Corresponding author:  
Nicholas J. Tustison, DSc  
Department of Radiology and Medical Imaging  
University of Virginia  
[ntustison@virginia.edu](mailto:ntustison@virginia.edu)

Data used in preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (<http://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/AD NI Acknowledgement List.pdf

# Abstract

The Advanced Normalizations Tools ecosystem, known as ANTsX, consists of multiple open-source software libraries which house top-performing algorithms used worldwide by scientific and research communities for processing and analyzing biological and medical imaging data. The base software library, ANTs, is built upon, and contributes to, the NIH-sponsored Insight Toolkit. Founded in 2008 with the highly regarded Symmetric Normalization image registration framework, the ANTs library has since grown to include additional functionality. Recent enhancements include statistical, visualization, and deep learning capabilities through interfacing with both the R statistical project (ANTsR) and Python (ANTsPy). Additionally, the corresponding deep learning extensions ANTsRNet and ANTsPyNet (built on the popular TensorFlow/Keras libraries) contain several popular network architectures and trained models for specific applications. One such comprehensive application is a deep learning analog for generating cortical thickness data from structural T1-weighted brain MRI. Not only does this significantly improve computational efficiency and provide comparable-to-superior accuracy the existing ANTs pipeline but it also illustrates the importance of the comprehensive ANTsX approach as a framework for medical image analysis.

# The ANTsX ecosystem: A brief overview

## Image registration origins

The Advanced Normalization Tools (ANTs) is a state-of-the-art, open-source software toolkit for image registration, segmentation, and other functionality for comprehensive biological and medical image analysis. Historically, ANTs is rooted in advanced image registration techniques which have been at the forefront of the field due to seminal contributions that date back to the original elastic matching method of Bajcsy and co-investigators1–3. Various independent platforms have been used to evaluate ANTs tools since their early development. In a landmark paper4, the authors reported an extensive evaluation using multiple neuroimaging datasets analyzed by fourteen different registration tools, including the Symmetric Normalization (SyN) algorithm5, and found that “ART, SyN, IRTK, and SPM’s DARTEL Toolbox gave the best results according to overlap and distance measures, with ART and SyN delivering the most consistently high accuracy across subjects and label sets.” 6,7 8,

## Current developments

Since its inception, though, ANTs has expanded significantly beyond its image registration origins. Other core contributions include template building9, segmentation10, image preprocessing (e.g., bias correction11 and denoising12), joint label fusion13,14, and brain cortical thickness estimation15,16 (cf Table ). Additionally, ANTs has been integrated into multiple, publicly available workflows such as fMRIprep17 and the Spinal Cord Toolbox18. Frequently used ANTs pipelines, such as cortical thickness estimation16, have been integrated into Docker containers and packaged as Brain Imaging Data Structure (BIDS)19 and FlyWheel applications (i.e., ``gears’’). It has also been independently ported for various platforms including Neurodebian20 (Debian OS), Neuroconductor21 (the R statistical project), and Nipype22 (Python). Even competing softwares, such as FreeSurfer23, have incorporated well-performing and complementary ANTs components11,12 into their own libraries.

Over the course of its development, ANTs has been extended to complementary frameworks resulting in the Python- and R-based ANTsPy and ANTsR toolkits, respectively. These ANTs-based interfaces with extremely popular, high-level, open-source programming platforms have significantly increased the user base of ANTs and facilitated research workflows The rapidly rising popularity of deep learning motivated further recent enhancement of ANTs and its extensions. Despite the existence of an abundance of online innovation and code for deep learning algorithms, much of it is disorganized and lacks a uniformity in structure and external data interfaces which would facilitate greater uptake. With this in mind, ANTsR spawned the deep learning ANTsRNet package which is a growing Keras/TensorFlow-based library of popular deep learning architectures and applications specifically geared towards medical imaging. Analogously, ANTsPyNet is an additional ANTsX complement to ANTsPy. Both, which we collectively refer to as “ANTsXNet”, are co-developed so as to ensure cross-compatibility such that training performed in one library is readily accessible by the other library. In addition to a variety of popular network architectures (which are implemented in both 2-D and 3-D), ANTsXNet contains a host of functionality for medical image analysis that have been developed in-house and collected from other open-source projects. For example, an extremely popular ANTsXNet application is a multi-modal brain extraction tool that uses different variants of the popular U-net24 architecture for segmenting the brain in multiple modalities. These modalities include conventional T1-weighted structural MRI as well as T2-weighted MRI, FLAIR, fractional anisotropy and BOLD. Demographic specialization also includes infant T1-weighted and/or T2-weighted MRI. Additionally, we have included other models and weights into our libraries such as a recent BrainAGE estimation model25, based on individuals; HippMapp3r26, a hippocampal segmentation tool; the winning entry of the MICCAI 2017 white matter hyperintensity segmentation competition27; MRI super resolution using deep-projection networks28; and NoBrainer, a T1-weighted brain extraction approach based on FreeSurfer (see Figure ).

