Received: 16 June 2022 | Revised: 15 November 2022 | Accepted: 16 November 2022

DOI: 10.1002/hbm.26165

TECHNICAL REPORT

WILEY

Fast three-dimensional image generation for healthy brain aging using diffeomorphic registration

Jingru Fu[1] | Antonios Tzortzakakis[2,3] | José Barroso[4] | Eric Westman[5,6] | Daniel Ferreira[5] | Rodrigo Moreno[1] | for the Alzheimer's Disease Neuroimaging Initiative

[1]Division of Biomedical Imaging, Department of Biomedical Engineering and Health Systems, KTH Royal Institute of Technology, Stockholm, Sweden

[2]Division of Radiology, Department for Clinical Science, Intervention and Technology (CLINTEC), Karolinska Institutet, Stockholm, Sweden

[3]Medical Radiation Physics and Nuclear Medicine, Functional Unit of Nuclear Medicine, Karolinska University Hospital, Huddinge, Stockholm, Sweden

[4]Department of Psychology, Faculty of Health Sciences, University Fernando Pessoa Canarias, Las Palmas, Spain

[5]Division of Clinical Geriatrics, Centre for Alzheimer Research, Department of Neurobiology, Care Sciences, and Society (NVS), Karolinska Institutet, Stockholm, Sweden

[6]Department of Neuroimaging, Centre for Neuroimaging Sciences, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

Correspondence
Jingru Fu, Division of Biomedical Imaging, KTH Royal Institute of Technology, Stockholm, Sweden.
Email: jingruf@kth.se

Funding information
Digital Futures, Grant/Award Number: project dBrain; VINNOVA, Grant/Award Number: 2108; Alzheimer's Disease Neuroimaging Initiative; National Institutes of Health, Grant/Award Number: U01 AG024904; DOD ADNI, Grant/Award Number: W81XWH-12-2-0012; National Institute on Aging; National Institute of Biomedical Imaging and Bioengineering; AbbVie; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EUROIMMUN Medizinische Labordiagnostika AG; F. Hoffmann-La Roche Ltd; Genentech, Inc.; Fujirebio US; GE Healthcare; IXICO Ltd; Janssen Alzheimer Immunotherapy Research & Development, LLC; Johnson & Johnson Pharmaceutical Research & Development LLC; Lumosity; Lundbeck; Merck & Co., Inc.; Meso

Abstract

Predicting brain aging can help in the early detection and prognosis of neurodegenerative diseases. Longitudinal cohorts of healthy subjects scanned through magnetic resonance imaging (MRI) have been essential to understand the structural brain changes due to aging. However, these cohorts suffer from missing data due to logistic issues in the recruitment of subjects. This paper proposes a methodology for filling up missing data in longitudinal cohorts with anatomically plausible images that capture the subject-specific aging process. The proposed methodology is developed within the framework of diffeomorphic registration. First, two novel modules are introduced within Synthmorph, a fast, state-of-the-art deep learning-based diffeomorphic registration method, to simulate the aging process between the first and last available MRI scan for each subject in three-dimensional (3D). The use of image registration also makes the generated images plausible by construction. Second, we used six image similarity measurements to rearrange the generated images to the specific age range. Finally, we estimated the age of every generated image by using the assumption of linear brain decay in healthy subjects. The methodology was evaluated on 2662 T1-weighted MRI scans from 796 healthy participants from 3 different longitudinal cohorts: Alzheimer's Disease Neuroimaging Initiative, Open Access Series of

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (https://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

Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; Transition Therapeutics; Canadian Institutes of Health Research; China Scholarship Council; Center for Innovative Medicine (CIMED); Stockholm County Council and Karolinska Institutet; Hjärnfonden; Alzheimerfonden; Demensfonden; Neurofonden; Stiftelsen För Gamla Tjänarinnor; Fundación Canaria Dr Manuel Morales; Fundación Cajacanarias; Estrategia de Especialización Inteligente de Canarias RIS3 from Consejería de Economía; Industria, Comercio y Conocimiento del Gobierno de Canarias; Programa Operativo FEDER Canarias 2014–2020, Grant/Award Number: ProID2020010063; Barncancerfonden, Grant/Award Number: MT2019-0019

Imaging Studies-3, and Group of Neuropsychological Studies of the Canary Islands (GENIC). In total, we generated 7548 images to simulate the access of a scan per subject every 6 months in these cohorts. We evaluated the quality of the synthetic images using six quantitative measurements and a qualitative assessment by an experienced neuroradiologist with state-of-the-art results. The assumption of linear brain decay was accurate in these cohorts ($R^2 \in [.924, .940]$). The experimental results show that the proposed methodology can produce anatomically plausible aging predictions that can be used to enhance longitudinal datasets. Compared to deep learning-based generative methods, diffeomorphic registration is more likely to preserve the anatomy of the different structures of the brain, which makes it more appropriate for its use in clinical applications. The proposed methodology is able to efficiently simulate anatomically plausible 3D MRI scans of brain aging of healthy subjects from two images scanned at two different time points.

KEYWORDS

brain aging, diffeomorphic registration, medical image generation, synthetic brain aging

1 | INTRODUCTION

Brain aging is usually associated with cognitive decline and an increased risk of neurological disorders such as Alzheimer's disease (AD) (Ma et al., 2022; Popescu et al., 2021). Analysis of longitudinal images of healthy brains can reveal the underlying spatiotemporal structure of brain changes due to aging (Giorgio et al., 2010; Peters, 2006), which can potentially be used for the early prognosis and the accurate diagnosis of diseases by serving as the reference of healthy aging (Alberdi et al., 2016; Lorenzi et al., 2015; Mueller et al., 2005).

In the last decades, the collection of longitudinal brain images has facilitated research on brain aging by enabling a noninvasive way to track brain changes and observe disease progression over time (Resnick et al., 2003). For example, one of the most-known data-sharing initiatives, the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Mueller et al., 2005) has collected images using different modalities of subjects at different stages of AD. Compared with other imaging modalities, magnetic resonance imaging (MRI) has superior soft-tissue contrast. This has fostered the use of MRI-based imaging biomarkers, alone or in combination with cerebrospinal fluid and blood biomarkers and psychological tests, for tracking aging or disease-related changes (Devic, 2012; Lockhart & DeCarli, 2014; MacDonald & Pike, 2021; Njeh, 2008; Poulakis et al., 2020, 2021; Schmidt & Payne, 2015).

An alternative approach to the extraction of imaging biomarkers for studying aging is to analyze the complete 3D image in order to detect shape changes of different anatomical structures due to aging (Oxtoby & Alexander, 2017). This approach has the potential advantage of capturing shape changes of the brain structures before they can be detected by specific imaging biomarkers. For example, the volume shrinkage of the brain might occur only in later stages of disease (Cury et al., 2016, 2019).

Recent studies have demonstrated the potential of using machine learning (ML) techniques to study brain aging (Anatürk et al., 2021; Choi et al., 2018; Cole et al., 2015; Ouyang et al., 2021; Popescu et al., 2021). However, it has rarely been straightforward to follow or analyze the age and disease progressions via those learning-based methods. On the one hand, ML-based methods, especially deep learning (DL)-based methods, are designed to distill knowledge from data. Therefore, the availability of ground truth data is crucial to feed the data-hungry DL models. Nevertheless, the sensitive nature of medical data makes it usually difficult to access. Moreover, brain aging research needs longitudinal data, which is even less available or unstable due to, for example, scans taken at different intervals of time or scanners. On the other hand, the high dimensionality of the brain images exponentially increases the resource demands of training DL models (Wegmayr et al., 2019). For this reason, many studies have limited their scope until now to generating two-dimensional (2D) slices extracted from the three-dimensional (3D) MRI scan (Bowles et al., 2018; Kim et al., 2021; Pathan & Hong, 2018). Although the 3D MRI scan can be reconstructed by concatenating 2D slices, it is difficult to assess its internal consistency, with a risk of losing its anatomical plausibility. For all these reasons, it is becoming increasingly necessary to develop *Medical Image Generation* (MIG) models, which aim at generating trusted and accurate synthetic 3D aging brain images in a computationally effective manner.

In this study, we propose a 3D MIG model based on diffeomorphic registration, aiming at synthesizing MRI scans with increasing age, in which subject-level predictions can be derived from individualized image pairs. With our model, the aging progression of the brain can be simulated rapidly with high-dimensional MRI scans. For example, it could synthesize brain atrophy progression from age 60 to age 80 as represented by MRI scans for a particular subject. The main contributions of the proposed method can be summarized as follows:

(i) we develop a new MIG pipeline that can synthesize subject-specific and anatomically plausible MRI series in a computationally efficient manner; (ii) we introduce an aging generative module (AGM) that does not require a training phase and can be applied to any framework that is based on diffeomorphic registration; (iii) we introduce the quality control module (QCM) working in conjunction with AGM, which is used to assess the quality of the synthetic images according to the input pair; (iv) we augment the existing longitudinal MRI scans with corresponding segmentations by around three times and provide the way to access them,[1] enabling the development of data-hungry Artificial Intelligence (AI)-driven healthcare tools, for instance, developing registration and segmentation algorithms for high-resolution predictions, which always require more data to increase their performance.

2 | RELATED WORK

The aging population has increased concern for age-related neurodegenerative diseases and so, aging cohorts and studies have attracted growing interest. MRI scans can clearly illustrate the anatomical structure inside the brain and thus have been used in research on aging. To analyze aging or chronic disease progressions of the brain, AI-based MIG models have been introduced to synthesize scans at different stages of diseases or at different ages.

A commonly used architecture in MIG models is generative adversarial networks (GANs) (Creswell et al., 2018). GANs are designed to generate new data from the same distribution, which consists of two parts: a discriminator to distinguish fake and real samples and a generator to learn new plausible samples to deceive the discriminator. Several GAN-based methods have been introduced to model the aging progression (Bowles et al., 2018; Kim et al., 2021; Wegmayr et al., 2019). Training GANs with 3D brain images is challenging mainly due to the high dimensionality of the brain images. As a result, most of the previous studies have simplified the problem by either using only a single slice per subject (Wegmayr et al., 2019) or by downsampling the original images, which might result in poor resolution predictions (Ravi et al., 2022). To alleviate these limitations (Jung et al., 2021) proposed a method to synthesize high-quality 3D medical images by introducing a 3D discriminator in a normal 2D GAN architecture. A depth-wise concatenation module was introduced to concatenate separate 2D slices into a whole 3D image. Apart from the technical challenges, the main issue of GAN-based methods is that they are unable to guarantee the anatomical plausibility of the generated images due to the lack of biologically informed constraints in the generation. This issue becomes relevant if the synthetic images are expected to be used for answering clinical questions. As an alternative to GANs, other methods using variational autoencoders (VAEs) (Kingma & Welling, 2013) have also been devised, such as Tudosiu et al. (2020). However, it has been reported that GANs tend to produce clearer images than VAEs in diffusion-weighted and T1-weighted images (Treder et al., 2022). More recently, latent diffusion models have been proposed for generating synthetic images (Pinaya et al., 2022).

Apart from the aforementioned issues when using generative models as the architecture, another critical factor for analyzing aging is individualization. Personalized healthcare and individualized medicine are important since each patient has different physical conditions that may cause the same disease. With regard to the aging-related processes, it is even more complicated since aging-related brain changes can be influenced or driven by several factors, such as AD (Song et al., 2022), traumatic brain injury (Cole et al., 2015), even different regions of the brain might follow a different aging pattern (Popescu et al., 2021). Image regression is therefore introduced as a means to encode personalized information in GANs.

Image regression was introduced with the aim of estimating images as a function of associated variables such as age (Niethammer et al., 2011; Beg et al., 2005). The complexity of analyzing age or disease progressions was alleviated by modeling regression approaches at the population level (Dukart et al., 2013; Huizinga et al. 2018). For example, those group-level methods aim at simulating spatiotemporal changes during aging across all subjects. Even though these methods capture the time-varying changes of a population well, the way to leverage and extrapolate this to the target subject is still under development (Campbell & Fletcher, 2017). The work of Pathan and Hong (2018) has addressed this extrapolation problem by incorporating the regression model based on the framework of large deformation diffeomorphic metric mapping (LDDMM) (Pathan & Hong, 2018) with convolutional neural networks and recurrent neural networks. Their model, however, generated a sequence of vector moments under the LDDMM framework before model training, so performance and the model were highly dependent on the LDDMM output.

The fundamental tool for performing image regression is diffeomorphic image registration, which aims at estimating spatial correspondences between images (Zitova & Flusser, 2003). Traditional methods for diffeomorphic image registration are very time-consuming, which has limited the application of image regression in different contexts. Recently, many deep-learning-based diffeomorphic registration methods have emerged (Balakrishnan et al., 2019; Chen et al., 2021; Dalca, Balakrishnan, et al., 2019; Fu et al., 2020; Hoffmann et al., 2022; Li & Fan, 2022) with the aim of reducing the computational burden. These methods have shown a comparable registration accuracy to that of the traditional methods, especially for brain images.

Inspired by the aforementioned methods for simulating brain aging on longitudinal data, we found it necessary to develop a computational effective and anatomically plausible MIG model for 3D scans. Most of the GAN-based methods use one input image to generate synthetic images. Instead, we use two scans at different time points. Aging is a complex process that is influenced by many factors, such as lifestyle factors, cognitive diseases, education, and so on. Using a pair of inputs from the same subject can provide a more accurate picture of the individual aging. With this feature, we can obtain images of a

[1] The synthetic scans can be found via https://github.com/Fjr9516/Synthetic-Brain-Aging/blob/main/README.md

higher quality which can be used to augment the available longitudinal datasets.

