \sum_{j} (y_{i,j} - \bar{y})^2 \tag{12}$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{13}$$

where s denotes the number of subjects, a number of atrophy levels. The BL_i stands for the mean CTh measured in a particular region of a synthetic baseline MRI (atrophy level = 0), $TP_{i,j}$ stands for the mean CTh measured in a particular region of a synthetic time point MRI (atrophy level > 0) and A_j stands for the introduced atrophy level. We denote the sum of squares of residuals between introduced and detected atrophy with SS_{res} and total sum of squares with SS_{tot}.

We selected R^2 as a suitable metric for CTh methods comparison since computing the mean absolute deviation between measured and introduced atrophy would not account for different CTh definitions (both in the synthesis method and the CTh estimation methods) which could introduce systematic bias. R^2 does not determine whether predictions are biased. Instead, it determines the goodness of linear fit between the ground truth and the predicted values. Since we aim to benchmark several CTh estimation methods with various CTh definitions, R^2 allows us to make a fair comparison where CTh definition related bias is excluded from the comparison. The R^2 is a suitable metric based on the assumption that the difference in CTh caused by using a different CTh definition can be modelled linearly.

We further focused on R^2 values computed in both introduced atrophy ranges, [0.01, 0.1] and [0.01, 1] mm for methods under test in several regions (parahippocampal gyrus, inferior, middle, superior, transverse temporal gyrus, posterior cingulate and temporal pole), in both hemispheres, relevant to early onset of AD.

4.4. Power analysis

To ensure the reliability of the benchmarking results (the experiment described in Section 4.3), we performed a power analysis. In this context, power is the likelihood that a CTh estimation method under test is sensitive to introduced atrophy level (difference between baseline and corresponding time point MRI), given that the atrophy level is present in the synthetic MRIs. In other words, we performed a single sample t-tests on the thickness difference by checking the null hypothesis (measured CTh difference does not equal zero). Thus we determined the minimal amount of the introduced atrophy (region-wise) that can be detected by a method under test for the specified power, significance level and sample size. For every method under test in Section 4.3, every region and two introduced atrophy ranges, [0.01, 0.1] mm and [0.01, 1] mm, we computed the minimal measurable atrophy level for a sample size of five, statistical power of 95% and significance level of 0.05, accounting for two tails. I.e. we measured the lowest atrophy level that a method under test can detect in every region defined in Desikan–Killiany atlas with five samples. To compute the minimal detectable atrophy level that fits the criteria mentioned above, we used G*Power 3.1 software package (Faul et al., 2009). We further mapped the computed atrophy levels on a template surface brain for both atrophy ranges.

4.5. Local atrophy evaluation

To illustrate the ability of our method to introduce localised atrophy (described in Section 3.4.), we generated synthetic baseline subjects and MRIs with atrophied superior temporal gyrus (left hemisphere only). For each subject in the test set, we synthesised ten MRIs with localised atrophy levels ranging between [0.1, 1] mm with the step of 0.1 mm. Then, we processed synthesised brain MRIs with FreeSurfer cross-sectional pipeline and measured evaluated atrophy. All surfaces were registered to a template using FreeSurfer surface-based group analysis pipeline, followed by a paired t-test on each vertex of reconstructed surfaces (baseline and corresponding time points).

5. Results

5.1. Synthetic brain MRI quality

The results of IQM computed to assess the image quality of synthetic MRI scans are presented in Table 1. According to Table 1, the mean SSIM index reaches 0.9451 on a [0, 1] range, NRMSE achieves 0.0133 (the NRMSE lower bound equals 0), and PSNR measures 36.82 dB where the higher PSNR values indicate an image of higher quality. Similarly, the mean NRMSE computed between PV-maps of GM, WM, and CSF derived from real and synthetic MRIs equals 0.0111, 0.0073 and 0.0043, respectively, where the lower bound equals 0. The obtained results are comparable to synthetic image quality measurements reported in our previous work (Rusak et al., 2021), where synthetic MRIs were generated by an HF-GAN. This result shows that the retrained HF-GAN is suitable for the purpose of generating synthetic brain MRIs with introduced atrophy.

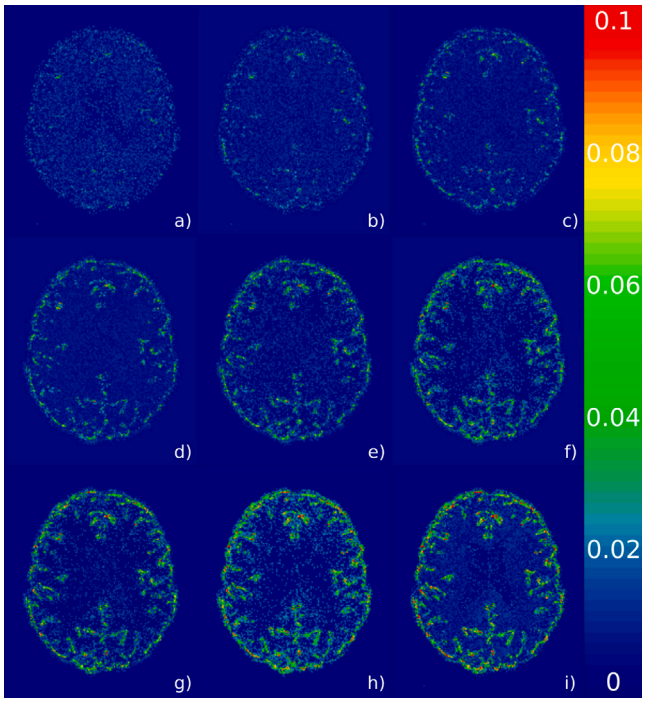

Fig. 7. Difference images between synthetic baseline (atrophy level = 0 mm) and synthesised time points (atrophy level > 0 mm) at different levels of atrophy: (a) 0.01 mm, (b) 0.02 mm, (c) 0.03 mm, (d) 0.04 mm, (e) 0.05 mm, (f) 0.06 mm, (g) 0.07 mm, (h) 0.08 mm and (i) 0.09 mm.

5.2. Introduced atrophy level visualisation

The results obtained from the evaluation of the introduced atrophy levels are presented in Fig. 7. According to the presented single-subject MRI difference (synthetic time points subtracted from the baseline), the image difference representation starts to form a brain contour at the introduced atrophy level of 0.03 mm. As expected, as the introduced atrophy level increases, the brain contour appears to be more defined with differences of higher magnitude.

5.3. Cortical thickness estimation methods benchmark results

The results of the experiment that measures sensitivity to atrophy changes (average detected CTh per atrophy level), across 20 subjects and 19 atrophy levels, in the range of [0.01, 0.1] mm are presented in Fig. 8, and in the range of [0.01, 1] mm are presented in Fig. 9.