## The ANTsXNet cortical thickness pipeline

The most recent ANTsX of our popular ANTs cortical thickness 16 29 pipeline within the ANTsXNet framework for, amongst other potential benefits, increased computational efficiency. , 16, 30 10

structural processing pipeline currently available as open-source within the ANTsXNet libraries which underwent a thorough evaluation using both cross-sectional and longitudinal data and discussed within the context of our previous evaluations16,29. Note that related work has been recently reported by external groups31,32. Fortunately, these overlapping contributions provide a context for comparison to motivate the utility of the ANTsX ecosystem.

# Results

## The original ANTs cortical thickness pipeline

The original ANTs cortical thickness pipeline16 consists of the following steps:

* preprocessing: denoising12 and bias correction33;
* brain extraction30;
* brain segmentation with spatial tissue priors10 comprising the
  + cerebrospinal fluid (CSF),
  + gray matter (GM),
  + white matter (WM),
  + deep gray matter,
  + cerebellum, and
  + brain stem; and
* cortical thickness estimation15.

Our recent longitudinal variant incorporates an additional step involving the construction of a single subject template (SST)9 coupled with the generation of tissue spatial priors of the SST for use with the processing of the individual time points as described above.

Although the resulting thickness maps are conducive to voxel-based34 and related analyses35, here we employ the well-known Desikan-Killiany-Tourville (DKT)36 labeling protocol (31 labels per hemisphere) to parcellate the cortex for averaging thickness values regionally. This allows us to 1) be consistent in our evaluation strategy for comparison with our previous work16,29 and 2) leverage an additional deep learning-based substitution within the proposed pipeline.

## Overview of cortical thickness via ANTsXNet

Note that the entire analysis/evaluation framework, from preprocessing to statistical analysis, is made possible through the ANTsX ecosystem and simplified through the open-source R and Python platforms. Preprocessing, image registration, and cortical thickness estimation are all available through the ANTsPy and ANTsR libraries whereas the deep learning steps are performed through networks constructed and trained via ANTsRNet/ANTsPyNet with data augmentation strategies and other utilities built from ANTsR/ANTsPy functionality.

The brain extraction, brain segmentation, and DKT parcellation deep learning components were trained using data derived from our previous work16. Specifically, the IXI37, MMRR38, NKI39, and OASIS40 data sets, and the corresponding derived data, comprising over 1200 subjects from age 4 to 94, were used for network training. Brain extraction employs a traditional 3-D U-net network24 with whole brain, template-based data augmentation41 whereas brain segmentation and DKT parcellation are processed via 3-D U-net networks with attention gating42 on image octant-based batches. We emphasize that a single model (27) was created for each of these steps and was used for all the experiments described below.