3 | METHOD

Figure 1 shows the framework of the proposed aging generative method. As shown, we assume that the input is an image pair of the same subject acquired at different time points. The objective of the aging generation is to synthesize images of aging overtime for that subject. The framework consists of three main parts as shown in Figure 1: (1) the skeleton of diffeomorphic registration; (2) the proposed AGM aiming at simulating linear aging MRI scans; and (3) the proposed QCM, which provides the subject-specific hyperparameter *s* for AGM and that is fulfilled by imposing an accurate-preserve constraint on the synthetic MRI scans.

3.1 | Diffeomorphic registration

Image registration is a fundamental step for analyzing medical images either in clinics or for downstream tasks such as segmentation, regression, or classification. In its simplest form, image registration involves estimating a smooth, continuous mapping between the points in one image and those in another. Specifically, given a moving image I_m and a fixed image I_f, the goal of image registration is to find a deformation field $\boldsymbol{\phi}$ to map I_m into I_f. Preserving the topology is crucial when registering biological image pairs in order to avoid the folding of tissues (e.g., tissues should not fold or even disappear with aging intraindividually). Diffeomorphic registration has the advantage that it computes deformation fields that are both differentiable and invertible, which means that it can preserve the topology.

Traditional diffeomorphic registration methods are computationally intensive. Benefiting from the vigorous development of ML, many ML-based registration methods have been proposed to shorten the registration time in the testing phase from tens of minutes to hours for the traditional methods, to a few minutes or even seconds for the ML-based ones (Fu et al., 2020). Several ML-based diffeomorphic registration methods have been proposed (Dalca, Balakrishnan, et al., 2019; Dalca, Rakic, et al., 2019; Hoffmann et al., 2022; Krebs et al., 2018, 2019; Li & Fan, 2022). We chose SynthMorph (Hoffmann et al., 2022) in this study because of its good performance for brain registration. The architecture of SynthMorph is summarized in the upper part of Figure 1. One of the issues of using ML for diffeomorphic registration is that it is difficult to make sure that the learned deformation fields are diffeomorphic. To solve this issue, SynthMorph divides the problem into two steps. First, a U Net-like neural network (Ronneberger et al., 2015) is trained to learn a stationary velocity field representation, $\boldsymbol{v}$, following a similar approach to the diffeomorphic anatomical registration using exponentiated lie algebra (DARTEL) method (Ashburner, 2007), in which a single velocity field is involved which remains constant over unit time. In a second step, this vector

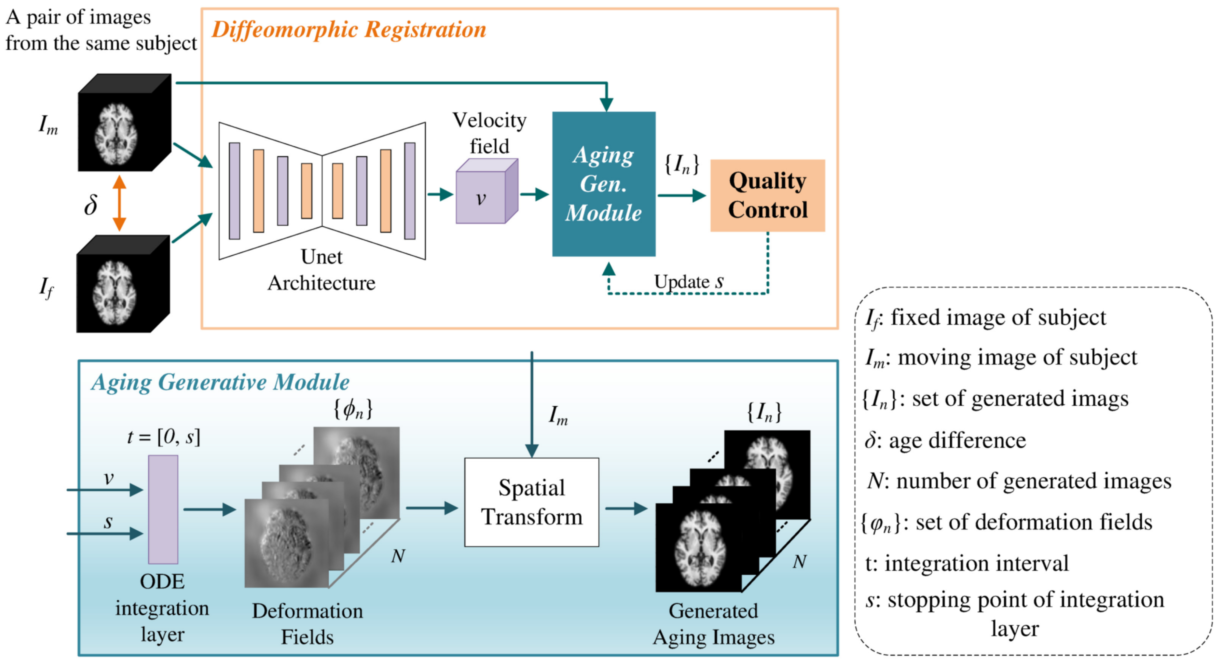


FIGURE 1 Architecture of the proposed Medical Image Generation (MIG) model. The input is an individualized image pair, where I_m stands for *moving* image, I_f stands for *fixed* image. The proposed two modules take *velocity field* as input. *Aging generative module* (*AGM*) and *quality control module* (*QCM*) are introduced within the skeleton of diffeomorphic registration. The details of AGM are shown in the bottom blue shadow part. The *deformation fields* can be derived from *velocity field* given the subject-specific stopping point *s* and the corresponding number of generated magnetic resonance imaging (MRI) scans *N*. At the end, aging MRI sequences can be derived through spatial transform. The *QCM* can provide the subject-specific *s* by applying quality measurements between generated and fixed MRI scans

field is used for estimating the actual diffeomorphic deformation field by solving the ordinary differential equation (ODE):

$$\frac{d\boldsymbol{\phi}^{(t)}}{dt} = \mathbf{v}\left(\boldsymbol{\phi}^{(t)}\right), \tag{1}$$

where $\boldsymbol{\phi}^{(0)}$ is initialized with an identity transform. The final deformation field $\boldsymbol{\phi}^{(1)}$ is obtained by integrating over unit time as follows:

$$\phi = \phi^{(1)} = \int_0^1 \mathbf{v}\left(\phi^{(t)}\right) dt. \tag{2}$$

From group theory, the velocity field $\boldsymbol{v}$ can be seen as a member of the Lie algebra, which is exponentiated in order to produce a deformation $\boldsymbol{\phi}^{(1)}$. The resulting deformation is a member of a Lie group: $\boldsymbol{\phi}^{(1)} = \mathrm{Exp}(\boldsymbol{v})$.

Equation (2) can be solved with the Euler method, which involves calculating a new solution after a series of successive small steps *h*.

$$\boldsymbol{\phi}^{(t+h)} = (\boldsymbol{p} + h\mathbf{v}) \circ \boldsymbol{\phi}^{(t)}, \tag{3}$$

where ∘ denotes the composition operation and $\boldsymbol{p}$ is a map of spatial locations. As an example, a relatively accurate solution can be obtained by using eight time-steps as follows:

$$\boldsymbol{\phi}^{(1/8)} = \boldsymbol{p} + \frac{\mathbf{v}(\boldsymbol{p})}{8}, \tag{4}$$

$$\boldsymbol{\phi}^{(2/8)} = \boldsymbol{\phi}^{(1/8)} \circ \boldsymbol{\phi}^{(1/8)}, \tag{5}$$

$$\boldsymbol{\phi}^{(3/8)} = \boldsymbol{\phi}^{(1/8)} \circ \boldsymbol{\phi}^{(2/8)}, \tag{6}$$

$$\ldots \qquad \ldots, \tag{7}$$

$$\boldsymbol{\phi}^{(1)} = \boldsymbol{\phi}^{(1/8)} \circ \boldsymbol{\phi}^{(7/8)}. \tag{8}$$

If the number of time steps is a power of 2, then it is called *scaling and squaring* (SS) (Arsigny et al., 2006; Ashburner, 2007). The main advantage of this implementation is its relatively low computational cost due to the simplifying of internal points.

Motivated by this, we consider extracting the output deformation fields in the middle of the integration and applying them to the spatial transform block. The evolution between two ages associated with two input images can be simulated.

3.2 | Aging generative module

Our approach for synthesizing images between the fixed and moving ones is to generate different deformation fields at different time steps between the paired input. In addition, the stopping point (i.e., *s*) of integration is set as the hyper-parameter to enable the input-specific focus. The main problem of the *scaling and squaring* approach for our purpose is that the deformation field is computed at irregular time steps (i.e., power of 2) in the integration range. Moreover, it is difficult to obtain fine-grained extrapolation points beyond the original stopping point when $t = 1$. Since our goal is to generate samples at more regular steps, it is beneficial to use a more standard ODE solver that allows us to generate deformation fields at any time *t*. In particular, we used the TensorFlow implementation of the ODE solver to obtain linear intermediate outputs. Specifically, we used the function "odeint" from the TensorFlow Scientific library, which implements a fifth-order Runge–Kuttab using the Dormand–Prince method (Shampine, 1986). We refer to this method as tfODE. This method is just slightly more computationally expensive than SS. The results comparing these two numerical methods can be found in Section 4.2.

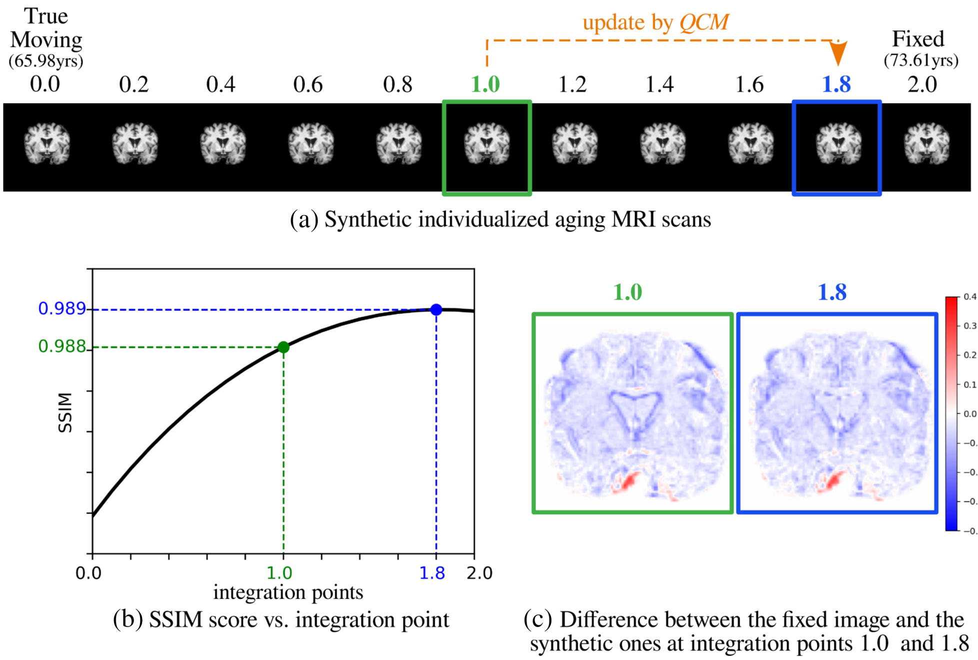
The bottom part of Figure 1 shows the AGM of the proposed method. First, given the velocity field $\boldsymbol{v}$ estimated with the neural network, we generate *N* deformation fields at different regular time points in the range [0, *s*], with *s* being a parameter. In the second step, we use the spatial transform block introduced by de Vos et al. (2017) to generate the images at different time points by warping the moving image I_{m} with the estimated deformation fields. Parameter *s* is referred to as the initial *stopping point*. This parameter is automatically adjusted later by the QCM as described in the next subsection.

3.3 | Quality control module

As already mentioned, in theory $\boldsymbol{\phi}^{(1)}$ should be used to map I_{m} into I_{f}. In practice, this might not happen when using ML-based methods. Although the neural networks learn the most likely vector field $\boldsymbol{v}$ for the input images, some inaccuracies are expected due to the fact that the testing I_{m} and I_{f} are, in general, not used during training. In other words, deformation fields at time points different than one can yield a better matching for registering the two images. As it will be discussed, it is important for the method to accurately estimate this stopping point since the age estimation of the synthetic images is adjusted with respect to that point.

In order to tackle this issue, we introduce the QCM whose aim is to adjust the initial stopping point *s* of the integration layer of the AGM. Our approach is to assess which integration time point yields the most similar generated image compared to I_{f}. Based on previous studies and medical image generation literature (Emami et al., 2018; Gu et al. 2019; Lei et al., 2019), we chose six different similarity measurements between two images I_1 and I_2, namely the mean absolute error (MAE), structural similarity index (SSIM), normalized cross-correlation (NCC), peak signal-to-noise ratio (PSNR), normalized Frobenius norm (NFN), and Dice score (DSC).