DL+DiReCT recovers the introduced atrophy levels in the [0.01, 1] mm atrophy range (Fig. 9) and its subrange [0.01, 0.1] mm (Fig. 8) better than FreeSurfer pipelines and HerstonNet. However, towards the upper bound of the introduced [0.01, 1] mm atrophy range, the DL+DiReCT atrophy measurements tend to be less accurate (greater mismatch between expected and DL+DiReCT measurement trend). The FreeSurfer cross-sectional pipeline atrophy measurements appear to follow the expected trend (denoted with a red x=y line) better than the FreeSurfer longitudinal (2 TPs) pipeline in both atrophy ranges (Figs. 8 and 9). In the [0.01, 0.1] mm atrophy subrange, the observed difference between FreeSurfer cross-sectional and longitudinal trends of recovering introduced atrophy is reflected in less variable FreeSurfer cross-sectional measurements. The FreeSurfer cross-sectional pipeline also detects introduced atrophy better than HerstonNet when the entire atrophy range is considered (Fig. 9). When comparing FreeSurfer cross-sectional pipeline and HerstonNet on the atrophy subrange only (Fig. 8), HerstonNet better recovers the introduced atrophy. In both atrophy ranges, [0.01, 0.1] and [0.01, 1] mm, HerstonNet measures the introduced atrophy better than the FreeSurfer longitudinal (2 TPs) pipeline. However, on the entire atrophy range (Fig. 9), both FreeSurfer pipelines and HerstonNet tend to underestimate the introduced atrophy, e.g. for uniform atrophy of 1 mm FreeSurfer cross-sectional pipeline measures the atrophy level of around 0.4 mm.

The obtained region-wise performance results (R^2) for methods under test, across 34 ROIs per hemisphere, are presented in Fig. 10. The numerical R^2 values are available in Supplementary materials, Table S1 for the range [0.01, 0.1] mm and Table S2 for the range [0.01, 1] mm.

Overall, DL+DiReCT achieves higher R^2 scores, compared to the other methods, in almost all brain regions for the atrophy range of [0.01, 0.1] mm (Fig. 10 - left) and in the vast majority of brain regions for the atrophy range of [0.01, 1] mm (Fig. 10 - right). DL+DiReCT and HerstonNet perform equally well in several brain regions for the [0.01, 1] mm atrophy range, e.g. superior parietal lobule (both hemispheres), while HerstonNet performs slightly better than DL+DiReCT for the same atrophy range in, e.g. frontal pole (LH) and inferior temporal cortex (LH). HerstonNet achieves higher R^2 than both FreeSurfer pipelines in all brain regions for the [0.01, 0.1] mm atrophy range. Conversely, for the [0.01, 1] mm atrophy range, on average, HerstonNet achieves lower R^2 than both FreeSurfer pipelines. On average, FreeSurfer cross-sectional achieves higher R^2 than the longitudinal (2 TPs) pipeline in most brain regions for both intervals (Fig. 10). The difference in R^2 between FreeSurfer pipelines is greater in the [0.01, 0.1] than [0.01, 1] mm atrophy range.

The region-wise performance results (R^2) in the regions relevant to the early-onset of AD are presented in Table 2 for the [0.01, 0.1] mm atrophy range and Table 3 for the [0.01, 1] mm atrophy range.

When considering only ROIs related to typical early AD atrophy, DL+DiReCT achieved the highest R^2 in all considered ROIs for both intervals (Tables 2 and 3). In the [0.01, 0.1] mm atrophy range (Table 2), HerstonNet achieves higher R^2 than both FreeSurfer pipelines in all brain regions, while FreeSurfer cross-sectional achieves higher R^2 than the longitudinal pipeline across brain regions. Conversely, in the [0.01, 1] mm atrophy range (Table 3), on average, FreeSurfer pipelines achieve higher R^2 than HerstonNet, even though that HerstonNet, together with DL+DiReCT, achieve the highest R^2 in two regions, inferior (LH) and middle temporal gyrus (LH). In the [0.01, 1] mm atrophy range (Table 3), FreeSurfer cross-sectional achieves R^2 higher than longitudinal (2 TPs) pipeline in all ROIs except temporal pole (both hemispheres).

5.4. Minimal detected atrophy level for five samples

The results of the power analysis are presented in Fig. 11. The numerical values of minimal detected atrophy levels are available in Supplementary materials, Table S3 for the range [0.01, 0.1] mm and Table S4 for the range [0.01, 1] mm. According to Fig. 11, DL+DiReCT can detect smaller changes in CTh than both FreeSurfer pipelines and HerstonNet in parietal, occipital lobe, insula and most regions of the frontal lobe, in both atrophy ranges [0.01, 0.1] mm and [0.01, 1] mm. In the caudal middle frontal, precentral gyrus, pars opercularis (left hemisphere), and frontal pole (right hemisphere), DL+DiReCT and HerstonNet can detect similar levels of atrophy. DL+DiReCT showed higher sensitivity to detect smaller levels of atrophy than other tested methods in all regions of the temporal lobe except the transverse temporal and parahippocampal gyrus. HerstonNet detected a lower level of atrophy than DL+DiReCT in the transverse temporal gyrus region (both hemispheres), while FreeSurfer cross-sectional pipeline and DL+DiReCT detected the same level of atrophy in the parahippocampal gyrus.

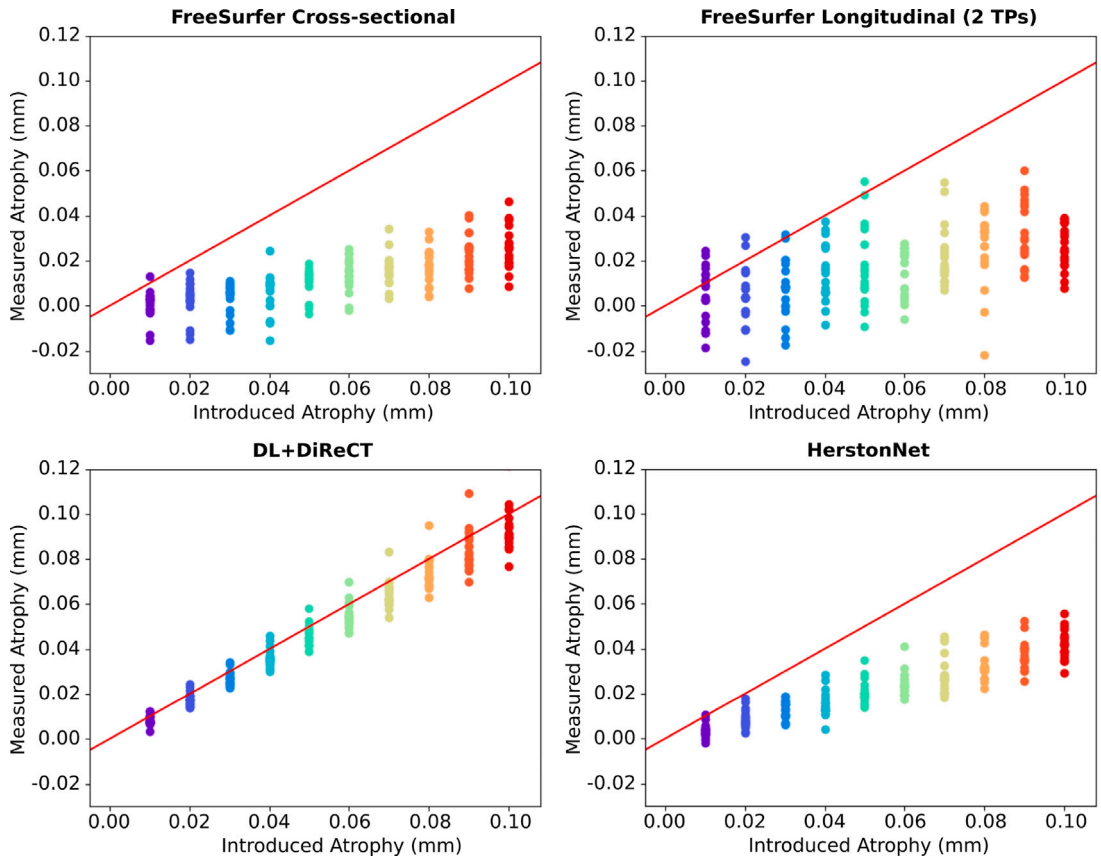

Fig. 8. Introduced vs measured uniform atrophy, measured by FreeSurfer Cross-sectional, FreeSurfer Longitudinal (2 TPs), DL+DiReCT and HerstonNet, respectively, in the [0.01, 0.1] mm range with 0.01 mm step between the atrophy levels. The red line (x=y) indicates the expected trend.