## Cross-sectional performance evaluation

Due to the absence of ground-truth, we utilize the evaluation strategy from our previous work16 where we used cross-validation to build and compare age prediction models from data derived from both the proposed ANTsXNet pipeline and the established ANTs pipeline. Specifically, we use “age” as a well-known and widely-available demographic correlate of cortical thickness43 and quantify the predictive capabilities of corresponding random forest classifiers44 of the form: with covariates and (i.e., total intracranial volume). is the average thickness value in the DKT region. Root mean square error (RMSE) between the actual and predicted ages are the quantity used for comparative evaluation. As we have explained previously16, we find these evaluation measures to be much more useful than other commonly applied criteria as they are closer to assessing the actual utility of these thickness measurements as biomarkers for disease45 or growth. For example, in recent work31 the authors employ correlation with FreeSurfer thickness values as the primary evaluation for assessing relative performance with ANTs cortical thickness16. This evaluation, unfortunately, is fundamentally flawed in that it is a prime example of a type of circularity analysis46 whereby data selection is driven by the same criteria used to evaluate performance. Specifically, the underying DeepSCAN network used for the tissue segmentation step employs training based on FreeSurfer results which directly influences thickness values as thickness/segmentation are highly correlated and vary characteristically between software packages. Relative performance with ANTs thickness (which does not use FreeSurfer for training) is then assessed by determining correlations with FreeSurfer thickness values. Almost as problematic is their use of repeatability, which they confusingly label as “robustness,” as an additional ranking criterion. Repeatability evaluations should be contextualized within considerations such as the bias-variance tradeoff and quantified using relevant metrics, such as the intra-class correlation coefficient which takes into account both inter- and intra-observer variability.

In addition to the training data listed above, to ensure generalizability, we also compared performance using the SRPB data set47 comprising over 1600 participants from 12 sites. Note that we recognize that we are processing a portion of the evaluation data through certain components of the proposed deep learning-based pipeline that were used to train the same pipeline components. Although this does not provide evidence for generalizability (which is why we include the much larger SRPB data set), it is still interesting to examine the results since, in this case, the deep learning training can be considered a type of noise reduction on the final results. It should be noted that training did not use age prediction (or any other evaluation or related measure) as a criterion to be optimized during network model training (i.e., circular analysis46).

The results are shown in Figure where we used cross-validation with 500 permutations per model per data set (including a “combined” set) and an 80/20 training/testing split. The ANTsXNet deep learning pipeline outperformed the classical pipeline16 in terms of age prediction in all data sets except for IXI. This also includes the cross-validation iteration where all data sets were combined. Importance plots ranking the cortical thickness regions and the other covariates of Equation (1) are shown in Figure . Rankings employ “MeanDecreaseAccuracy” which quantifies the decrease in model accuracy based on the exclusion of a specific random forest regressor. Additionally, repeatability assessment on the MMRR data set yielded ICC values (“average random rater”) of 0.99 for both pipelines.

## Longitudinal performance evaluation

Given the excellent performance and superior computational efficiency of the proposed ANTsXNet pipeline for cross-sectional data, we evaluated its performance on longitudinal data using the longitudinally-specific evaluation strategy and data we employed with the introduction of the longitudinal version of the ANTs cortical thickness pipeline29. 1014.

The ADNI-1 data used for our previous longitudinal performance evaluation29 consisted of over 600 subjects (197 cognitive normals, 324 LMCI subjects, and 142 AD subjects) with one or more follow-up image acquisition sessions every 6 months (up to 36 months) for a total of over 2500 images. In addition to the ANTsXNet pipelines for the current evaluation, our previous work included the FreeSurfer23 cross-sectional (“FSCross”) and longitudinal (“FSLong”) streams, the ANTs cross-sectional pipeline (“ANTsCross”) in addition to two longitudinal ANTs-based variants (“ANTsNative” and “ANTsSST”). Two evaluation measurements, one unsupervised and one supervised, were used to assess comparative performance between all five pipelines. We add the results of the ANTsXNet pipeline evaluations in relation to these other pipelines to provide a comprehensive overview of relative performance.

First, the supervised evaluation employed Tukey post-hoc analyses with false discovery rate (FDR) adjustment to test the significance of the LMCI-CN, AD-LMCI, and AD-CN diagnostic contrasts. This is provided by the following LME model Here, is the change in thickness of the DKT region from baseline (bl) thickness with random intercepts for both the individual subject () and the acquisition site. The subject-specific covariates , status, , , , and were taken directly from the ADNIMERGE package.