Figure 2 shows an example of how *s* is adjusted for the specific case of SSIM. We have chosen hyperparameter *s* as 2 in this case to facilitate understanding. As shown, the image at $t = 1.8$ is more similar to the fixed image than the one at $t = 1.0$ according to SSIM. Although the difference in SSIM is slight, the morphological



(a) Synthetic individualized aging MRI scans

(b) SSIM score vs. integration point

(c) Difference between the fixed image and the synthetic ones at integration points 1.0 and 1.8

FIGURE 2 Adjustment of the stopping point *s*. (a) Series of generated images for a specific subject for different integration points. The center coronal slice from the three-dimensional (3D) volumes is depicted in these images. (b) Structural similarity index (SSIM) between the fixed image and the synthetic ones is maximum at $t = 1.8$. (c) Local differences between the fixed image and the synthetic ones at $t = 1.0$ and the optimal stopping point at $t = 1.8$

differences are visible (see for example the ventricles). The six measurements are computed as follows:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\, I_1(i) - I_2(i)\,|, \tag{9}$$

where N is the number of voxels.

$$\mathrm{SSIM} = \frac{(2\mu_1\mu_2 + C_1)(2\sigma_{12} + C_2)}{(\mu_1^2 + \mu_2^2 + C_1)(\sigma_1^2 + \sigma_2^2 + C_2)}, \tag{10}$$

where μ_i and σ_i stand for the mean and standard deviation of image i, σ_{12} is the covariance, and the parameters $C_1 = (k_1 Q)^2$ and $C_2 = (k_2 Q)^2$ are used to stabilize divisions with weak denominators, with Q being the dynamic range of the MRI scans. We used $k_1 = 0.01$ and $k_2 = 0.02$ in the experiments.

$$\mathrm{NCC} = \frac{\left|\sum_{j=1}^{N}\left(\hat{I}_1(j)\hat{I}_2(j)\right)\right|}{\left[\sum_{j=1}^{N}\hat{I}_1{}^2(j)\sum_{j=1}^{N}\hat{I}_2{}^2(j)\right]^{1/2}}, \tag{11}$$

with $\hat{I}_i(j) = I_i(j) - \mu_i$.

$$\mathrm{PSNR} = 10\log_{10}\left(\frac{Q^2}{\mathrm{MSE}}\right), \tag{12}$$

with Q being the dynamic range of the MRI scans and MSE is the mean squared error between the two images.

Furthermore, we use the normalized Frobenius Norm (NFN) (also known as the sum of squared differences) (Van Loan & Golub, 2013) between the two images:

$$\mathrm{NFN} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|I_1(i) - I_2(i)|^2}. \tag{13}$$

Whenever segmentation masks are available, the Dice score is applied to the segmentation. In order to get segmentation masks for the generated images, the same estimated deformation fields are used to warp the segmentation masks of the moving image. In this case, nearest-neighbor interpolation is used instead of linear used in the spatial transform. The formula is as follows:

$$\mathrm{DSC} = \frac{2 \times (A \cap B)}{A + B}, \tag{14}$$

with A and B being the two segmentation masks. The values range from 0 to 1, 1 representing a perfectly overlapping segmentation.

Lastly, we combine the six similarity measurements by computing the mean updated s of the individual methods.

Once the stopping point s is adjusted, we can re-generate the synthetic images with this more accurate input-specific hyperparameter.

3.4 | Age estimation

When conducting research on aging, age is a valuable piece of information. Once the images are synthesized, the next step is to estimate the age of every synthetic MRI scan. This step is important in order to match the synthesized images with real ones by age.

It is vital to know how anatomy changes with age when it comes to the age estimation of synthetic images. Walhovd et al. (2005) showed that the contraction of brain structures is linear with age.

Moreover, Dukart et al. (2013) found linear decreasing age-related changes in one voxel considering GM volume at the age of 50 years as a baseline. Based on these findings, we assume that a linear increase in the integration time will lead to a linear change in the brain structures. Thus, the age of the synthetic image at time t, I_t, is computed as:

$$\mathrm{Age}(I_t) = \mathrm{Age}(I_m) + \frac{t}{s}[\mathrm{Age}(I_f) - \mathrm{Age}(I_m)]. \tag{15}$$

Notice that this age estimation depends on the stopping point s, which can be different depending on the applied measurement from the previous section. Since there are subjects with more than two acquired images in the datasets, it is possible to use the intermediate acquisitions to assess the error in the age estimation. With this, it is possible to determine which measurement is more appropriate for simulating aging with the proposed methodology.

4 | EXPERIMENTAL RESULTS

4.1 | Datasets

We evaluated the generative performance of the proposed methodology on three datasets: two publicly available datasets, the ADNI (Jack et al., 2008) and the Open Access Series of Imaging Studies-3 (OASIS-3) dataset (LaMontagne et al., 2019); and our own dataset, named Group of Neuropsychological Studies of the Canary Islands (GENIC). They all are 3D brain-MRI datasets. We focus only on the T1-w MRI scans in this study. The ADNI[2] 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 MRI, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. OASIS-3[3] is a retrospective compilation of data from more than 1000 participants, including 609 cognitively normal adults and 489 individuals at various stages of cognitive decline. It contains more than 2000 MR sessions and includes T1-w scans, among other sequences. GENIC is a population-based prospective longitudinal study from the Canary Islands in Spain, which it was started in 2004 and is currently on-going (Machado et al., 2018; Nemy et al., 2020). It includes T1-w scans, among other sequences.

4.1.1 | Data setup

First, FreeSurfer (Fischl, 2012) was applied to all datasets. Image processing and data management of ADNI and GENIC were done in the Hive database system (Muehlboeck et al., 2014), while OASIS FreeSurfer data were obtained from https://www.oasis-brains.org/#access. FreeSurfer performs skull-stripping and bias field correction. After that, we affine registered the images into FreeSurfer's Talairach space using the *talairach.xfm* atlas transform generated by recon-all. Affine registration is necessary since we adopt the stationary velocity model, in which the evolution of the diffeomorphism is not invariant with respect to the affine transformations (Ashburner, 2007). To harmonize medical data for the DL-based architecture, it is important to resample the intensity of images to a common shape and scale between 0 and 1. We also cropped the images to [160,160,192] in the experiments. The segmentations of three datasets were extracted from the file *aparc* + *aseg.mgz* obtained with the widely used FreeSurfer software.

Many neurodegenerative diseases can affect brain aging (Popescu et al., 2021). For example, it has been previously shown that the brains of patients with AD tend to look older than the brains they would have expected when healthy (Franke et al., 2012; Popescu et al., 2020). Based on this, it is reasonable to separate healthy patients from diseased patients, especially for age estimation. Therefore, we only used images of cognitively healthy subjects, resulting in 1489 images in ADNI, 1310 images in OASIS-3, and 406 in GENIC. Furthermore, the proposed methodology requires images acquired in at least two time points as input, so subjects with sessions fewer than two were excluded. Then 1393 images were left in ADNI, 1066 images in OASIS-3, and 203 images in the GENIC dataset. Details about the data included in the experiment appear in Table 1.

4.2 | Image generation

As mentioned, SynthMorph (Hoffmann et al., 2022) was used as the backbone of the diffeomorphic registration due to its state-of-the-art performance for DL-based diffeomorphic registration. We used pretrained weights which were trained with a set of brain-anatomy label maps (*sm-brains*).[4] We also conducted an experiment comparing the two numerical ODE solvers (i.e., SS and tfODE) in terms of quantitative measurements used in Section 3.3 in two images from the OASIS-3 dataset. We found that the two methods could achieve comparable results even over a wide time range (see Figure S1), which is in line with the findings of Dalca, Balakrishnan, et al. (2019). The number of images generated for each subject, N_i, is determined by the age difference (i.e., δ shown in Figure 1) between the youngest and oldest sessions in the dataset, namely $N_i = 2 \times \delta_i$ for each subject i. We chose the two MRI scans with the greatest age gap from a subject for two reasons: (i) it will result in the longest aging simulation and relatively large augmented data pool that can be used for developing data-hungry AI-enable tools, such as registration and segmentation of 3D MRI scans; (ii) the remaining intermediate MRI scans can be used for the evaluation part. It was chosen here to use double age difference because we consider the acquisition of longitudinal data to be

[2] http://adni.loni.usc.edu/about/
[3] https://www.oasis-brains.org
[4] https://surfer.nmr.mgh.harvard.edu/ftp/data/voxelmorph/synthmorph/brains-dice-vel-0.5-res-16-256f.h5

TABLE 1 Summary of the datasets

	Complete dataset		Selected images			Synthetic images	
Dataset	# Images	# Healthy	# Sessions >1	# Subjects	Age range	# Images	Size increase (%)
ADNI	5097	1489	1393	347	59–95	2500	179
OASIS-3	2044	1310	1066	353	42–95	3948	370
GENIC	539	406	203	96	34–79	1100	542
Total	7680	3205	2662	796	34–95	7548	284

Abbreviations: ADNI, Alzheimer's Disease Neuroimaging Initiative; OASIS-3, Open Access Series of Imaging Studies-3.

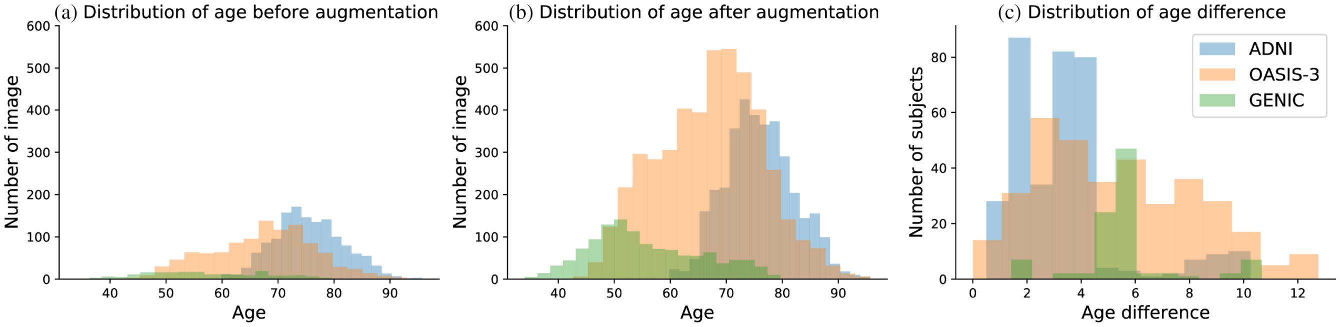


FIGURE 3 Age distributions in the three datasets. (a,b) The histogram of age before and after augmentation. (c) The distribution of age difference between the first and last acquired magnetic resonance imaging (MRI) scan per subject

suitable every 6 months. The initial stopping point s is set as three in the experiments. The summary of the datasets and the generated synthetic images can be found in Table 1. It is worth mentioning that the original datasets can be augmented with high-quality MRI scans by 284%.

Figure 3 shows the age distributions of the three datasets before and after the generation of synthetic data. As shown, GENIC contains younger subjects, ADNI older, and OASIS-3 subjects in the middle. The figure also shows that OASIS-3 covers a larger range of age differences between the first and last MRI acquisition compared to the ADNI, with GENIC in between. These differences determine that the number of synthetic images per subject in ADNI is on average smaller than in the other two datasets.

4.3 | QCM validation

Introducing QCM is one of the contributions of this work since it can take the quality of synthetic MRI scans into consideration by adjusting stopping points in the AGM at the inference phase, thus mitigating the effects of domain shift between training and test cohorts. Six similarity measurements are introduced in the QCM. The validity of QCM is evaluated from three perspectives: (i) selected measurements reflect aging-related changes; (ii) the “optimal” stopping point is mostly beyond the fixed one (i.e., $t = 1$); (iii) significant differences are found before and after applying QCM in terms of six criteria.

As discussed, the stopping point can be different depending on the quality measures used for comparing the acquired images with the synthetic ones. To show this, we randomly chose a subject from OASIS-3 that was scanned five times and used them to assess their corresponding closest synthetic images according to the different criteria. Every curve in Figure 4 shows the evolution of the different measurements for the five images with the integration points. Notice in Figure 4 that the integration point of the closest synthetic image is always growing with the age of the acquired image for all measurements. Similar behavior was observed in subjects with more than two acquired images. This means that the chosen quality measurements are consistent and can capture aging-related changes.

Figure 5 shows box plots of the adjusted stopping points s per dataset for the different quality measurements. As shown, the stopping point is higher than 1.0 in the vast majority of the cases. ADNI tends to have lower values of s closer to theoretical stopping point 1 compared to the other datasets. One hypothesis for this is that ADNI was the only dataset that was included in the training phase of SynthMorph (Hoffmann et al., 2022). Notice that the value of s is relatively similar for all quality measurements. To further evaluate the difference among datasets from a quantitative perspective, we computed the Fréchet inception distance (FID) (Heusel et al., 2017) between each pair of datasets. FID is widely used in the GAN literature and is a popular metric for measuring the feature distance between two distributions, which also shows sensitivity to image quality and good correspondence with human perception (Treder et al., 2022). We randomly selected 200 samples of each dataset, took a representative slice (the central slice in the sagittal direction), and computed FID between the datasets. This procedure was repeated

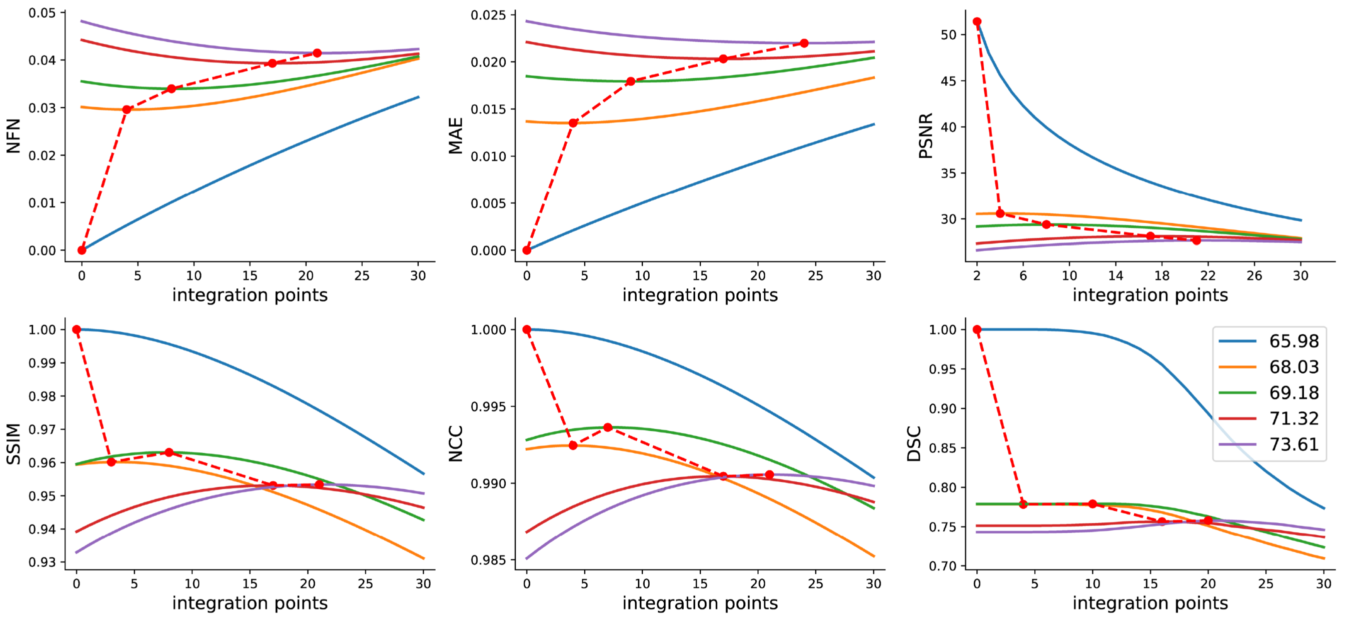


FIGURE 4 Trends of the six quality criteria for five images of the same subject. The legends indicate the *true ages* of the corresponding real magnetic resonance imaging (MRI) scans. For each curve, we calculate the "extreme" values and positions and connect them to a dashed red line. For normalized Frobenius norm (NFN) and mean absolute error (MAE), the generated MRI scans that are most similar to the real ones are at the point of minimum value; for other measurements, it is at the maximum point. Note that the peak signal-to-noise ratio (PSNR) plot is not shown from 0 since it is not defined for that value