Table 2
R^2 values for the seven brain regions with cortical atrophy present from the early stages of Alzheimer's disease. The R^2 values are computed for regions on both, left (LH) and right (RH) hemispheres in the [0.01, 0.1] mm range. The bold values denote the highest measured R^2 value region and hemisphere-wise.

Brain regions (R^2)	FreeSurfer Cross-sectional	FreeSurfer Longitudinal (2 TP)	DL+DiReCT	HerstonNet
Parahippocampal — LH	0.28	0.24	**0.9**	0.45
Parahippocampal — RH	0.29	0.19	**0.95**	0.3
Posterior cingulate — LH	0.4	0.21	**0.94**	0.49
Posterior cingulate — RH	0.26	0.17	**0.94**	0.46
Inferior temporal — LH	0.52	0.27	**1**	0.93
Inferior temporal — RH	0.49	0.31	**0.99**	0.91
Middle temporal — LH	0.51	0.33	**1**	0.85
Middle temporal — RH	0.69	0.3	**0.99**	0.94
Superior temporal — LH	0.69	0.36	**1**	0.93
Superior temporal — RH	0.74	0.31	**0.99**	0.94
Transverse temporal — LH	0.5	0.33	**0.95**	0.94
Transverse temporal — RH	0.46	0.35	**0.97**	0.93
Temporal pole — LH	0.2	0.23	**0.98**	0.76
Temporal pole — RH	0.19	0.34	**0.95**	0.77

5.5. Evaluation of localised atrophy

The t-maps as results of the per-vertex paired t-test of reconstructed surfaces (baseline and corresponding time points) are shown in Fig. 12. It shows that FreeSurfer can detect local atrophy introduced in the superior temporal gyrus (left hemisphere). The darker shades of red indicate the detection of higher atrophy levels. The red indicator fades away at the edges of the superior temporal gyrus region, where the changes were introduced. The reason for that is the uneven introduction of atrophy across the superior temporal gyrus region since the attenuation function (Eq. (4)) was used to fix the boundary of the ROI when introducing atrophy.

6. Discussion

This paper presents a method for generating synthetic brain MRIs with known CTh changes (global and local) of multiple levels. The proposed method introduces changes in CTh by displacing pial towards WM surface mesh, vertex-by-vertex, deriving PV-maps from the displaced mesh and using HF-GAN to synthesise realistic-looking brain MRIs with known atrophy changes. We show that the synthetic dataset, generated by the proposed method, enables region-wise inter-and intra-method performance comparison, which provides a more-complete evaluation of different CTh estimation methods. We illustrated the benefits of CTh estimation methods evaluation on our synthetic dataset

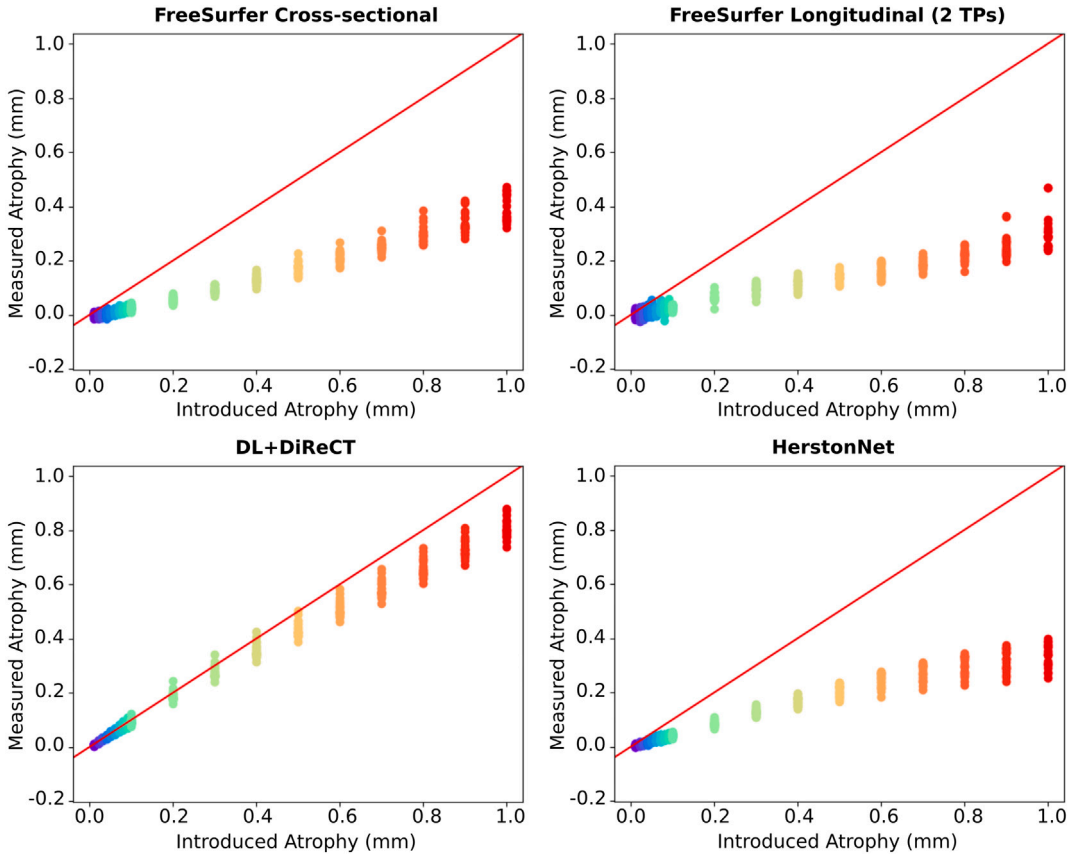

Fig. 9. Introduced vs measured uniform atrophy, measured by FreeSurfer Cross-sectional, FreeSurfer Longitudinal (2 TPs), DL+DiReCT and HerstonNet, respectively, in the [0.01, 1] mm range with 0.01 mm step between the atrophy levels in the [0.01, 0.1) and 0.1 mm in the [0.1, 1] mm subrange. The red line (x=y) indicates the expected trend.