Second, LME48 modeling was used to quantify between-subject and residual variabilities, the ratio of which provides an estimate of the effectiveness of a given biomarker for distinguishing between subpopulations. In order to assess this criteria while accounting for changes that may occur through the passage of time, we used the following Bayesian LME model: where denotes the individual’s cortical thickness measurement corresponding to the region of interest at the time point indexed by and specification of variance priors to half-Cauchy distributions reflects commonly accepted best practice in the context of hierarchical models49. The ratio of interest, , per region of the between-subject variability, , and residual variability, is where the posterior distribution of was summarized via the posterior median.

Results for both longitudinal evaluation scenarios are shown in Figure . Log p-values are provided in Figure (a) which demonstrate excellent LMCI-CN and AD-CN differentiation and comparable AD-LMCI diffierentiation relative to the other pipelines. Figure (b) shows significantly better performance for the longitudinal ANTsXNet pipeline where, in a longitudinal setting, we prefer to see lower values for residual variability and higher values for between-subject variability, leading to a larger variance ratio. In contrast, cross-sectional ANTsXNet performs remarkably poorly for these measures.

# Discussion

The ANTsX software ecosystem provides a comprehensive framework for quantitative biological and medical imaging. Although ANTs, the original core of ANTsX, is still at the forefront of image registration technology, it has moved signicantly beyond its image registration origins. This expansion is not confined to technical contributions (of which there are many) but also consists of facilitating access to a wide range of users who can use ANTsX tools (whether through bash, Python, or R scripting) to construct tailored pipelines for their own studies or to take advantage of our pre-fabricated pipelines. And given the open-source nature of the ANTsX software, usage is not limited, for example, to academic institutions—a common constraint characteristic of other packages.

One of our most widely used pipelines is the estimation of cortical thickness from neuroimaging. This is understandable given the widespread usage of regional cortical thickness as a biomarker for developmental or pathological trajectories of the brain. In this work, we used this well-vetted ANTs tool to provide training data for producing alternative variants which leverage deep learning for improved computational efficiency and also provides superior performance with respect to previously proposed evaluation measures for both cross-sectional16 and longitudinal scenarios29. In addition to providing the tools which generated the original training data for the proposed ANTsXNet pipeline, the ANTsX ecosystem provides a full-featured platform for the additional steps such as preprocessing (ANTsR/ANTsPy); data augmentation (ANTsR/ANTsPy); network construction and training (ANTsRNet/ANTsPyNet); and visualization and statistical analysis of the results (ANTsR/ANTsPy).

In contrast, related work31 described and evaluated a similar thickness measurement pipeline. However, due to the lack of a complete processing and analysis framework, training data was generated using the FreeSurfer stream, deep learning-based brain segmentation employed DeepSCAN50 (in-house software), and cortical thickness estimation15 was generated using the ANTs toolkit.

In terms of future work, the recent surge and utility of deep learning in medical image analysis has significantly guided the areas of active ANTsX development. As demonstrated in this work with our widely used cortical thickness pipelines, there are many potential benefits of deep learning analogs to existing ANTs tools as well as the development of new ones. Performance is comparable-to-superior relative to existing pipelines depending on the evaluation metric. We see possible additional longitudinal extensions incorporating subject ID and months as additional network inputs.

# Methods

Software, average DKT regional thickness values for all data sets, and the scripts to perform both the analysis and obtain thickness values for a single subject are provided as open-source. Specifically, all the ANTsX libraries are hosted on GitHub (<https://github.com/ANTsX>). The cross-sectional data and analysis code are available as .csv files and R scripts at the GitHub repository dedicated to this paper (<https://github.com/ntustison/PaperANTsX>) whereas the longitudinal data and evaluation scripts are organized with the repository associated with our previous work29 (<https://github.com/ntustison/CrossLong>).