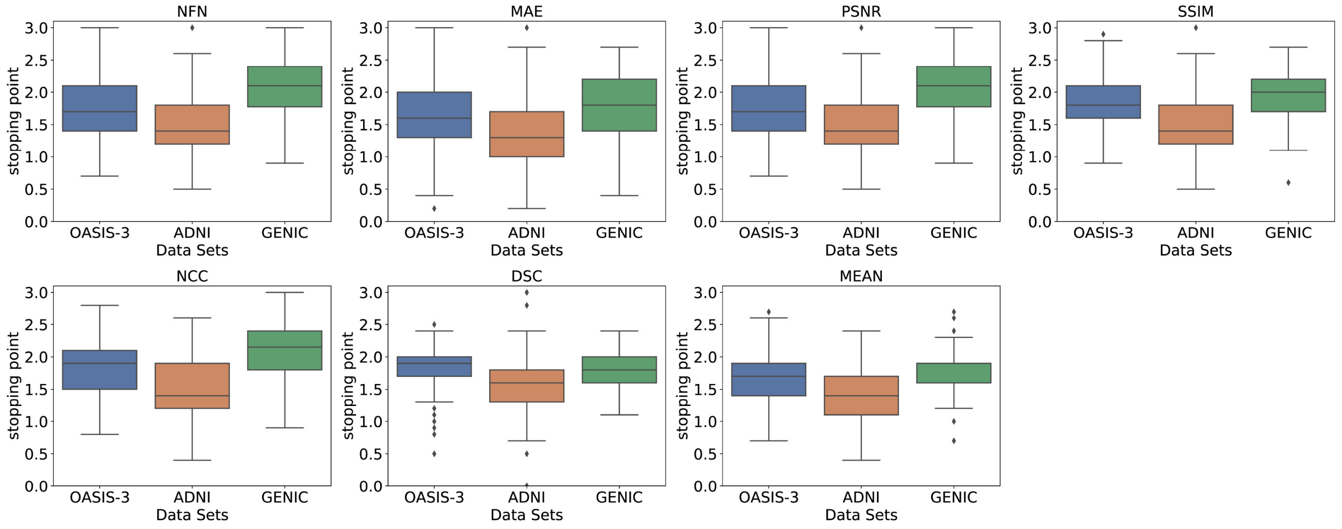


FIGURE 5 Distribution of the stopping point per dataset for the tested quality measurements

10 times to get the average FID and standard deviation between each pair of datasets. These results are summarized in Table 2. According to the table, OASIS-3 is relatively close to ADNI, while GENIC is far away from both OASIS-3 and ADNI. This finding supports that OASIS-3 had optimal stopping points closer to the ones from ADNI.

Although the computational cost of adjusting the image generation to the optimal point s is low, one relevant question is if such a procedure contributes to get better images. For answering this question, we conducted an experiment on GENIC in which we compared the images generated with QCM (i.e., adjusting s) and

TABLE 2 Comparison of FID measurements between paired two datasets. Standard deviations are shown in parentheses

Datasets	GENIC	OASIS-3	ADNI
GENIC	1305 (239)	20,089 (1074)	17,108 (1111)
OASIS-3	19,382 (1244)	1754 (192)	4393 (453)
ADNI	17,013 (1054)	4227 (530)	1880 (109)

Abbreviations: ADNI, Alzheimer's Disease Neuroimaging Initiative; FID, Fréchet inception distance; OASIS-3, Open Access Series of Imaging Studies-3.

without QCM (i.e., with $s = 1$). As shown in Table 3, the values for the six measurements are similar. Still, we found that SSIM and DSC scores are statistically different ($p = .018$ for SSIM and $p = .03$ for DSC). Since GENIC contains a younger population, the changes due to aging are less pronounced. Thus, the differences between using QCM or not are expected to be larger with older subjects (as well as in pathological subjects). Notice that the computational time for these methods is very similar. Thus, it is advantageous to adjust the optimal stopping point since it is computationally inexpensive.

4.4 | Validation of age estimation

At this point, it is not clear which quality measurement is the most appropriate for age estimation. To answer this question, we used the true age of the intermediate images, which were not used in the image generation, as ground truth to test the error in the age estimation. We used the root mean square error (RMSE) between the true ages and the estimated ages as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathrm{Age}_{\mathrm{true}}(I_i) - \mathrm{Age}_{\mathrm{estimated}}\left(\hat{I}_i\right)\right)^2}, \qquad (16)$$

where N is the number of images in the ground truth, I_i is the ith image in the ground truth, and $\hat{I}_i$ is its closest synthetic image according to the tested quality measurement.

TABLE 3 Quantitative comparisons of images generated with the adjustment of the stopping point s (with quality control module [QCM]) and without that adjustment ($s = 1$) on the GENIC dataset. The same six quality measures were computed for these images

Quality measurement	With QCM	Without QCM
NFN	**0.035 (0.008)**	0.036 (0.008)
MAE	**0.015 (0.004)**	**0.015 (0.004)**
PSNR	**29.33 (1.82)**	29.04 (1.83)
SSIM	**0.955 (0.012)**	0.950 (0.014)
NCC	**0.982 (0.008)**	0.980 (0.008)
DSC	**0.729 (0.022)**	0.722 (0.025)

Note: The best value on each row is marked in bold.
Abbreviations: DSC, Dice score; MAE, mean absolute error; NCC, normalized cross-correlation; NFN, normalized Frobenius norm; PSNR, peak signal-to-noise ratio; QCM, quality control module; SSIM, structural similarity index.

Table 4 shows the RMSE for the tested quality measurements. The corresponding box plots are shown in Figure 6. We observe that the RMSE is around 2 years when the three datasets are combined. Again, the error is lower for ADNI. As shown, NFN, PSNR, and NCC are the best options for age estimation. Notice that the RMSE estimations might be affected by quantization errors since we generate images to simulate increments of 0.5 years of age.

As mentioned, previous studies have found that the brain changes linearly with age (Dukart et al., 2013; Walhovd et al., 2005). In order to assess the validity of that hypothesis in our datasets, we performed linear regressions between the real age of the images used as ground truth and the estimated as described in Section 3.4. Figure 7 shows these plots for the tested quality criteria. According to the coefficient of determination R^2, the linear regression is valid since all measurements are higher than .9, which usually indicates a strong correlation between variables thus demonstrating the goodness of this fit. Ideally, the fitting lines should be $y = x$. As shown, the slopes are close to one in all cases, which validates the hypothesis that the linear changes within the brain will cause linear age increments. However, the intersection varies between 0.15 (i.e., 1.8 months) and 1.23 (i.e., 1 year and 2.8 months). To assess the effect of the intersect in the estimation of age, we added an extra column in Table 4 where the regression was used instead of the linear estimation of age. The improvement is in the range of 0.07–0.12, which is equivalent to 0.84–1.44 months.

4.5 | Quantitative quality assessment

Table 5 and Figure 8 show the different quality measurements computed on the generated images that are most similar to the ground truth images. The obtained values cannot be directly compared with

TABLE 4 RMSE of the age estimation for the tested quality measurements. Columns 1–4 show the error of the linear model, while the last one shows the RMSE of the fitted regression lines of Figure 7

Quality measurement	OASIS-3	ADNI	GENIC	Three datasets	Regressions of Figure 7
NFN	2.78	1.67	2.80	**2.13**	**2.02**
MAE	2.94	1.84	2.69	2.28	2.18
PSNR	2.78	1.67	2.80	**2.13**	**2.02**
SSIM	2.93	1.80	2.88	2.26	2.14
NCC	2.80	1.67	2.83	**2.13**	**2.02**
DSC	3.00	1.92	2.73	2.35	2.27
MEAN	2.83	1.72	2.72	*2.17*	*2.10*

Note: Bold is the best and italic is the second best.
Abbreviations: ADNI, Alzheimer's Disease Neuroimaging Initiative; DSC, Dice score; MAE, mean absolute error; NCC, normalized cross-correlation; NFN, normalized Frobenius norm; OASIS-3, Open Access Series of Imaging Studies-3; PSNR, peak signal-to-noise ratio; RMSE, root mean square error; SSIM, structural similarity index.

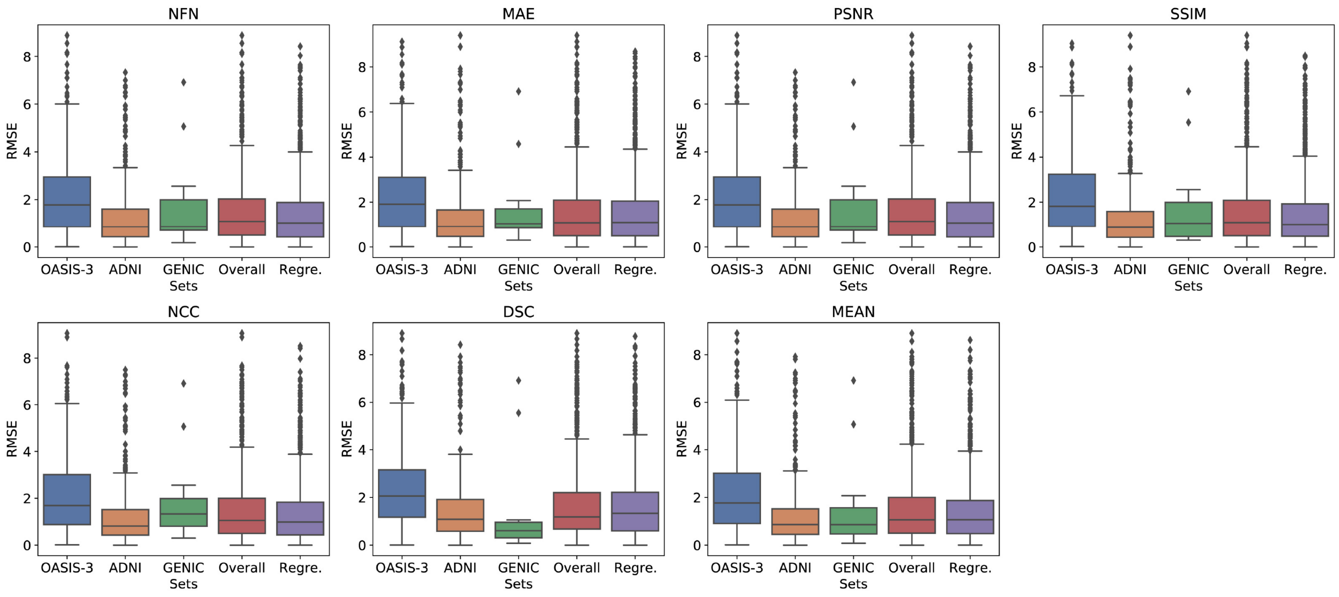


FIGURE 6 Box plots of the root mean square error (RMSE) of the age estimation for the tested quality measurements

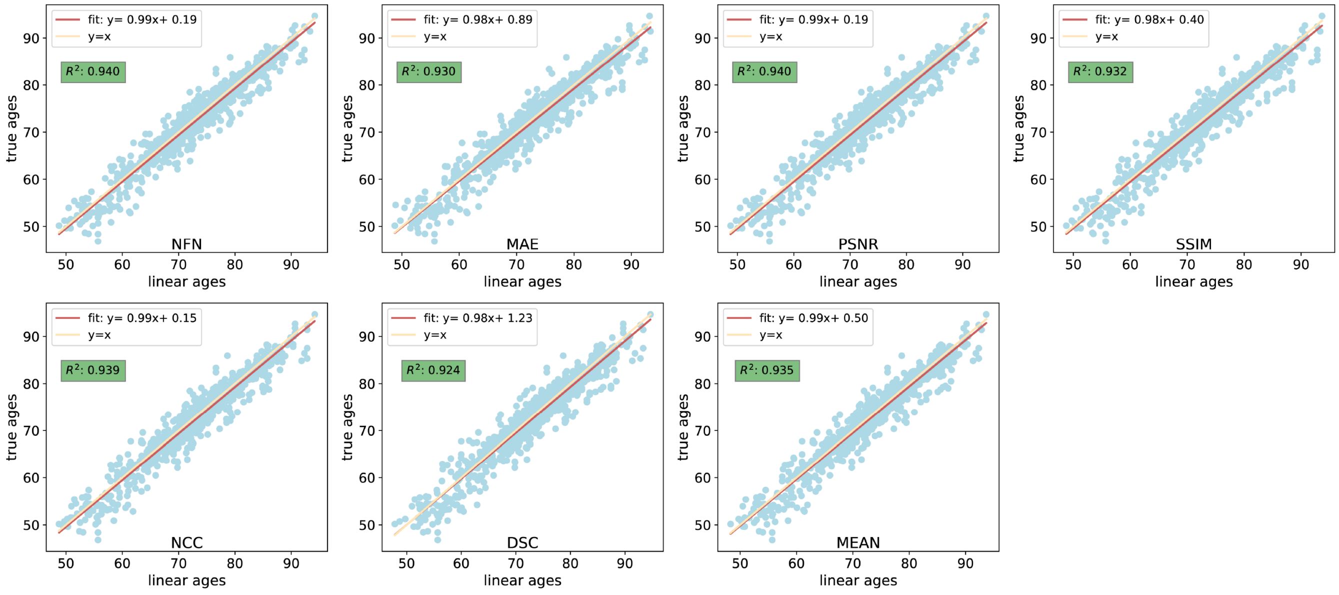


FIGURE 7 Correlation plots between the estimated linear ages and the true ages of the ground truth images. The coefficient of determination R^2 is used to determine the performance of the fitting

other generative models from the literature because of different experimental settings. Still, these values are competitive or superior to those from previous studies (Emami et al., 2018; Gu et al. 2019; Lei et al., 2019). For example, in these studies, PSNR was around 28 (we got 27.87), SSIM was around 0.85 (we got 0.936), and NCC was around 0.93 (we got 0.982). We observed that NFN and MAE are very small as well.