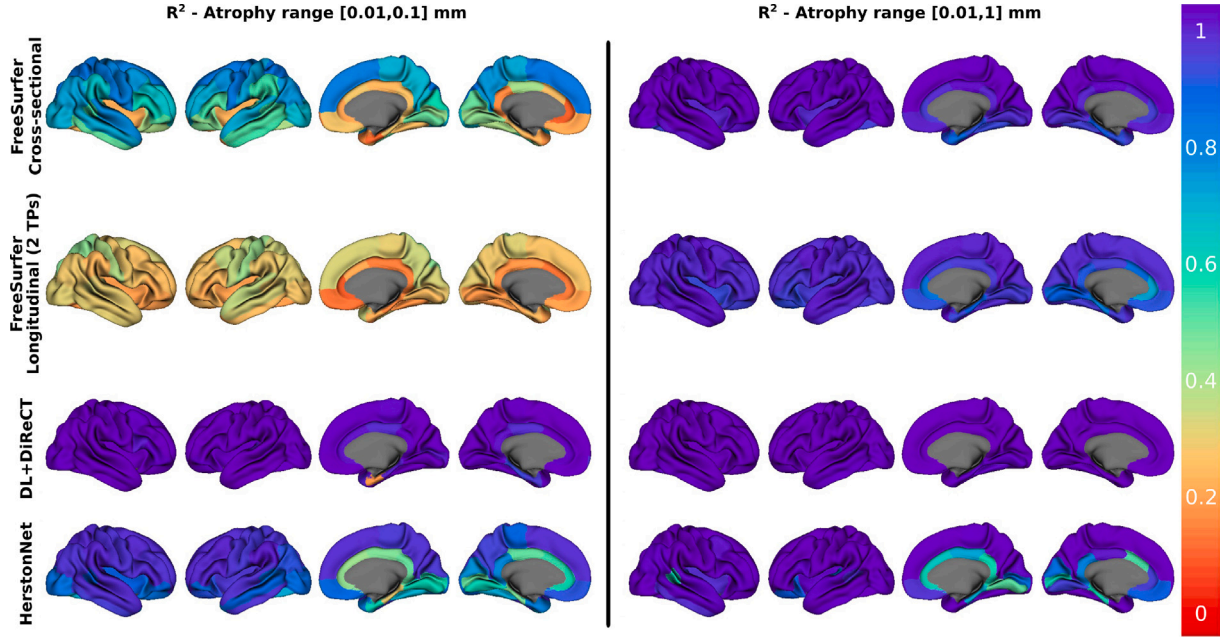

Fig. 10. R^2 computed per region and mapped on the template mesh. R^2 was computed on two atrophy intervals, [0.01, 0.1] mm (left) and [0.01, 1] mm (right).

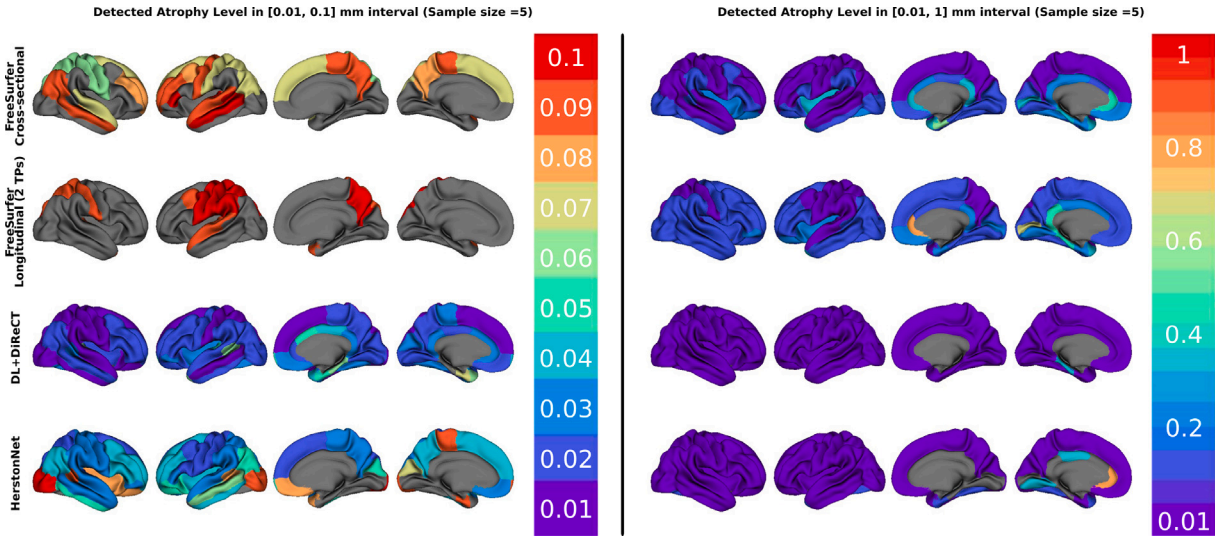

Fig. 11. Atrophy level computed by performing power analysis with the power of 0.95, significance of 0.05 and sample size of five. Atrophy levels were computed on two atrophy intervals, [0.01, 0.1] mm (left) and [0.01, 1] mm (right) and mapped on a template surface brain.

Table 3
R^2 values for the seven brain regions with cortical atrophy present from the early stages of Alzheimer's disease. The R^2 values are computed for regions on both, left (LH) and right (RH) hemispheres in the [0.01, 1] mm range. The bold values denote the highest measured R^2 value region and hemisphere-wise.

Brain regions (R^2)	FreeSurfer Cross-sectional	FreeSurfer Longitudinal (2 TP)	DL+DiReCT	HerstonNet
Parahippocampal — LH	0.87	0.77	**0.99**	0.56
Parahippocampal — RH	0.88	0.85	**0.97**	0.68
Posterior cingulate — LH	0.98	0.91	**1**	0.96
Posterior cingulate — RH	0.97	0.92	**1**	0.69
Inferior temporal — LH	0.97	0.94	**0.98**	**0.98**
Inferior temporal — RH	0.97	0.96	**0.99**	0.94
Middle temporal — LH	0.98	0.96	**0.99**	**0.99**
Middle temporal — RH	0.99	0.96	**1**	0.98
Superior temporal — LH	0.99	0.96	**1**	0.98
Superior temporal — RH	0.99	0.96	**1**	0.97
Transverse temporal — LH	0.98	0.94	**1**	0.77
Transverse temporal — RH	0.98	0.96	**1**	0.9
Temporal pole — LH	0.91	0.94	**1**	0.99
Temporal pole — RH	0.88	0.96	**1**	0.99

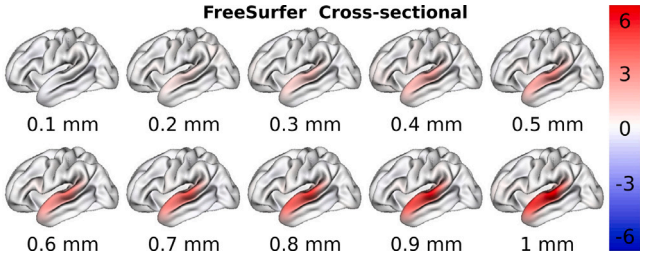

Fig. 12. T-map from the per-vertex paired t-test, between the baseline and time points (atrophied superior temporal gyrus on left hemisphere), mapped on a template surface brain.

by benchmarking FreeSurfer pipelines (cross-sectional and longitudinal), DL+DiReCT and HerstonNet. Further, we computed region-wise minimal detected atrophy level on a synthetic dataset and compared the results among the previously-mentioned methods. Finally, we evaluated and compared the detection results of locally synthesised atrophy.