## Implementation

In Listing 1, we show the ANTsPy/ANTsPyNet code snippet for cross-sectional processing a single subject which starts with reading the T1-weighted MRI input image, through the generation of the Atropos-style six-tissue segmentation and probability images, application of ants.kelly\_kapowski (i.e., DiReCT), DKT cortical parcellation, subsequent label propagation through the cortex, and, finally, regional cortical thickness tabulation. antspynet.cortical\_thickness and antspynet.longitudinal\_cortical\_thickness, Note that there are precise, line-by-line R-based analogs available through ANTsR/ANTsRNet.

Both the ants.deep\_atropos and antspynet.desikan\_killiany\_tourville\_labeling functions perform brain extraction using the antspynet.brain\_extraction function. Internally, antspynet.brain\_extraction contains the requisite code to build the network and assign the appropriate hyperparameters. The model weights are automatically downloaded from the online hosting site <https://figshare.com> (see the function get\_pretrained\_network in ANTsPyNet or getPretrainedNetwork in ANTsRNet for links to all models and weights) and loaded to the constructed network. antspynet.brain\_extraction performs a quick translation transformation to a specific template (also downloaded automatically) using the centers of intensity mass, a common alignment initialization strategy. This is to ensure proper gross orientation. Following brain extraction, preprocessing for the other two deep learning components includes ants.denoise\_image and ants.n4\_bias\_correction and an affine-based reorientation to a version of the MNI template51.

We recognize the presence of some redundancy due to the repeated application of certain preprocessing steps. Thus, each function has a do\_preprocessing option to eliminate this redundancy for knowledgeable users but, for simplicity in presentation purposes, we do not provide this modified pipeline here. Although it should be noted that the time difference is minimal considering the longer time required by ants.kelly\_kapowski. ants.deep\_atropos returns the segmentation image as well as the posterior probability maps for each tissue type listed previously. antspynet.desikan\_killiany\_tourville\_labeling returns only the segmentation label image which includes not only the 62 cortical labels but the remaining labels as well. The label numbers and corresponding structure names are given in the program description/help. Because the DKT parcellation will, in general, not exactly coincide with the non-zero voxels of the resulting cortical thickness maps, we perform a label propagation step to ensure the entire cortex, and only the non-zero thickness values in the cortex, are included in the tabulated regional values.

antspynet.longitudinal\_cortical\_thickness, ants.deep\_atropos ants.atropos ants.kelly\_kapowski

ants.kelly\_kapowski quick antspynet.desikan\_killiany\_tourville\_labeling,

## Training details

Training differed slightly between models and so we provide details for each of these components below. For all training, we used ANTsRNet scripts and custom batch generators. Although the network construction and other functionality is available in both ANTsPyNet and ANTsRNet (as is model weights compatibility), we have not written such custom batch generators for the former (although this is on our to-do list). In terms of hardware, all training was done on a DGX (GPUs: 4X Tesla V100, system memory: 256 GB LRDIMM DDR4).

**T1-weighted brain extraction.** A whole-image 3-D U-net model24 was used in conjunction with multiple training sessions employing a Dice loss function followed by categorical cross entropy. 30. A center-of-mass-based transformation to a standard template was used to standardize such parameters as orientation and voxel size. However, to account for possible different header orientations of input data, a template-based data augmentation scheme was used41 whereby forward and inverse transforms are used to randomly warp batch images between members of the training population (followed by reorientation to the standard template). A digital random coin flipping for possible histogram matching52 between source and target images further increased data augmentation. Although not detailed here, training for brain extraction in other modalities was performed similarly.