Regarding the DSC, Hoffmann et al. (2022) reported values around 0.75, compared to the mean of 0.727 of our experiments. DSC is commonly used for assessing the performance of image segmentation methods. In such applications, DCS values of 0.70–0.75 are not considered accurate. However, it is important to consider that many brain structures are small, which usually has a direct impact on the DSC. Considering that, Hoffmann et al. (2022) used the 26 large brain structures for computing the DSC. We decided to report DSC considering 113 structures in total, on Table 5, to get a more comprehensive overview of the performance estimated with DSC.

4.6 | Comparison with baselines

We compared our results on GENIC with two basic baselines: (i) Baseline$_{\text{deformation}}$ in which the longitudinal scans are synthesized by linearly scaling the deformation field; (ii) Baseline$_{velocity}$ in which the longitudinal scans are synthesized by linearly scaling the velocity field. Notice that these two strategies cannot guarantee diffeomorphism. Table 6 shows the six image quality measurements of the proposed method compared with the two baselines for the GENIC dataset. From the results, we can observe that the improvements are relatively small but using integration (our proposed method) we can achieve the best results for all six criteria. The visual assessment shows that our method gives better results (see Figures S2 and S3). As mentioned, the two baselines cannot guarantee diffeomorphism. The errors are expected to increase with age where anatomical changes in the brain are more visible than in the younger population.

TABLE 5 Quality measurements of the most similar generated images compared to the ground truth images for different datasets and criteria

Quality measurement	OASIS-3	ADNI	GENIC	Average
NFN	0.056	0.040	**0.035**	0.044
MAE	0.029	0.020	**0.016**	0.022
PSNR	25.81	28.73	**29.07**	27.87
SSIM	0.917	**0.946**	0.945	0.936
NCC	0.981	**0.985**	0.981	0.982
DSC	0.715	**0.749**	0.718	0.727

Note: The best value on each row is marked in bold except for the average column.

Abbreviations: ADNI, Alzheimer's Disease Neuroimaging Initiative; DSC, Dice score; MAE, mean absolute error; NCC, normalized cross-correlation; NFN, normalized Frobenius norm; OASIS-3, Open Access Series of Imaging Studies-3; PSNR, peak signal-to-noise ratio; RMSE, root mean square error; SSIM, structural similarity index.

4.7 | Qualitative quality assessment

Figure 9 shows an example subject of the generated aging images. The comparison between real-aging MRI scans and synthetic aging MRI scans shows the good quality of the synthetic images. Figure 10 shows the estimated aging progression of this subject, where the aging MRI scans are shown in three directions (coronal, sagittal, and axial), ranging from 51.7 to 63.7 years old. We report a magnified region to show the aging progression in the sagittal direction. As shown, the ventricles expand with increasing age. This can be seen as an indication that the proposed method is consistent with what is expected in the aging brain.

Although the synthetic images look realistic to untrained eyes, it is necessary to perform validation with a neuroradiologist in order to assess the quality of the generated images. This step is important toward the use of synthetic data for answering clinical questions.

In order to accomplish this, we designed a discrimination task for the neuroradiologist (A. T.), who has 16 years of experience with neurological images. The discrimination task consisted of distinguishing real images from synthetic ones. From the pool of mixed generated and real images, 200 MRI scans were randomly selected. Because we are interested in knowing if there was any bias in different datasets and age difference between the oldest and youngest image of the subject used to generate the image, we selected the 200 samples proportionally for each subcategory. By dividing the age difference into six ranges (i.e., 2 years per range), we got [0–2, 2–4, 4–6, 6–8, 8–10, 10+]. We included the generated MRI scans at time point *s* as the generated one in the pool.

The neuroradiologist was completely blinded to the purpose and design of the study, as well as to the demographic and clinical characteristics of the study participants. The neuroradiologist used the 3D

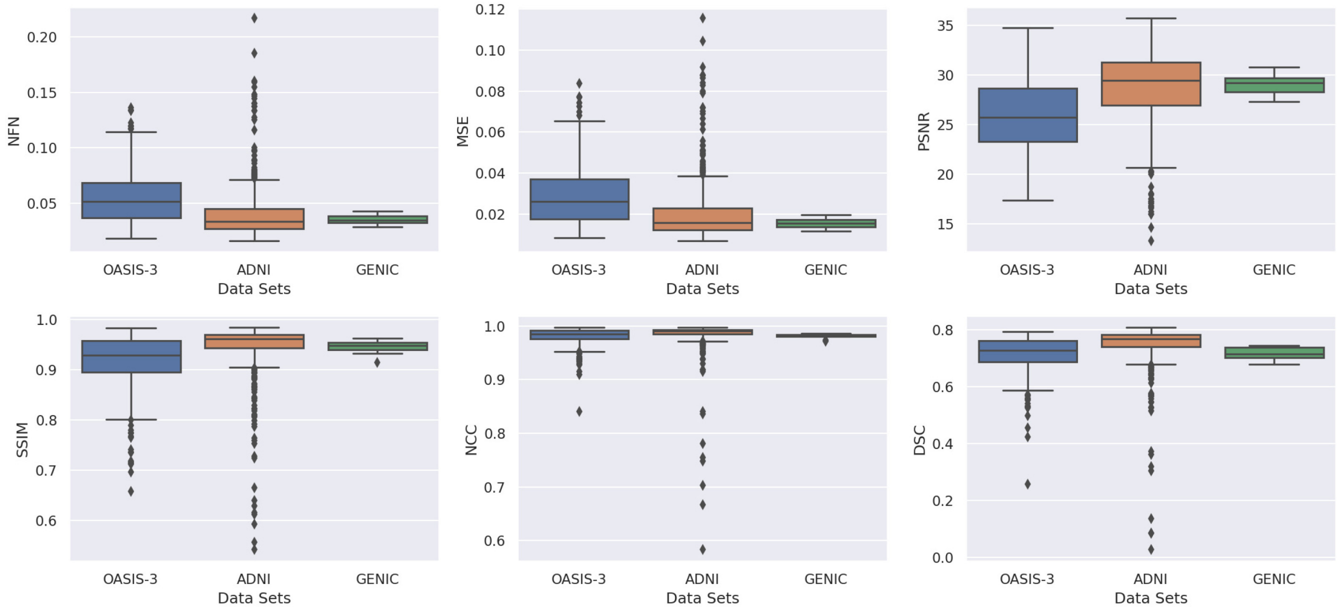


FIGURE 8 Box plots of the quality measurements of the most similar generated images compared to the ground truth images for different datasets and criteria

TABLE 6 Comparison with baselines: Quantitative comparison between the most similar generated images and the ground truth images for different strategies for obtaining deformation fields on the GENIC dataset.

Quality measurements	OURS$_{\text{integration}}$	Baseline$_{\text{deformation}}$	Baseline$_{\text{velocity}}$
NFN	**0.035 (0.004)**	0.036 (0.005)	0.036 (0.005)
MAE	**0.016 (0.003)**	**0.016 (0.003)**	**0.016 (0.003)**
PSNR	**29.07 (1.02)**	28.97 (1.09)	28.97 (1.09)
SSIM	**0.945 (0.013)**	0.943 (0.016)	0.943 (0.016)
NCC	**0.981 (0.004)**	0.980 (0.005)	0.980 (0.005)
DSC	**0.718 (0.002)**	0.715 (0.002)	0.716 (0.002)

Note: The best value on each row is marked in bold.
Abbreviations: DSC, Dice score; MAE, mean absolute error; NCC, normalized cross-correlation; NFN, normalized Frobenius norm; PSNR, peak signal-to-noise ratio; RMSE, root mean square error; SSIM, structural similarity index.

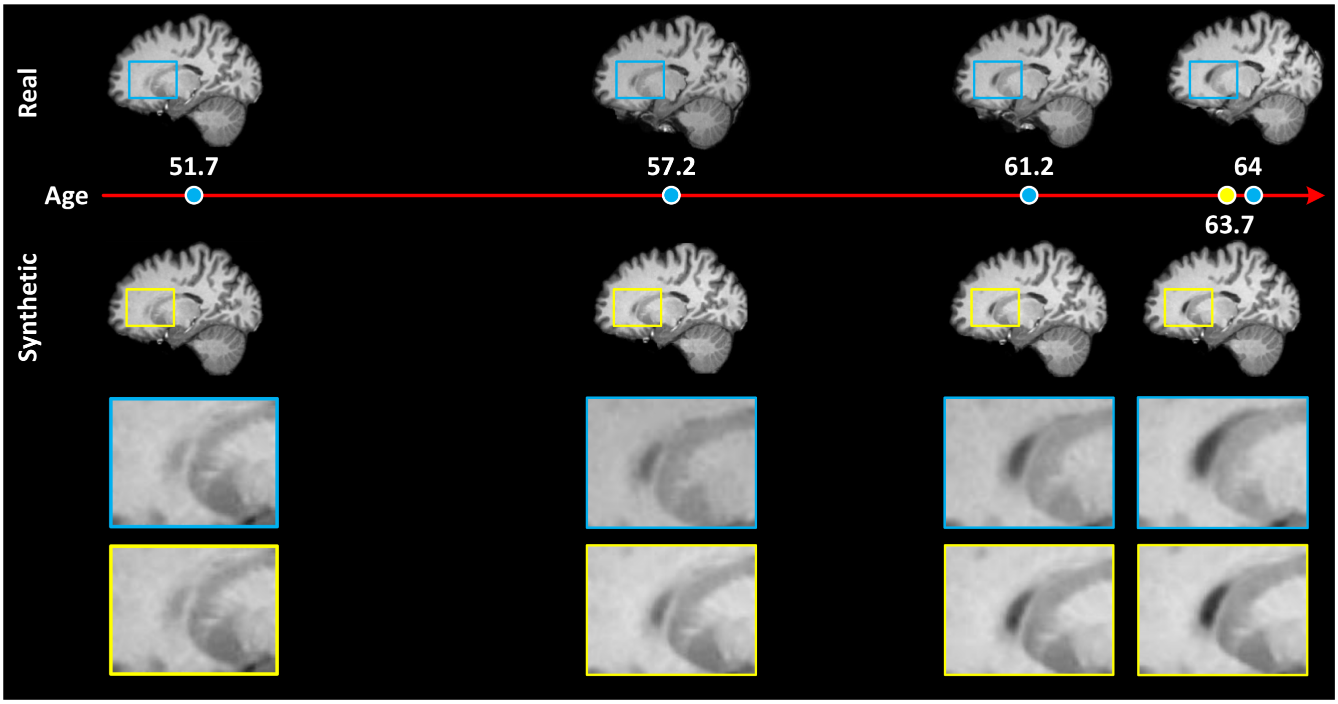


FIGURE 9 Qualitative assessment of the quality of the synthetic magnetic resonance imaging (MRI) scans versus the real MRI scans. The first row indicates the real MRI scans in the longitudinal dataset, and the second row shows synthetic aging MRI scans at different estimated ages. We also show a magnified region at the bottom of the figure for each row respectively (color figure online). As the synthetic MRI scans are acquired along with the same interval of 6 months, the last obtained brain age is 63.7 in this case, whereas the counterpart in real data is 64

slicer[5] for classifying the 200 selected MRI scans as real or synthetic. He completed the task in four consecutive days—1 h of assessment per day. Before doing the task, the neuroradiologist was exposed to three true MRI scans, one per dataset, in order to help the neuroradiologist to build a template of a real image in this dataset. Motivated by the experiments by Ravi et al. (2022), for each case, the expert was asked to assign a confidence level from the given list:

- None: “I have no idea, I am guessing the class of this scan.”
- Low: “I have low confidence in my answer.”
- Medium: “I am reasonably confident in my answer.”
- High: “I am absolutely sure in my answer.”

Table 7 shows the confusion matrix of the assessment in which the real class is marked as positive (P) and synthetic as negative (N). As shown, highly skilled experts struggle to recognize the real and synthetic images with an *accuracy* of 63.5%. Specifically, while even the expert can reach an accuracy rate of 63.5%, the precision for detecting real images (45.6%) is poor, suggesting our synthetic images look realistic.

Figure 11 shows the distribution of confidence levels reported by the neuroradiologist in the experiment. As we can see, the expert has high confidence in only 13% of the cases. Moreover, the neuroradiologist has more uncertainty for synthetic images (i.e., none confidence in synthetic was 11.9% compared to 3% for real images, p-value $=$ .042).