6.1. Cortical thickness estimation methods benchmark

DL+DiReCT achieved the best performance overall, in terms of introduced atrophy detection, among the other methods under test on both atrophy ranges [0.01, 0.1] mm and [0.01, 1] mm, showing the best correlation and least amount of bias compared to the introduced atrophy level. The ability of DL+DiReCT to recover introduced changes is a strong indicator that displacements made on the pial surface mesh are preserved in the synthetic MRIs. While DL+DiReCT shows overall limited bias in the [0.01,0.6] mm range, the bias tends to increase above 0.6 mm. This happens due to the CTh of some regions being smaller than the introduced level of atrophy. Since some brain regions are thinner than 1 mm, the distance between corresponding pial and WM vertices get capped at 0 mm when the introduced atrophy exceeds the existing distance. Consequently, the expected level of atrophy introduced in these regions is overestimated and manifests as the increased bias in CTh measurements for a particular atrophy subrange. These results are in agreement with the work of Rebsamen et al. (2020a), where DL+DiReCT was found to be more sensitive to changes in CTh than FreeSurfer.

A possible reason for the worse FreeSurfer cross-sectional performance versus DL+DiReCT in detecting introduced atrophy is the higher similarity of the CTh definition used in DL+DiReCT to the CTh definition used in our method compared to the CTh definition used in FreeSurfer. As stated in Section 3.7, FreeSurfer estimates CTh as an average minimum distance between points on the GM and the WM surface mesh, while DL+DiReCT estimates CTh as a distance between corresponding points on CSF/GM and GM/WM interfaces. Since our

method simulates quantifiable atrophy by deforming a vertex on the pial surface towards the corresponding vertex on the WM surface mesh, the CTh measure between the two corresponding vertices resembles more the DL+DiReCT than the FreeSurfer CTh definition. The difference in the CTh definitions could also explain why FreeSurfer (and by extension HerstonNet, since it was trained on the FreeSurfer definition of thickness) tend to underestimate the introduced atrophy.

The results between FreeSurfer cross-sectional and longitudinal (2 TPs) pipeline, (Figs. 8–10), indicate that the cross-sectional pipeline can better detect atrophy changes in our synthetic dataset. Such an observation is unexpected since the aim of the longitudinal pipeline is to decrease the variability and increase the accuracy of longitudinal CTh measurements. FreeSurfer longitudinal (2 TPs) pipeline is run on each pair of data independently, where the CTh of the baseline scan is computed for each data pair, i.e., 19 times per subject. In contrast, in the case of the FreeSurfer cross-sectional pipeline, we compute the CTh of the baseline scan only once per subject. This difference in testing frameworks of these two methods may create a source of uncertainty in the comparison. To make a more fair FreeSurfer pipeline comparison and to avoid a biased evaluation (due to baseline CTh variability), we also ran the FreeSurfer longitudinal pipeline on both the baseline and all synthetic time points (19 atrophy levels) together. Consequently, we ensured that the CTh of the baseline scan was computed only once while running FreeSurfer longitudinal pipeline (20 TPs). We then compared the obtained results with cross-sectional and longitudinal (2 TPs) pipelines. The obtained results show a similar bias of measured atrophy for both the FreeSurfer cross-sectional and longitudinal pipelines for both atrophy ranges, [0.01, 0.1] mm (Figure S2) and [0.01, 1] mm (Figure S3). However, in the [0.01, 0.1] mm atrophy range, FreeSurfer longitudinal (20 TPs) exhibits a slightly smaller spread of measured atrophy (Figure S2) and higher R^2 (Figure S4 — left and Table S1) than the longitudinal pipeline (2 TPs), but a larger spread of measured atrophy and lower R^2 than the FreeSurfer cross-sectional pipeline. In the [0.01, 1] mm atrophy range, all three FreeSurfer pipelines achieve high R^2 with minor regional differences. FreeSurfer longitudinal pipeline (20 TPs), on average, achieves slightly higher R^2 than FreeSurfer cross-sectional pipeline in temporal and occipital lobes as well as insula (Figure S4 — right, Table S2). On the other hand, in the [0.01, 1] mm atrophy range, FreeSurfer longitudinal pipeline (20 TPs) achieves moderately lower R^2 than FreeSurfer cross-sectional pipeline in frontal and parietal lobes (Figure S4 — right, Table S2). In both atrophy ranges ([0.01, 0.1] and [0.01, 1] mm), FreeSurfer longitudinal (2 TPs) achieves lower R^2 than FreeSurfer cross-sectional and longitudinal (20 TPs) pipelines. The power analysis results show that for the [0.01, 0.1] mm atrophy range (Figure S5 — left, Table S3), FreeSurfer longitudinal pipeline (20 TPs) managed to detect lower atrophy levels than FreeSurfer longitudinal pipeline (2 TPs) on the right, while it failed to detect any atrophy level on the left hemisphere. In the [0.01, 1] mm atrophy range (Figure S5 — right, Table S4), the longitudinal FreeSurfer pipelines appear to detect similar atrophy levels in most regions, while FreeSurfer cross-sectional detects lower atrophy levels, especially in frontal and parietal lobes. Overall, the obtained results indicate that FreeSurfer cross-sectional does not perform better than the longitudinal pipeline due to our testing framework since FreeSurfer longitudinal (20 TPs) pipeline, as expected, performs better than the longitudinal (2 TPs) but worse than the cross-sectional pipeline. One significant difference between FreeSurfer cross-sectional and longitudinal pipelines is that the cross-sectional pipeline keeps all operations in the image native space, whereas in the longitudinal pipeline, all images are resampled in the same subject-specific space. As we introduce small atrophy changes, these changes might be lost or diluted by the resampling, resulting in the longitudinal (2 TPs) pipeline being less able to detect small atrophy changes compared to the cross-sectional one.

HerstonNet recovers introduced atrophy in the range [0.01, 0.1] mm better than in the range of [0.01, 1] mm. That is reflected in Figs. 8–9 and explains why, in the introduced atrophy range of [0.01, 1] mm, HerstonNet on average achieves an R^2 lower than both FreeSurfer pipelines but recovers smaller atrophy levels better than both FreeSurfer pipelines. Since HerstonNet recovers atrophy levels in the [0.01, 0.1] mm well, as presented in Fig. 11 - left, and as the same recovered atrophy levels are also included in the [0.01, 1] mm atrophy range, presented in Fig. 11 - right, it appears that HerstonNet performs better than FreeSurfer pipelines. However, HerstonNet performs less well compared to the other methods in the [0.01, 1] mm atrophy range when considering the overall performance. That is especially prominent for atrophy greater than 0.1 mm where, on average, HerstonNet achieves R^2 lower than other methods. Such behaviour can be explained by the population used for HerstonNet training, where the training set may not include subjects with atrophy greater than 0.1 mm in particular regions. On the other hand, according to the obtained results, HerstonNet generalises well in the [0.01, 0.1] mm atrophy range since it recovers small atrophy better than FreeSurfer cross-sectional pipeline, even though it was trained on FreeSurfer CTh measurements.