**Deep Atropos.** Dealing with 3-D data presents unique barriers for training that are often unique to medical imaging. Various strategies are employed such as minimizing the number of layers and/or the number of filters at the base layer of the U-net architecture (as we do for brian extraction). However, we found this to be too limiting for capturing certain brain structures such as the cortex. 2-D and 2.5-D approaches are often used with varying levels of success but we also found better performance using full 3-D information. This led us to try randomly selected 3-D patches of various sizes. However, for both the six-tissue segmentations and DKT parcellations, we found that an octant-based patch strategy yielded the desired results. Specifically, after a brain extracted affine normalization to the MNI template, the normalized image is cropped to a size of [160, 190, 160]. Overlapping octant patches of size [112, 112, 112] were extracted from each image and trained using a batch size of 12 such octant patches with weighted categorical cross entropy as the loss function. As we point out in our earlier work16, obtaining proper brain segmentation is perhaps the most critical step to estimating thickness values that have the greatest utility as a potential biomarker. In fact, the first and last authors (NT and BA, respectively) spent much time during the original ANTs pipeline development16 trying to get the segmentation correct which required manually looking at many images and manually adjusting where necessary. This fine-tuning is often omitted or not considered when other groups31,53,54 use components of our cortical thickness pipeline which can be potentially problematic55. Fine-tuning for this particular workflow was also performed between the first and last authors using manual variation of the weights in the weighted categorical cross entropy. Ultimately, we settled on a weight vector of for the CSF, GM, WM, Deep GM, brain stem, and cerebellum, respectively. Other hyperparameters can be directly inferred from explicit specification in the actual code. As mentioned previously, training data was derived from application of the ANTs Atropos segmentation10 during the course of our previous work16. Data augmentation included small affine and deformable perturbations using antspynet.randomly\_transform\_image\_data and random contralateral flips.

**Desikan-Killiany-Tourville parcellation.** Preprocessing for the DKT parcellation training was similar to the Deep Atropos training. However, the number of labels and the complexity of the parcellation required deviation from other training steps. First, labeling was split into an inner set and an outer set. Subsequent training was performed separately for both of these sets. For the cortical labels, a set of corresponding input prior probability maps were constructed from the training data (and are also available and automatically downloaded, when needed, from <https://figshare.com>). Training occurred over multiple sessions where, initially, categorical cross entropy was used and then subsquently refined using a Dice loss function. Whole-brain training was performed on a brain-cropped template size of [96, 112, 96]. Inner label training was performed similarly to our brain extraction training where the number of layers at the base layer was reduced to eight. Training also occurred over multiple sessions where, initially, categorical cross entropy was used and then subsquently refined using a Dice loss function. Other hyperparameters can be directly inferred from explicit specification in the actual code. Training data was derived from application of joint label fusion13 during the course of our previous work16. When calling antspynet.desikan\_killiany\_tourville\_labeling, inner labels are estimated first followed by the outer, cortical labels.

## Acknowledgments

Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee 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.

# References

1. Bajcsy, R. & Broit, C. Matching of deformed images. in *Sixth International Conference on Pattern Recognition (ICPR’82)* 351–353 (1982).

2. Bajcsy, R. & Kovacic, S. Multiresolution elastic matching. *Computer Vision, Graphics, and Image Processing* **46**, 1–21 (1989).

3. Gee, J., Sundaram, T., Hasegawa, I., Uematsu, H. & Hatabu, H. Characterization of regional pulmonary mechanics from serial magnetic resonance imaging data. *Acad Radiol* **10**, 1147–52 (2003).

4. Klein, A. *et al.* Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *Neuroimage* **46**, 786–802 (2009).

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

6. Murphy, K. *et al.* Evaluation of registration methods on thoracic CT: The EMPIRE10 challenge. *IEEE Trans Med Imaging* **30**, 1901–20 (2011).

7. Menze, B., Reyes, M. & Van Leemput, K. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imaging* (2014) doi:[10.1109/TMI.2014.2377694](https://doi.org/10.1109/TMI.2014.2377694).

8. Tustison, N. J., Avants, B. B. & Gee, J. C. Learning image-based spatial transformations via convolutional neural networks: A review. *Magn Reson Imaging* **64**, 142–153 (2019).

9. Avants, B. B. *et al.* The optimal template effect in hippocampus studies of diseased populations. *Neuroimage* **49**, 2457–66 (2010).

10. Avants, B. B., Tustison, N. J., Wu, J., Cook, P. A. & Gee, J. C. An open source multivariate framework for -tissue segmentation with evaluation on public data. *Neuroinformatics* **9**, 381–400 (2011).

11. Tustison, N. J. & Gee, J. C. N4ITK: Nick’s N3 ITK implementation for MRI bias field correction. *The Insight Journal* (2009).