Table 8 shows the accuracy and $F1$-score of the neuroradiologist in distinguishing between real and synthetic images. Moreover, we independently report the $F1$-scores for the synthetic and real images in Table 8. To evaluate if the algorithm shows differences among

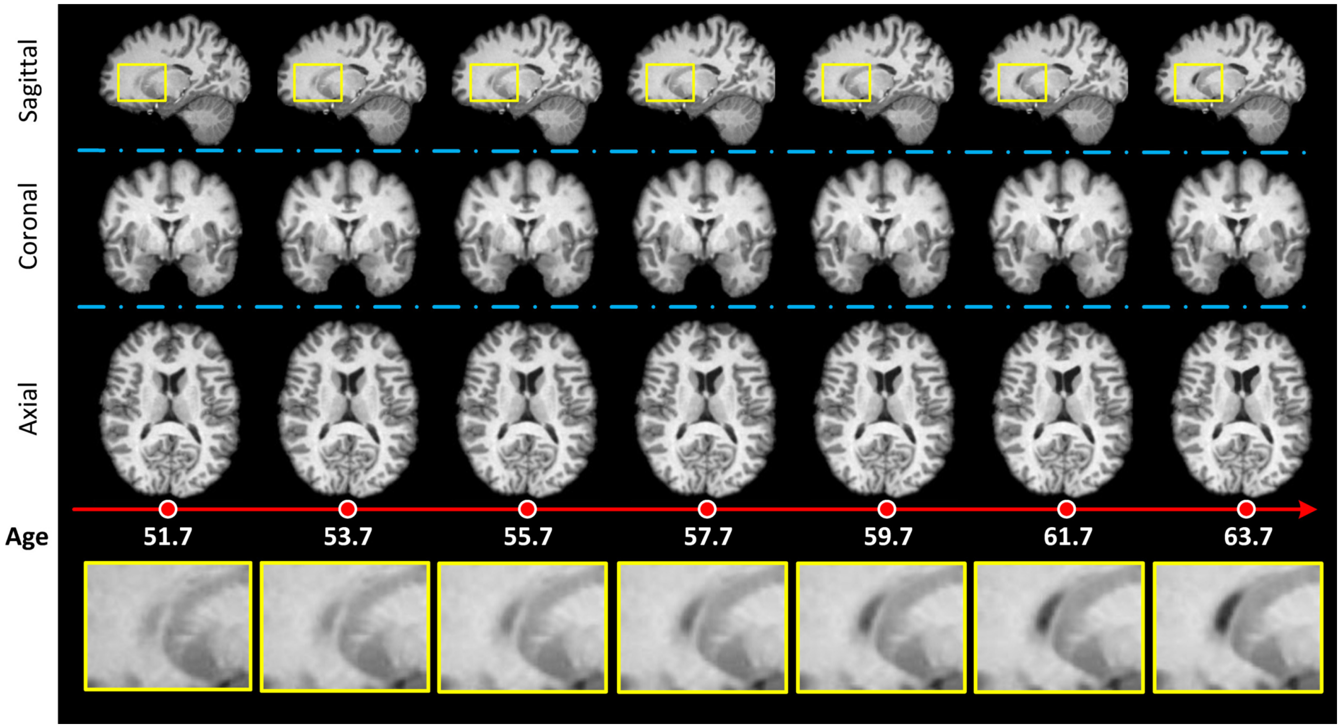


FIGURE 10 The aging simulations were synthesized using our methodology for a healthy subject from age 51.7 to age 63.7. The three-dimensional (3D) magnetic resonance imaging (MRI) scans are shown in three directions sagittal, coronal, and axial, respectively. A magnified region is highlighted at the bottom of the figure for a better illustration (color figure online)

TABLE 7 Confusion matrix on the discrimination task performed by the neuroradiologist

	Neuroradiologist assessment	
	Real (P)	**Synthetic (N)**
Real	True positive = 41	False negative = 24
Synthetic	False positive = 49	True negative = 86

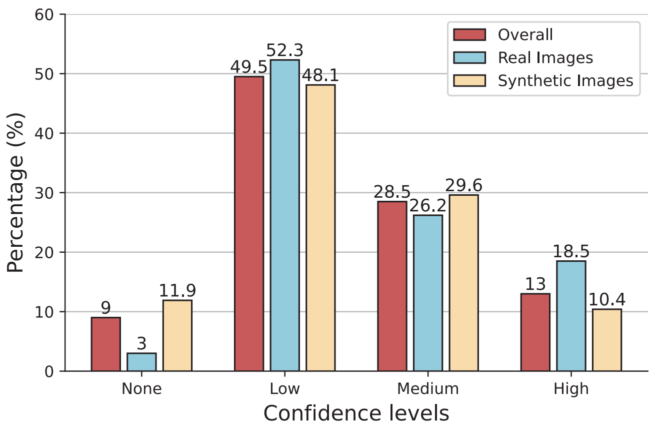


FIGURE 11 The distribution of confidence levels

different datasets and confidence levels, we also calculate the metrics for each subcategory. It is also worthwhile to analyze the age difference between the youngest and oldest images from the subjects since in our experiments we chose to generate *N* images for each subject based on such age difference. According to Figure 3, there is a large age difference between the three datasets, so we decided to divide the range of age differences into six small intervals, each of which contains 2 years.

As expected, the neuroradiologist was able to distinguish better the two classes when his confidence was high. Regarding confidence levels between none to medium, the performance was better for synthetic images for the same level of confidence, because the neuroradiologist tends to be more synthetic-oriented as demonstrated by the precision for detecting real images of only 45.6%. The accuracy was similar for different datasets (62 ± 5%), whereas the *F*1-score on overall for GENIC is about 0.2 lower than the other two datasets. One possible reason for this is that images from GENIC come from younger subjects that might have fewer visual distinctions due to age. This can make the deformation fields to be small, which increases the chances of synthetic ones being more similar to the real ones.

We conducted proportion hypothesis tests on the age differences in accuracy and corrected the resulting *p*-values by multiple comparisons ($p < .05$). Based on the results, we observed that there is no significant difference among subgroups even over a period of 10 years, thus indicating the validity of QCM and robustness of the proposed method.

5 | DISCUSSION

We presented a method to efficiently generate 3D MRI scans of aging brains with the aim of augmenting the current longitudinal datasets with high-quality images. Our method is able to leverage DL-based methods to generate synthetic images while avoiding the time-

TABLE 8 Comparison of the accuracy and *F*1-scores of the assessment performed by the neuroradiologist

		Confidence levels				Datasets			Age differences					
Criteria	Overall	None	Low	Medium	High	OASIS-3	ADNI	GENIC	0–2	2–4	4–6	6–8	8–10	10+
Accuracy	0.64	0.56	0.66	0.58	0.73	0.67	0.62	0.57	0.46	0.67	0.66	0.74	0.50	0.71
*F*1-score (overall)	0.62	0.45	0.62	0.57	0.73	0.65	0.62	0.42	0.41	0.66	0.61	0.73	0.49	0.71
*F*1-score (real)	0.53	0.20	0.51	0.50	0.74	0.58	0.56	0.13	0.24	0.60	0.47	0.70	0.44	0.67
*F*1-score (synthetic)	0.70	0.69	0.73	0.64	0.72	0.72	0.67	0.71	0.58	0.72	0.75	0.77	0.55	0.75

Abbreviations: ADNI, Alzheimer's Disease Neuroimaging Initiative; OASIS-3, Open Access Series of Imaging Studies-3.

consuming training process. The use of image registration also makes the generated images plausible by construction. In addition, we propose a series of strategies to ensure that the model can provide predictions of age for the generated images.

5.1 | Image generation

The proposed method was evaluated on 2662 T1-w MRI scans from 796 participants collected in three different datasets. Although T1-w MRI is the only modality used, the methodology is applicable to any other modality. We synthesized 7548 high-quality images from the three datasets, which corresponds to an increase of 284% in the number of images. Notice that the synthetic images were generated in the range between the youngest and oldest image per subject. It is actually straightforward to generate more images by extrapolating the images beyond the oldest one. In fact, we discarded approximately one-third of the synthetic images because we first generated images up to the integration point of three to estimate the stopping point *s*, which was close to two, so images beyond *s* were discarded. The reason for this decision is that we were focused on creating a dataset of high-quality images that can be combined with existing datasets for brain aging analysis. Although we think images beyond the stopping point *s* are also of high quality, it is more difficult to assess that at this point. Future works may expand our current method in that direction.

Notice that we decided to exclude AD patients or subjects with MCI in this study. Indeed, it is straightforward to apply the method to these subjects. However, the aging patterns of these patient groups coexist with disease-related patterns, making the age estimation more difficult than in cognitively healthy individuals. This means that a synthetic dataset of aging in patients must consider this problem in order to make it useful for aging studies. Thus, we decided to develop and demonstrate the goodness of our method in healthy individuals, and future developments shall consider applying the method to disease populations such as AD.

5.2 | QCM verification

Generally speaking, DL models always face a problem that the performance of the model is limited to the distribution of training data. When the distribution of the prediction data and the distribution of the training are different, the performance of the model tends to drop. This can be supported by FID scores in Table 2. The table also shows that GENIC is slightly more homogeneous (see the diagonal values of the table), which might have contributed to getting slightly lower variances seen in Figure 5 (plot with the title MEAN). From Figure 3, the range of ages of GENIC is very different from those of both ADNI and OASIS-3. Thus, a possible explanation for the difference in FID for GENIC is that its range of ages is very different from the other two datasets (see Figure 3). We tackled that problem by using quality measurements to estimate the most similar generated image to the acquired ones. As shown in Figure 5, the closest image was almost always at a stopping point higher than one regardless of the used quality measurement, while theoretically that should be 1.0. This way, we managed to benefit from the speedup provided by DL-based registration methods while keeping image quality as high as possible.

We selected six different quality measurements for estimating the stopping point *s* and to estimate the quality of the images. All criteria gave consistent results as the closest synthetic image was always growing with the age of the acquired image for all criteria (Figure 4).

The results also show that these measurements perform similarly for age estimation, with NFN, PSNR, and NCC being slightly better (cf. Table 4). Except for DSC, these measurements also show that the generated images are of good quality (cf. Table 5). However, the obtained DSC values are similar to the ones from previous studies, especially Hoffmann et al. (2022). Moreover, the small size of some regions of the brain can be biased in the estimation of DSC.

Although the focus of this paper is to generate T1-w images, the method can be used for other modalities (e.g., T2w, FLAIR, etc.) without any change. More interesting would be to generate images of one modality using images from another modality. For example, let us assume that in the first session, T1-w and T2w images were acquired, but only a T1-w was acquired in the second session. Then, the two T1-w images can be used to generate the deformation fields that can be applied to the only available T2w image. Another example would be if only a T1-w is available from the first session and only a T2w image is available for the second. In this case, it is necessary to use quality measurements that can deal with multimodality data, for example, DSC or mutual information.

5.3 | Qualitative assessment

From the experiments, the synthetic images are of high quality. The assessment performed by the neuroradiologist (A.T.) shows that it is difficult to distinguish between synthetic images and real ones, especially when the neuroradiologist's confidence was not high. As expected, the images generated between smaller age differences were more difficult for the neuroradiologist, since the deformation fields in those cases are very small as well as age-related brain changes like small silent infarcts, enlarged perivascular spaces, cortical atrophy, white matter changes, and microbleed were absent. An interesting result was that the neuroradiologist tends to think that real images from GENIC look synthetic. Something that we have to consider is that the real images were preprocessed with Freesurfer (e.g., they are bias-corrected and skull-stripped), which can make the visual assessment of the neuroradiologist slightly different compared with everyday clinical praxis in a neuroradiology department.

It is worth mentioning that in Ravi et al. (2022), the same discrimination task was introduced. They found that neuroradiologists can achieve an accuracy of 68.0 ± 7.1% on the synthetic images generated by their method, while our method can achieve a 64% accuracy on the generated images. Although these numbers cannot be directly compared, this indicates that our performances might at least be comparable to the results in Ravi et al. (2022).

5.4 | Age estimation

Based on the literature, we assumed that the brain changes due to age were linear. Such an assumption was supported by the regression analysis of Figure 7. After correcting the age estimation with the fitted lines, the age estimation was just slightly better (cf. Table 4 and Figure 6). Notice that the research community has been very active in estimating age from images (Lund et al., 2022; Sajedi & Pardakhti, 2019). Notice that the current best-performing method in the state-of-the-art for these methods has a MAE in the order of 2.13 years (Bintsi et al., 2020; Dartora et al., 2022), that is, not too different from our results. Thus, we expect similar results if we change the linear model by a brain age estimation method. An advantage of using the linear model is that we were able to assess the validity of that model, as shown in Figure 7. Beyond the use of different metrics (RMSE in our case and MAE in Bintsi et al., 2020), it is clear that it is not possible to compare the performance of the used linear method with the state-of-the-art of brain age estimation since the latter methods use only one, while we use two images from the same subject (the youngest and oldest). It might be possible to extend brain age estimation methods by having more than one image as an input. However, that is beyond the aim of the paper.

5.5 | Dependency on training data

An important feature of Synthmorph is that it was trained on brain parcellations instead of raw images. This makes it less sensitive to the different imaging characteristics of the different scanners and datasets compared to GAN- or VAE-based methods. Still, we found a connection between the stopping point *s* and the dataset, as shown in Figure 5. Our hypothesis is that rather than the imaging statistics of a dataset, its demographical characteristics can influence *s*. The results of Table 2 align with this hypothesis. Notice that adjusting the image generation with the stopping point *s* removes the dependency of the method with the dataset.