6.2. Minimal detected atrophy level for a given sample size

Based on the results presented in Fig. 11, all methods can be employed to recover sub-millimetre atrophy, with DL+DiReCT showing higher sensitivity to the CTh changes compared to the other methods. Yet, in the [0.01, 0.1] mm range, DL+DiReCT and HerstonNet are more sensitive to CTh changes than FreeSurfer pipelines, where DL+DiReCT shows higher sensitivity than HerstonNet. This experiment illustrates the diversity of information that may be revealed by benchmarking CTh estimation methods on a synthetic dataset generated by the proposed method. Therefore, when conducting a study where the sensitivity of the CTh estimation method is crucial, understanding the method's performance on a target atrophy range can be extremely helpful to select the most suitable CTh estimation method.

6.3. Detection of synthesised local atrophy

The results, as presented in Fig. 12, indicate that the locally introduced changes are preserved in synthetic MRIs and that FreeSurfer cross-sectional pipeline can recover locally induced atrophy in the superior temporal gyrus. Further, the gradual transition from atrophied brain region to non-atrophied regions indicates that vertices, between the brain region centre and boundary, with attenuated atrophy, are preserved in the synthetic images; and recovered by the cross-sectional pipeline. This example shows the potential of our method to be utilised for more realistic cortical atrophy modelling, disease progression or brain aging.

6.4. Clinical relevance of results

The clinical relevance of our method is in the evaluation of well established and emerging CTh estimation methods against ground truth (known difference in CTh between synthesised baseline and time point subjects). Our CTh evaluation framework facilitates the selection of a CTh estimation method suitable for the detection of sub-voxel atrophy levels either locally or globally. The evaluation frameworks such as group separation, test-retest, and metrics computed against 'silver-standard' ground truth (e.g. ICC) do not enable regional inter-and intra-method performance evaluation. In contrast, our method bridges that gap and allows a closer look into methods' performance in ROIs. Future clinical studies may benefit from our method when selecting a CTh estimation method suitable for a certain pathological context (e.g. early detection of AD).

6.5. Limitations and avenues for improvement

Several relevant limitations emerge from this study, such as:

- Cortical thickness definition used for atrophy synthesis

The mismatch between CTh definitions used for data synthesis and CTh estimation methods may lead to measurement bias. In other words, due to the absence of a standardised CTh definition, the quantifiable synthesised atrophy may be indistinguishable with a CTh estimation method. For fair method comparison on the same synthetic test set, data synthesis methods and all CTh estimation methods must use a single CTh definition, which would require reimplementation of methods under test. However, despite the CTh definition mismatch, one could expect a strong correlation between introduced and measured atrophy. While a single definition of CTh was evaluated in this work, the proposed framework is sufficiently flexible to allow evaluating different CTh definitions.

- Synthesised atrophy range

We restricted the introduced atrophy to the sub-voxel range of [0.01, 1] mm. The rationale behind the atrophy level not being higher than 1 mm is that certain cortex regions are thinner than 1 mm. Further, simulation of an atrophy level higher than 1 mm makes synthetic MRIs appear unrealistic. Nevertheless, if needed, our method supports the introduction of higher atrophy levels. The ability to introduce atrophy in the sub-voxel range is of high importance for clinical applications related to early AD detection. Since structural cortical changes caused by AD progression happen slowly and on a smaller (sub-voxel) scale, the atrophy range used in this work covers the atrophy range of interest for early AD detection applications.

- Inability to model ventricles, cerebellum, deep grey matter, hippocampus and amygdala

The FreeSurfer pipeline, which was used to derive the PV-maps through the rasterisation of the surface meshes, does not provide surface meshes for the ventricles, deep grey matter and cerebellum. Therefore, we construct PV-maps by combining these structures extracted from PV-maps created by an EM-algorithm followed by PV-maps estimation with FreeSurfer derived PV-maps. Similarly, since FreeSurfer does not accurately mesh the hippocampus and amygdala,[7] these structures were also derived from the EM PV-maps. As a result, these structures remained constant for all introduced atrophy levels. While the ventricles, deep grey matter and cerebellum are not involved in the computation of CTh, their lack of change relative to the introduced cortical atrophy could potentially introduce bias in deep learning models. These structures might indirectly contribute to the cortex segmentation (such as DL+DiReCT) or the computation of CTh (such as HerstonNet). However, both deep learning approaches outperformed the more traditional Freesurfer pipelines. Therefore, it is unlikely that not introducing atrophy in these structures had a negative impact on the deep learning approaches.

- Inter-CTh estimation methods parcellation variabilities

Anatomical segmentation (parcellation) plays an essential role in information extraction from brain MRIs and is a prerequisite for quantitative image analysis (Heckemann et al., 2010). While all tested CTh estimation methods use Desikan–Killiany atlas for whole-brain parcellation, the parcellation maps may still vary across methods due to parcellation protocols and implementation differences (Mikhael et al., 2018; Popovych et al., 2021). Since we introduce atrophy uniformly across all ROIs, the impact of the induced parcellation protocol-specific bias in the context of per-region inter-and intra-method regional comparison should be minimal.

[7] https://surfer.nmr.mgh.harvard.edu/fswiki/UserContributions/FAQ

- HF-GAN generates skull-stripped images

This method is not suitable for benchmarking CTh estimation methods that require brain MRI scans with a skull, e.g. FastSurfer (Henschel et al., 2020). For this study, we considered FastSurfer and included it in the list of methods under test but obtained poor results. FastSurfer exhibited poor CTh estimation performance mainly because our synthetic dataset is skull stripped. Since, FastSurferCNN, a FastSurfer DL-based component responsible for whole-brain segmentation, is trained on brain MRIs with a skull, the whole brain segmentation of our skull-stripped synthetic MRIs does not give accurate results. We plan to overcome this obstacle, in future work, by training a conditional GAN that generates a skull for a given skull-stripped brain MRI.

- Lack of diversity between appearance of synthetic MRI time points

The differences in the appearance between baseline and time point scans in our synthetic dataset do not reflect all real-world scenarios (e.g. scanner type, movement artefacts). Therefore the proposed method does not account for the ability of CTh estimation methods to deal with movement and bias field artefacts as well as the contrast variability due to the protocol differences. As a result, our dataset offers a best-case scenario where the differences between time points are limited to CTh changes only. However, a conditional GAN could be used to further model these variabilities and offer a dataset with more realistic changes between time points.

- Method evaluation on a single dataset

Within the scope of this study, we developed and evaluated our method on the ADNI dataset only. Nevertheless, our framework can be generalised on other T1-w brain MRI datasets, e.g., The Australian Imaging, Biomarker & Lifestyle Flagship Study of Aging (AIBL) (Ellis et al., 2009). For this purpose, the HF-GAN, presented in Rusak et al. (2021), needs to be trained on the AIBL training subset while synthetic brain MRI scans with quantifiable atrophy can be constructed as in Section 3 by using AIBL instead of ADNI subset. Evaluating our method on multiple datasets would account for various MRI scan appearances, scanner-related contrasts and artefacts. While robustness to diverse appearances is a notable aspect of CTh estimation methods evaluation, the main aim of this work is to control quantifiable atrophy synthesis in the context of evaluating CTh estimation methods.

6.6. Future work

While this work was limited to using our synthetic dataset as a testing platform for existing CTh estimation methods, we aim to evaluate its use as a training set for improving CTh estimation models. The HerstonNet CTh estimation results strongly suggest that brain regions with higher variability in CTh were estimated better than regions with lower CTh variability. Therefore, the investigation into the applicability of synthetic brain MRIs in CTh estimation model training may reveal beneficial insights for CTh estimation precision.