5.6 | Limitations

As stated, the current method aims to fill up the missing data and augment the current longitudinal cohorts with high-quality scans in 3D. In order to avoid sources of uncertainty and to guarantee the quality of the images, we limited the generation of images to intrasubject paired inputs. Thus, we did not investigate other alternatives for generating images. For example, although the generation of images from intersubject pairs is technically possible, it would be difficult to assess whether the changes in the images are due to aging or to the morphological difference between subjects. Also, we only focused on healthy subjects in order to avoid disease-related factors that can be confounded with normal aging. Moreover, the plausibility of images synthesized beyond the oldest image (i.e., beyond the stopping point *s*) has not been investigated and cannot be guaranteed. Besides, due to practical obstacles such as focusing on 2D or the absence of code, the comparison with previous brain image generation methods has not been conducted. Instead, the source codes generated for this study are provided to other researchers for comparison in the future.

One important limitation of the method is that it requires two images, unlike generative approaches such as GANs and VAEs. Our current research is focused on removing that restriction to generate images from a single image. An additional limitation is that the imaging statistics of the synthetic images are expected to be similar to the ones used for performing the registration. However, this limitation might be tackled by using domain adaptation methods.

Finally, the proposed method can suffer from identity leakage. It has been shown that the cortical and subcortical surfaces can be used to identify subjects (Wachinger et al., 2015). Thus, by using the proposed method, it is potentially possible to find the subject used for image generation from a synthetic image since the method relies on diffeomorphic registration. Although this issue is not a problem for public datasets, this privacy issue is something that should be considered if the proposed method is used to generate synthetic images from private datasets.

5.7 | Future work

Beside tackling the mentioned limitations, we foresee many potential applications and improvement directions for the proposed method. First, as mentioned, the method can be used to synthesize high-quality and high-resolution images from two images from different modalities by only changing to a suitable similarity measurement. Second, our method could be applied to estimate the progression of neurodegenerative brain diseases, not just for normal aging. For example, with our methodology, we can synthesize subject-specific temporal estimations of undergoing neurodegeneration, which can then be compared with the healthy templates to provide cross-sectional comparisons that shall aid clinical diagnoses. Third, the augmented longitudinal data as well as the corresponding segmentations can be used in the training phase of ML-based segmentation or classification tasks, or as a reliable reference to validate or interpret AI-enabled models.

6 | CONCLUSION

In this work, we proposed a methodology with the aim of simulating subject-specific aging in brain MRIs given two 3D images acquired at different time points. DL-based diffeomorphic registration was used as a backbone to generate deformation fields at different integration points. Quality measurements were used for controlling the age estimation of the generated images by using a linear assumption. The results show good performance from both quantitative and qualitative perspectives regarding both the image quality of the synthetic MRI scans and the estimation of age.

AUTHOR CONTRIBUTIONS

Jingru Fu: conceptualization; formal analysis; investigation; methodology; software; validation; visualization; writing—original draft; writing—review and editing. **Antonios Tzortzakakis:** validation; investigation; writing—review and editing. **José Barroso:** resources; data curation; writing—review and editing. **Eric Westman:** resources; data curation; writing—review and editing. **Daniel Ferreira:** conceptualization; methodology; resources; data curation; writing—review and editing; supervision; funding acquisition. **Rodrigo Moreno:** conceptualization; methodology; visualization; resources; project administration; writing—review and editing; supervision; funding acquisition.

ACKNOWLEDGMENTS

Data were provided in part by ADNI. Data collection and sharing for this project were 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 organization 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. Data were provided in part by OASIS-3: principal investigators: T. Benzinger, D. Marcus, J. Morris; NIH P50 AG00561, P30 NS09857781, P01 AG026276, P01 AG003991, R01 AG043434, UL1 TR000448, R01 EB009352. AV-45 doses were provided by Avid Radiopharmaceuticals, a wholly owned subsidiary of Eli Lilly. The authors would like to thank Dr Antonio Rodríguez for providing access to participants in GENIC and the Servicio de Resonancia Magnética para Investigaciones Biomédicas del SEGAI (University of La Laguna, Spain). Data used in preparation of this article are part of the GENIC database (Group of Neuropsychological Studies of the Canary Islands, University of La Laguna, Spain; principal investigator: Professor José Barroso; contact: Dr Daniel Ferreira, daniel.ferreira.padilla@ki.se). The following collaborators contributed to the GENIC database but did not participate in analysis or writing of this report (in alphabetic order by family name): Nira Cedrés, Rut Correia, Patricia Diaz, Aida Figueroa, Nerea Figueroa, Eloy García, Lissett González, Teodoro González, Zaira González, Cathaysa Hernández, Edith Hernández, Nira Jiménez, Judith López, Cándida Lozano, Alejandra Machado, María Antonieta Nieto, María Sabucedo, Elena Sirumal, Marta Suárez, Manuel Urbano, and Pedro Velasco.

FUNDING INFORMATION

This study has been partially funded by Vinnova through AIDA, project ID: 2108, by the China Scholarship Council (CSC) for PhD studies at KTH Royal Institute of Technology, by Barncancerfonden (MT2019-0019), by Digital Futures, project dBrain, by the Center for Innovative Medicine (CIMED), the regional agreement on medical training and clinical research (ALF) between Stockholm County Council and Karolinska Institutet, Hjärnfonden, Alzheimerfonden, Demensfonden, Neurofonden, Stiftelsen För Gamla Tjänarinnor, Fundación Canaria Dr Manuel Morales (calls in 2012, 2014, and 2017); Fundación Cajacanarias; and the Estrategia de Especialización Inteligente de Canarias RIS3 from Consejería de Economía, Industria, Comercio y Conocimiento del Gobierno de Canarias, co-funded by the Programa Operativo FEDER Canarias 2014–2020 (ProID2020010063). The funders of the study had no role in the study design nor the collection, analysis, and interpretation of data, writing of the report, or decision to submit the manuscript for publication.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in ADNI, OASIS and GENIC at http://doi.org/10.1002/jmri.21049, http://doi.org/10.1101/2019.12.13.19014902 and https://doi.org/10.1016/j.neurobiolaging.2018.07.017, reference number (Jack et al., 2008; LaMontagne et al., 2019; Machado et al., 2018). The augmented 3D MRI scans are publicly available on https://github.com/Fjr9516/Synthetic-Brain-Aging/blob/main/README.md. The source codes generated for this study are available on https://github.com/Fjr9516/Synthetic-Brain-Aging.

ORCID

Jingru Fu https://orcid.org/0000-0003-4175-395X
Antonios Tzortzakakis https://orcid.org/0000-0001-7563-732X
Eric Westman https://orcid.org/0000-0002-3115-2977
Daniel Ferreira https://orcid.org/0000-0001-9522-4338
Rodrigo Moreno https://orcid.org/0000-0001-5765-2964

REFERENCES

Alberdi, A., Aztiria, A., & Basarab, A. (2016). On the early diagnosis of Alzheimer's disease from multimodal signals: A survey. *Artificial Intelligence in Medicine*, *71*, 1–29. https://doi.org/10.1016/j.artmed.2016.06.003

Anatürk, M., Kaufmann, T., Cole, J. H., Suri, S., Griffanti, L., Zsoldos, E., Filippini, N., Singh-Manoux, A., Kivimäki, M., Westlye, L. T., Ebmeier, K. P., & Lange, A.-M. G. (2021). Prediction of brain age and cognitive age: Quantifying brain and cognitive maintenance in aging. *Human Brain Mapping*, *42*(6), 1626–1640. https://doi.org/10.1002/hbm.25316

Arsigny, V., Commowick, O., Pennec, X., & Ayache, N. (2006). A log-Euclidean framework for statistics on diffeomorphisms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Vol. 9, pp. 924–931). Springer. https://doi.org/10.1007/11866565_113

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage*, *38*(1), 95–113. https://doi.org/10.1016/j.neuroimage.2007.07.007

Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., & Dalca, A. V. (2019). Voxelmorph: A learning framework for deformable medical image registration. *Institute of Electrical and Electronics Engineers Transactions on Medical Imaging*, *38*(8), 1788–1800. https://doi.org/10.1109/tmi.2019.2897538

Beg, M. F., Miller, M. I., Trouvé, A., & Younes, L. (2005). Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International Journal of Computer Vision*, *61*(2), 139–157. https://doi.org/10.1023/B:VISI.0000043755.93987.aa

Bintsi, K.-M., Baltatzis, V., Kolbeinsson, A., Hammers, A., & Rueckert, D. (2020). Patch-based brain age estimation from MR images. In *Machine learning in clinical neuroimaging and radiogenomics in neuro-oncology* (pp. 98–107). Springer. https://doi.org/10.1007/978-3-030-66843-3_10

Bowles, C., Gunn, R., Hammers, A., & Rueckert, D. (2018). Modelling the progression of Alzheimer's disease in MRI using generative adversarial networks. *Proceedings of the Medical Imaging: Image Processing*, *10574*, 105741K. https://doi.org/10.1117/12.2293256

Campbell, K. M., & Fletcher, P. T. (2017). Efficient parallel transport in the group of diffeomorphisms via reduction to the lie algebra. In *Graphs in biomedical image analysis, computational anatomy and imaging genetics* (pp. 186–198). Springer. https://doi.org/10.1007/978-3-319-67675-3_17

Chen, J., Yong, D., He, Y., Segars, W. P., Li, Y., & Frey, E. C. (2021). Transmorph: Transformer for unsupervised medical image registration. *arXiv preprint arXiv:2111.10480*. https://doi.org/10.48550/arXiv.2111.10480.

Choi, H., Kang, H., Lee, D. S., & Alzheimer's Disease Neuroimaging Initiative. (2018). Predicting aging of brain metabolic topography using variational autoencoder. *Frontiers in Aging Neuroscience*, *10*, 212. https://doi.org/10.3389/fnagi.2018.00212

Cole, J. H., Leech, R., Sharp, D. J., & Alzheimer's Disease Neuroimaging Initiative. (2015). Prediction of brain age suggests accelerated atrophy after traumatic brain injury. *Annals of Neurology*, *77*(4), 571–581. https://doi.org/10.1002/ana.24367

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, *35*(1), 53–65. https://doi.org/10.1109/MSP.2017.2765202

Cury, C., Durrleman, S., Cash, D. M., Lorenzi, M., Nicholas, J. M., Bocchetta, M., van Swieten, J. C., Borroni, B., Galimberti, D., Masellis, M., Tartaglia, M. C., Rowe, J. B., Graff, C., Tagliavini, F., Frisoni, G. B., Laforce, R., Finger, E., de Mendonça, A., Sorbi, S., … Tang-W, D. (2019). Spatiotemporal analysis for detection of pre-symptomatic shape changes in neurodegenerative diseases: Initial application to the GENFI cohort. *NeuroImage*, *188*, 282–290. https://doi.org/10.1016/j.neuroimage.2018.11.063

Cury, C., Lorenzi, M., Cash, D., Nicholas, J. M., Routier, A., Rohrer, J., Ourselin, S., Durrleman, S., & Modat, M. (2016). Spatio-temporal shape analysis of cross-sectional data for detection of early changes in neurodegenerative disease. In *International Workshop on Spectral and Shape Analysis in Medical Imaging* (pp. 63–75). Springer.

Dalca, A., Rakic, M., Guttag, J., & Sabuncu, M. (2019). Learning conditional deformable templates with convolutional networks. *Proceedings of the Neural Information Processing Systems*, *32*, 806–818. https://doi.org/10.48550/arXiv.1908.02738

Dalca, A. V., Balakrishnan, G., Guttag, J., & Sabuncu, M. R. (2019). Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical Image Analysis*, *57*, 226–236. https://doi.org/10.1016/j.media.2019.07.006

Dartora, C., Marseglia, A., Mårtensson, G., Rukh, G., Dang, J., Muehlboeck, J.-S., Wahlund, L.-O., Moreno, R., Barroso, J., Ferreira, D., Schiöth, H. B., Westman, E., the Alzheimer's Disease Neuroimaging Initiative, the Australian Imaging Biomarkers, & Lifestyle Flagship Study. (2022). Predicting the age of the brain with minimally processed T1-weighted MRI data. *medRxiv*. https://doi.org/10.1101/2022.09.06.22279594

de Vos, B. D., Berendsen, F. F., Viergever, M. A., Staring, M., & Išgum, I. (2017). End-to-end unsupervised deformable image registration with a convolutional neural network. In *Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (pp. 204–212). Springer. https://doi.org/10.1007/978-3-319-67558-9_24

Devic, S. (2012). MRI simulation for radiotherapy treatment planning. *Medical Physics*, *39*(11), 6701–6711. https://doi.org/10.1118/1.4758068

Dukart, J., Kherif, F., Mueller, K., Adaszewski, S., Schroeter, M. L., Frackowiak, R. S. J., Draganski, B., & Alzheimer's Disease Neuroimaging Initiative. (2013). Generative FDG-PET and MRI model of aging and disease progression in Alzheimer's disease. *Public Library of Science Computational Biology*, *9*(4), e1002987. https://doi.org/10.1371/journal.pcbi.1002987

Emami, H., Dong, M., Nejad-Davarani, S. P., & Glide-Hurst, C. K. (2018). Generating synthetic CTs from magnetic resonance images using generative adversarial networks. *Medical Physics*, *45*(8), 3627–3636. https://doi.org/10.1002/mp.13047

Fischl, B. (2012). Freesurfer. *NeuroImage*, *62*(2), 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021

Franke, K., Luders, E., May, A., Wilke, M., & Gaser, C. (2012). Brain maturation: Predicting individual BrainAGE in children and adolescents using structural MRI. *NeuroImage*, *63*(3), 1305–1312. https://doi.org/10.1016/j.neuroimage.2012.08.001

Fu, Y., Lei, Y., Wang, T., Curran, W. J., Liu, T., & Yang, X. (2020). Deep learning in medical image registration: A review. *Physics in Medicine & Biology*, *65*(20), 20TR01. https://doi.org/10.1088/1361-6560/ab843e

Giorgio, A., Santelli, L., Tomassini, V., Bosnell, R., Smith, S., De Stefano, N., & Johansen-Berg, H. (2010). Age-related changes in grey and white matter structure throughout adulthood. *NeuroImage*, *51*(3), 943–951. https://doi.org/10.1016/j.neuroimage.2010.03.004

Gu, X., Knutsson, H., Nilsson, M., & Eklund, A. (2019). Generating diffusion MRI scalar maps from T1 weighted images using generative adversarial networks. In *Scandinavian Conference on Image Analysis* (pp. 489–498). Springer. https://doi.org/10.1007/978-3-030-20205-7_40

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Proceedings of the Neural Information Processing Systems*, *30*, 6629–6640. https://doi.org/10.48550/arXiv.1706.08500

Hoffmann, M., Billot, B., Greve, D. N., Iglesias, J. E., Fischl, B., & Dalca, A. V. (2022). Synthmorph: Learning contrast-invariant registration without acquired images. *Institute of Electrical and Electronics Engineers Transactions on Medical Imaging*, *41*(3), 543–558. https://doi.org/10.1109/TMI.2021.3116879

Huizinga, W., Poot, D. H. J., Vernooij, M. W., Roshchupkin, G. V., Bron, E. E., Ikram, M. A., Rueckert, D., Niessen, W. J., Klein, S., & Alzheimer's Disease Neuroimaging Initiative. (2018). A spatio-temporal reference model of the aging brain. *NeuroImage*, *169*, 11–22. https://doi.org/10.1016/j.neuroimage.2017.10.040

Jack Jr, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., Whitwell, J. L., Ward, C., Dale, A. M., Felmlee, J. P., Gunter, J. L., Hill, D. L., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., … Weiner, M. W. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, *27*(4), 685–691. https://doi.org/10.1002/jmri.21049

Jung, E., Luna, M., & Park, S. H. (2021). Conditional GAN with an attention-based generator and a 3D discriminator for 3D medical image generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 318–328). Springer. /10.1007/978-3-030-87231-1_31

Kim, S. T., Küçükaslan, U., & Navab, N. (2021). Longitudinal brain MR image modeling using personalized memory for Alzheimer's disease. *Institute of Electrical and Electronics Engineers Access*, *9*, 143212–143221. https://doi.org/10.1109/ACCESS.2021.3121609

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*. https://doi.org/10.48550/arXiv.1312.6114.

Krebs, J., Delingette, H., Mailhé, B., Ayache, N., & Mansi, T. (2019). Learning a probabilistic model for diffeomorphic registration. *Institute of Electrical and Electronics Engineers Transactions on Medical Imaging*, *38*(9), 2165–2176. https://doi.org/10.1109/TMI.2019.2897112

Krebs, J., Mansi, T., Mailhé, B., Ayache, N., & Delingette, H. (2018). Unsupervised probabilistic deformation modeling for robust diffeomorphic registration. In *Proceedings of the deep learning in medical image analysis and multimodal learning for clinical decision support* (pp. 101–109). Springer-Verlag. https://doi.org/10.1007/978-3-030-00889-5_12

LaMontagne, P. J., Benzinger, T. L. S., Morris, J. C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A., et al. (2019). OASIS-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *MedRxiv*. https://doi.org/10.1101/2019.12.13.19014902

Lei, Y., Harms, J., Wang, T., Liu, Y., Shu, H.-K., Jani, A. B., Curran, W. J., Mao, H., Liu, T., & Yang, X. (2019). MRI-only based synthetic CT generation using dense cycle consistent generative adversarial networks. *Medical Physics*, *46*(8), 3565–3581. https://doi.org/10.1002/mp.13617

Li, H., & Fan, Y. (2022). Mdreg-net: Multi-resolution diffeomorphic image registration using fully convolutional networks with deep self-supervision. *Human Brain Mapping*, *43*, 1–14. https://doi.org/10.1002/hbm.25782

Lockhart, S. N., & DeCarli, C. (2014). Structural imaging measures of brain aging. *Neuropsychology Review*, *24*(3), 271–289. https://doi.org/10.1007/s11065-014-9268-3

Lorenzi, M., Pennec, X., Frisoni, G. B., Ayache, N., & Alzheimer's Disease Neuroimaging Initiative. (2015). Disentangling normal aging from Alzheimer's disease in structural magnetic resonance images. *Neurobiology of Aging*, *36*, S42–S52. https://doi.org/10.1016/j.neurobiolaging.2014.07.046

Lund, M. J., Alnæs, D., de Lange, A.-M. G., Andreassen, O. A., Westlye, L. T., & Kaufmann, T. (2022). Brain age prediction using fMRI network coupling in youths and associations with psychiatric symptoms. *NeuroImage: Clinical*, *33*, 102921. https://doi.org/10.1016/j.nicl.2021.102921

Ma, R., Kutchy, N. A., Chen, L., Meigs, D. D., & Hu, G. (2022). Primary cilia and ciliary signaling pathways in aging and age-related brain disorders. *Neurobiology of Disease*, *163*, 105607. https://doi.org/10.1016/j.nbd.2021.105607

MacDonald, M. E., & Pike, G. B. (2021). MRI of healthy brain aging: A review. *NMR in Biomedicine*, *34*(9), e4564. https://doi.org/10.1002/nbm.4564

Machado, A., Barroso, J., Molina, Y., Nieto, A., Díaz-Flores, L., Westman, E., & Ferreira, D. (2018). Proposal for a hierarchical, multidimensional, and multivariate approach to investigate cognitive aging. *Neurobiology of Aging*, *71*, 179–188. https://doi.org/10.1016/j.neurobiolaging.2018.07.017

Muehlboeck, J.-S., Westman, E., & Simmons, A. (2014). TheHiveDB image data management and analysis framework. *Frontiers in Neuroinformatics*, *7*, 49. https://doi.org/10.3389/fninf.2013.00049

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C. R., Jagust, W., Trojanowski, J. Q., Toga, A. W., & Beckett, L. (2005). Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's disease neuroimaging initiative (ADNI). *Alzheimer's & Dementia*, *1*(1), 55–66. https://doi.org/10.1016/j.jalz.2005.06.003

Nemy, M., Cedres, N., Grothe, M. J., Muehlboeck, J.-S., Lindberg, O., Nedelska, Z., Stepankova, O., Vyslouzilova, L., Eriksdotter, M., Barroso, J., Teipel, S., Westman, E., & Ferreira, D. (2020). Cholinergic white matter pathways make a stronger contribution to attention and memory in normal aging than cerebrovascular health and nucleus basalis of Meynert. *NeuroImage*, *211*, 116607. https://doi.org/10.1016/j.neuroimage.2020.116607

Niethammer, M., Huang, Y., & Vialard, F.-X. (2011). Geodesic regression for image time-series. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Vol. 14, pp. 655–662). Springer. https://doi.org/10.1007/978-3-642-23629-7_80

Njeh, C. F. (2008). Tumor delineation: The weakest link in the search for accuracy in radiotherapy. *Journal of Medical Physics*, *33*(4), 136. https://doi.org/10.4103/0971-6203.44472

Ouyang, J., Zhao, Q., Adeli, E., Sullivan, E. V., Pfefferbaum, A., Zaharchuk, G., & Pohl, K. M. (2021). Self-supervised longitudinal neighbourhood embedding. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Vol. 12902, pp. 80–89). Springer. https://doi.org/10.1007/978-3-030-87196-3_8

Oxtoby, N. P., & Alexander, D. C. (2017). Imaging plus x: Multimodal models of neurodegenerative disease. *Current Opinion in Neurology*, *30*(4), 371–379. https://doi.org/10.1097/WCO.0000000000000460

Pathan, S., & Hong, Y. (2018). Predictive image regression for longitudinal studies with missing data. In *1st Conference on Medical Imaging with Deep Learning (MIDL)*. https://doi.org/10.48550/arXiv.1808.07553

Peters, R. (2006). Ageing and the brain. *Postgraduate Medical Journal*, *82*(964), 84–88. https://doi.org/10.1136/pgmj.2005.036665

Pinaya, W. H. L., Tudosiu, P.-D., Dafflon, J., Da Costa, P. F., Fernandez, V., Nachev, P., Ourselin, S., & Cardoso, M. J. (2022). *Brain imaging generation with latent diffusion models* (pp. 117–126). Springer. https://doi.org/10.1007/978-3-031-18576-2_12

Popescu, S. G., Glocker, B., Sharp, D. J., & Cole, J. H. (2021). Local brain-age: A U-net model. *Frontiers in Aging Neuroscience*, *13*, 838. https://doi.org/10.3389/fnagi.2021.761954

Popescu, S. G., Whittington, A., Gunn, R. N., Matthews, P. M., Glocker, B., Sharp, D. J., Cole, J. H., & Alzheimer's Disease Neuroimaging Initiative. (2020). Nonlinear biomarker interactions in conversion from mild cognitive impairment to Alzheimer's disease. *Human Brain Mapping*, *41*(15), 4406–4418. https://doi.org/10.1002/hbm.25133

Poulakis, K., Ferreira, D., Pereira, J. B., Smedby, Ö., Vemuri, P., & Westman, E. (2020). Fully Bayesian longitudinal unsupervised learning for the assessment and visualization of AD heterogeneity and progression. *Aging (Albany NY)*, *12*(13), 12622–12647. https://doi.org/10.18632/aging.103623

Poulakis, K., Pereira, J. B., Muehlboeck, J., Wahlund, L.-O., Smedby, Ö., Volpe, G., Masters, C. L., Ames, D., Niimi, Y., Iwatsubo, T., Ferreira, D., Westman, E., AddNeuroMed Consortium and Group, Alzheimer's Disease Neuroimaging Initiative, & Japanese Alzheimer's Disease Neuroimaging Initiative. (2021). Stage vs. Subtype hypothesis in Alzheimer's disease: A multi-cohort and longitudinal Bayesian clustering study. *The Lancet*. https://doi.org/10.2139/ssrn.3797614

Ravi, D., Blumberg, S. B., Ingala, S., Barkhof, F., Alexander, D. C., Oxtoby, N. P., & Alzheimer's Disease Neuroimaging Initiative. (2022). Degenerative adversarial neuroimage nets for brain scan simulations: Application in ageing and dementia. *Medical Image Analysis*, *75*, 102257. https://doi.org/10.1016/j.media.2021.102257

Resnick, S. M., Pham, D. L., Kraut, M. A., Zonderman, A. B., & Davatzikos, C. (2003). Longitudinal magnetic resonance imaging studies of older adults: A shrinking brain. *Journal of Neuroscience*, *23*(8), 3295–3301. https://doi.org/10.1523/JNEUROSCI.23-08-03295.2003

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

Sajedi, H., & Pardakhti, N. (2019). Age prediction based on brain MRI image: A survey. *Journal of Medical Systems*, *43*(8), 1–30. https://doi.org/10.1007/s10916-019-1401-7

Schmidt, M. A., & Payne, G. S. (2015). Radiotherapy planning using MRI. *Physics in Medicine & Biology*, *60*(22), R323–R361. https://doi.org/10.1088/0031-9155/60/22/R323

Shampine, L. F. (1986). Some practical Runge–Kutta formulas. *Mathematics of Computation*, *46*(173), 135–150. https://doi.org/10.1090/S0025-5718-1986-0815836-3

Song, L., Li, H., & Fan, J. (2022). Longitudinal structural MRI data prediction in nondemented and demented older adults via generative adversarial convolutional network. *Neural Processing Letters*. https://doi.org/10.1007/s11063-022-10922-6

Treder, M. S., Codrai, R., & Tsvetanov, K. A. (2022). Quality assessment of anatomical MRI images from generative adversarial networks: Human assessment and image quality metrics. *Journal of Neuroscience Methods*, *374*, 109579. https://doi.org/10.1016/j.jneumeth.2022.109579

Tudosiu, P.-D., Varsavsky, T., Shaw, R., Graham, M., Nachev, P., Ourselin, S., Sudre, C. H., & Cardoso, M. J. (2020). Neuromorphologicaly-preserving volumetric data encoding using vq-vae. *arXiv preprint arXiv:2002.05692*. https://doi.org/10.48550/arXiv.2002.05692.

Van Loan, C. F., & Golub, G. (2013). *Matrix computations* (4th ed.). The Johns Hopkins University Press.

Wachinger, C., Golland, P., Kremen, W., Fischl, B., & Reuter, M. (2015). Brainprint: A discriminative characterization of brain morphology. *NeuroImage*, *109*, 232–248. https://doi.org/10.1016/j.neuroimage.2015.01.032

Walhovd, K. B., Fjell, A. M., Reinvang, I., Lundervold, A., Dale, A. M., Eilertsen, D. E., Quinn, B. T., Salat, D., Makris, N., & Fischl, B. (2005). Effects of age on volumes of cortex, white matter and subcortical structures. *Neurobiology of Aging*, *26*(9), 1261–1270. https://doi.org/10.1016/j.neurobiolaging.2005.05.020

Wegmayr, V., Hörold, M., & Buhmann, J. M. (2019). Generative aging of brain MR-images and prediction of Alzheimer progression. In *German Conference on Pattern Recognition. Lecture notes in computer science* (pp. 247–260). Springer International Publishing. https://doi.org/10.1007/978-3-030-33676-9_17

Zitova, B., & Flusser, J. (2003). Image registration methods: A survey. *Image and Vision Computing*, *21*(11), 977–1000. https://doi.org/10.1016/S0262-8856(03)00137-9

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Fu, J., Tzortzakakis, A., Barroso, J., Westman, E., Ferreira, D., Moreno, R., & for the Alzheimer's Disease Neuroimaging Initiative (2023). Fast three-dimensional image generation for healthy brain aging using diffeomorphic registration. *Human Brain Mapping*, *44*(4), 1289–1308. https://doi.org/10.1002/hbm.26165