7. Conclusions

In this paper, we proposed a GAN-based method for quantifiable cortex atrophy synthesis. By using the proposed method, we synthesised a test set composed of 20 subjects with 19 quantifiable atrophy levels for each subject. Then, we showed that the synthesised test set is suitable for CTh estimation methods evaluation by per-region (34 regions/hemisphere) benchmarking of four CTh estimation methods: FreeSurfer cross-sectional, FreeSurfer longitudinal (2 TPs), DL+DiReCT and HerstonNet. Moreover, we made our synthetic dataset publically available to encourage and support researchers in the thorough evaluation of their CTh estimation methods.

The purpose of this work is by no means to identify the superiority of any CTh estimation method under test over the others, but rather propose a method that provides a complementary perspective on evaluation and comparison of the CTh estimation methods. Our proposed method may be relevant for researchers developing CTh estimation methods and conducting clinical studies involving CTh measurements. The ability to compare the sensitivity to atrophy of several CTh estimation methods on vertex-level gives the opportunity to gain useful insights into their performance. Further, it facilitates choosing a suitable CTh estimation method for clinical studies (depending on the ROI). According to the obtained results, DL+DiReCT has higher sensitivity to detect introduced atrophy levels compared to the other CTh estimation methods under test in most regions.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Acknowledgement of country

We acknowledge and pay our respect to the Turrbal and Yuggera (Brisbane, Queensland, Australia) Peoples as the Traditional Owners and ongoing custodians of the land and seas on which this research was undertaken. We acknowledge that Aboriginal and Torres Strait Islander Peoples are Australia's first scientists, first educators, and first healers of illness. We acknowledge their elders whom have and continue to pave the path for all Aboriginal and Torres Strait Islander communities of today. We also want to recognise that this land has always been and always will be Aboriginal and Torres Strait Islander land.

Acknowledgement of resources and computing expertise

This project was supported by resources and expertise provided by CSIRO IMT Scientific Computing.

Funding

This work was funded in part through an Australian Department of Industry, Energy and Resources CRC-P project between CSIRO, Maxwell Plus and I-Med Radiology Network.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. 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 organisation is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2022.102576.

References

Acosta, O., Bourgeat, P., Zuluaga, M.A., Fripp, J., Salvado, O., Ourselin, S., Initiative, A.D.N., et al., 2009. Automated voxel-based 3D cortical thickness measurement in a combined Lagrangian–Eulerian PDE approach using partial volume maps. Med. Image Anal. 13 (5), 730–743.

Andersson, J.L., Jenkinson, M., Smith, S., et al., 2007. Non-linear registration, aka spatial normalisation FMRIB technical report TR07ja2. FMRIB Anal. Group Univ. Oxf. 2 (1), e21.

Avants, B.B., Tustison, N., Song, G., et al., 2009. Advanced normalization tools (ANTS). Insight J. 2 (365), 1–35.

Bergouignan, L., Chupin, M., Czechowska, Y., Kinkingnéhun, S., Lemogne, C., Le Bastard, G., Lepage, M., Garnero, L., Colliot, O., Fossati, P., 2009. Can voxel based morphometry, manual segmentation and automated segmentation equally detect hippocampal volume differences in acute depression? Neuroimage 45 (1), 29–37.

Bernal, J., Valverde, S., Kushibar, K., Cabezas, M., Oliver, A., Lladó, X., 2021. Generating longitudinal atrophy evaluation datasets on brain magnetic resonance images using convolutional neural networks and segmentation priors. Neuroinformatics 1–16.

Besag, J., 1986. On the statistical analysis of dirty pictures. J. R. Stat. Soc. Ser. B Stat. Methodol. 48 (3), 259–279.

Burggren, A.C., Zeineh, M., Ekstrom, A.D., Braskie, M.N., Thompson, P.M., Small, G.W., Bookheimer, S.Y., 2008. Reduced cortical thickness in hippocampal subregions among cognitively normal apolipoprotein E e4 carriers. Neuroimage 41 (4), 1177–1183.

Camara, O., Schweiger, M., Scahill, R.I., Crum, W.R., Sneller, B.I., Schnabel, J.A., Ridgway, G.R., Cash, D.M., Hill, D.L., Fox, N.C., 2006. Phenomenological model of diffuse global and regional atrophy using finite-element methods. IEEE Trans. Med. Imaging 25 (11), 1417–1430.

Casey, B., Cannonier, T., Conley, M.I., Cohen, A.O., Barch, D.M., Heitzeg, M.M., Soules, M.E., Teslovich, T., Dellarco, D.V., Garavan, H., et al., 2018. The adolescent brain cognitive development (ABCD) study: Imaging acquisition across 21 sites. Dev. Cogn. Neurosci. 32, 43–54.

Clarkson, M.J., Cardoso, M.J., Ridgway, G.R., Modat, M., Leung, K.K., Rohrer, J.D., Fox, N.C., Ourselin, S., 2011. A comparison of voxel and surface based cortical thickness estimation methods. Neuroimage 57 (3), 856–865.

Das, S.R., Avants, B.B., Grossman, M., Gee, J.C., 2009. Registration based cortical thickness measurement. Neuroimage 45 (3), 867–879.

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., et al., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31 (3), 968–980.

Ellis, K.A., Bush, A.I., Darby, D., De Fazio, D., Foster, J., Hudson, P., Lautenschlager, N.T., Lenzo, N., Martins, R.N., Maruff, P., et al., 2009. The Australian imaging, biomarkers and lifestyle (AIBL) study of aging: Methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. Int. Psychogeriatr. 21 (4), 672–687.

Faul, F., Erdfelder, E., Buchner, A., Lang, A.-G., 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. Behav. Res. Methods 41 (4), 1149–1160.

Fein, G., Di Sclafani, V., Cardenas, V., Goldmann, H., Tolou-Shams, M., Meyerhoff, D.J., 2002. Cortical gray matter loss in treatment-naive alcohol dependent individuals. Alcohol. Clin. Exp. Res. 26 (4), 558–564.

Fischl, B., 2012. Freesurfer. Neuroimage 62 (2), 774–781.

Fischl, B., Dale, A.M., 2000. Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proc. Natl. Acad. Sci. 97 (20), 11050–11055.

Fischl, B., Sereno, M.I., Dale, A.M., 1999. Cortical surface-based analysis: II: Inflation, flattening, and a surface-based coordinate system. Neuroimage 9 (2), 195–207.

Fox, N.C., Schott, J.M., 2004. Imaging cerebral atrophy: Normal ageing to Alzheimer's disease. Lancet 363 (9406), 392–394.

Gavish, B., Graves, S.C., 1978. The Travelling Salesman Problem and Related Problems. Massachusetts Institute of Technology, Operations Research Center.

Geveci, B., Schroeder, W., Brown, A., Wilson, G., 2012. VTK. Architecture Open Source Appl. 1, 387–402.

Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., Busa, E., Pacheco, J., Albert, M., Killiany, R., et al., 2006. Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. Neuroimage 32 (1), 180–194.

Heckemann, R.A., Keihaninejad, S., Aljabar, P., Rueckert, D., Hajnal, J.V., Hammers, A., Initiative, A.D.N., et al., 2010. Improving intersubject image registration using tissue-class information benefits robustness and accuracy of multi-atlas based anatomical segmentation. Neuroimage 51 (1), 221–227.

Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B., Reuter, M., 2020. Fastsurfer-A fast and accurate deep learning based neuroimaging pipeline. NeuroImage 117012.

Hutton, C., De Vita, E., Ashburner, J., Deichmann, R., Turner, R., 2008. Voxel-based cortical thickness measurements in MRI. Neuroimage 40 (4), 1701–1710.

Jack Jr., C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L. Whitwell, J., Ward, C., et al., 2008. The alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J. Magn. Reson. Imaging Off. J. Int. Soc. Magn. Reson. Med. 27 (4), 685–691.

Jernigan, T.L., Archibald, S.L., Fennema-Notestine, C., Gamst, A.C., Stout, J.C., Bonner, J., Hesselink, J.R., 2001. Effects of age on tissues and regions of the cerebrum and cerebellum. Neurobiol. Aging 22 (4), 581–594.

Karaçali, B., Davatzikos, C., 2006. Simulation of tissue atrophy using a topology preserving transformation model. IEEE Trans. Med. Imaging 25 (5), 649–652.

Khanal, B., Ayache, N., Pennec, X., 2016a. Simulating realistic synthetic longitudinal brain MRIs with known volume changes. NeuroImage 12.

Khanal, B., Lorenzi, M., Ayache, N., Pennec, X., 2016b. A biophysical model of brain deformation to simulate and analyze longitudinal MRIs of patients with Alzheimer's disease. NeuroImage 134, 35–52.

Larson, K.E., Oguz, I., 2021. Synthetic atrophy for longitudinal surface-based cortical thickness measurement. In: Medical Imaging 2021: Image Processing, Vol. 11596. International Society for Optics and Photonics, p. 115963K.

McKinley, R., Wepfer, R., Aschwanden, F., Grunder, L., Muri, R., Rummel, C., Verma, R., Weisstanner, C., Reyes, M., Salmen, A., et al., 2021. Simultaneous lesion and brain segmentation in multiple sclerosis using deep neural networks. Sci. Rep. 11 (1), 1–11.

Mikhael, S., Hoogendoorn, C., Valdes-Hernandez, M., Pernet, C., 2018. A critical analysis of neuroanatomical software protocols reveals clinically relevant differences in parcellation schemes. Neuroimage 170, 348–364.

Pieperhoff, P., Südmeyer, M., Hömke, L., Zilles, K., Schnitzler, A., Amunts, K., 2008. Detection of structural changes of the human brain in longitudinally acquired MR images by deformation field morphometry: Methodological analysis, validation and application. NeuroImage 43 (2), 269–287.

Popovych, O.V., Jung, K., Manos, T., Diaz-Pier, S., Hoffstaedter, F., Schreiber, J., Yeo, B.T., Eickhoff, S.B., 2021. Inter-subject and inter-parcellation variability of resting-state whole-brain dynamical modeling. NeuroImage 118201.

Rebsamen, M., Rummel, C., Reyes, M., Wiest, R., McKinley, R., 2020a. Direct cortical thickness estimation using deep learning-based anatomy segmentation and cortex parcellation. Hum. Brain Mapp. 41 (17), 4804–4814.

Rebsamen, M., Suter, Y., Wiest, R., Reyes, M., Rummel, C., 2020b. Brain morphometry estimation: From hours to seconds using deep learning. Front. Neurol. 11, 244.

Reuter, M., Rosas, H.D., Fischl, B., 2010. Highly accurate inverse consistent registration: A robust approach. Neuroimage 53 (4), 1181–1196.

Reuter, M., Schmansky, N.J., Rosas, H.D., Fischl, B., 2012. Within-subject template estimation for unbiased longitudinal image analysis. Neuroimage 61 (4), 1402–1418.

Rowe, C.C., Ellis, K.A., Rimajova, M., Bourgeat, P., Pike, K.E., Jones, G., Fripp, J., Tochon-Danguy, H., Morandeau, L., O'Keefe, G., et al., 2010. Amyloid imaging results from the Australian imaging, biomarkers and lifestyle (AIBL) study of aging. Neurobiol. Aging 31 (8), 1275–1283.

Rusak, F., Cruz, R.S., Smith, E., Fripp, J., Fookes, C., Bourgeat, P., Bradley, A., 2021. Detail matters: High-frequency content for realistic synthetic MRI generation. In: International Workshop on Simulation and Synthesis in Medical Imaging. Springer, pp. 3–13.

Rusak, F., Santa Cruz, R., Bourgeat, P., Fookes, C., Fripp, J., Bradley, A., Salvado, O., 2020. 3D brain MRI GAN-based synthesis conditioned on partial volume maps. In: International Workshop on Simulation and Synthesis in Medical Imaging. Springer, pp. 11–20.

Santa Cruz, R., Lebrat, L., Bourgeat, P., Doré, V., Dowling, J., Fripp, J., Fookes, C., Salvado, O., 2021. Going deeper with brain morphometry using neural networks. In: 2021 IEEE 18th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 711–715.

Sharma, S., Noblet, V., Rousseau, F., Heitz, F., Rumbach, L., Armspach, J.-P., 2010. Evaluation of brain atrophy estimation algorithms using simulated ground-truth data. Med. Image Anal. 14 (3), 373–389.

Shattuck, D.W., Sandor-Leahy, S.R., Schaper, K.A., Rottenberg, D.A., Leahy, R.M., 2001. Magnetic resonance image tissue classification using a partial volume model. NeuroImage 13 (5), 856–876.

Shi, F., Fan, Y., Tang, S., Gilmore, J.H., Lin, W., Shen, D., 2010. Neonatal brain image segmentation in longitudinal MRI studies. Neuroimage 49 (1), 391–400.

Sluimer, J., Vrenken, H., Blankenstein, M., Fox, N., Scheltens, P., Barkhof, F., Van Der Flier, W., 2008. Whole-brain atrophy rate in alzheimer disease: Identifying fast progressors. Neurology 70 (19 Part 2), 1836–1841.

Smith, A.D.C., Crum, W.R., Hill, D.L., Thacker, N.A., Bromiley, P.A., 2003. Biomechanical simulation of atrophy in MR images. In: Medical Imaging 2003: Image Processing, Vol. 5032. International Society for Optics and Photonics, pp. 481–490.

Tohka, J., 2014. Partial volume effect modeling for segmentation and tissue classification of brain magnetic resonance images: A review. World J. Radiology 6 (11), 855.

Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 1999a. Automated model-based bias field correction of MR images of the brain. IEEE Trans. Med. Imaging 18 (10), 885–896.

Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 1999b. Automated model-based tissue classification of MR images of the brain. IEEE Trans. Med. Imaging 18 (10), 897–908.

Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Green, R.C., Harvey, D., Jack Jr., C.R., Jagust, W., Morris, J.C., et al., 2017. The alzheimer's disease neuroimaging initiative 3: Continued innovation for clinical trial improvement. Alzheimer's Dementia 13 (5), 561–571.

Whitwell, J.L., Przybelski, S.A., Weigand, S.D., Knopman, D.S., Boeve, B.F., Petersen, R.C., Jack Jr., C.R., 2007. 3D maps from multiple MRI illustrate changing atrophy patterns as subjects progress from mild cognitive impairment to Alzheimer's disease. Brain 130 (7), 1777–1786.