12. Manjón, J. V., Coupé, P., Martí-Bonmatí, L., Collins, D. L. & Robles, M. Adaptive non-local means denoising of MR images with spatially varying noise levels. *J Magn Reson Imaging* **31**, 192–203 (2010).

13. Wang, H. & Yushkevich, P. A. Multi-atlas segmentation with joint label fusion and corrective learning-an open source implementation. *Front Neuroinform* **7**, 27 (2013).

14. Wang, H. *et al.* Multi-atlas segmentation with joint label fusion. *IEEE Trans Pattern Anal Mach Intell* **35**, 611–23 (2013).

15. Das, S. R., Avants, B. B., Grossman, M. & Gee, J. C. Registration based cortical thickness measurement. *Neuroimage* **45**, 867–79 (2009).

16. Tustison, N. J. *et al.* Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. *Neuroimage* **99**, 166–79 (2014).

17. Esteban, O. *et al.* FMRIPrep: A robust preprocessing pipeline for functional MRI. *Nat Methods* **16**, 111–116 (2019).

18. De Leener, B. *et al.* SCT: Spinal cord toolbox, an open-source software for processing spinal cord MRI data. *Neuroimage* **145**, 24–43 (2017).

19. Gorgolewski, K. J. *et al.* The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci Data* **3**, 160044 (2016).

20. Halchenko, Y. O. & Hanke, M. Open is not enough. Let’s take the next step: An integrated, community-driven computing platform for neuroscience. *Front Neuroinform* **6**, 22 (2012).

21. Muschelli, J. *et al.* Neuroconductor: An R platform for medical imaging analysis. *Biostatistics* **20**, 218–239 (2019).

22. Gorgolewski, K. *et al.* Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in python. *Front Neuroinform* **5**, 13 (2011).

23. Fischl, B. FreeSurfer. *Neuroimage* **62**, 774–81 (2012).

24. Falk, T. *et al.* U-net: Deep learning for cell counting, detection, and morphometry. *Nat Methods* **16**, 67–70 (2019).

25. Bashyam, V. M. *et al.* MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14,468 individuals worldwide. *Brain* **143**, 2312–2324 (2020).

26. Goubran, M. *et al.* Hippocampal segmentation for brains with extensive atrophy using three-dimensional convolutional neural networks. *Hum Brain Mapp* **41**, 291–308 (2020).

27. Li, H. *et al.* Fully convolutional network ensembles for white matter hyperintensities segmentation in mr images. *Neuroimage* **183**, 650–665 (2018).

28. Haris, M., Shakhnarovich, G. & Ukita, N. Deep back-projection networks for super-resolution. in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 1664–1673 (2018). doi:[10.1109/CVPR.2018.00179](https://doi.org/10.1109/CVPR.2018.00179).

29. Tustison, N. J. *et al.* Longitudinal mapping of cortical thickness measurements: An Alzheimer’s Disease Neuroimaging Initiative-based evaluation study. *J Alzheimers Dis* (2019) doi:[10.3233/JAD-190283](https://doi.org/10.3233/JAD-190283).

30. Avants, B. B., Klein, A., Tustison, N. J., Woo, J. & Gee, J. C. Evaluation of open-access, automated brain extraction methods on multi-site multi-disorder data. in *16th annual meeting for the organization of human brain mapping* (2010).

31. Rebsamen, M., Rummel, C., Reyes, M., Wiest, R. & McKinley, R. Direct cortical thickness estimation using deep learning-based anatomy segmentation and cortex parcellation. *Hum Brain Mapp* (2020) doi:[10.1002/hbm.25159](https://doi.org/10.1002/hbm.25159).

32. Henschel, L. *et al.* FastSurfer - a fast and accurate deep learning based neuroimaging pipeline. *Neuroimage* **219**, 117012 (2020).

33. Tustison, N. J. *et al.* N4ITK: Improved N3 bias correction. *IEEE Trans Med Imaging* **29**, 1310–20 (2010).

34. Ashburner, J. & Friston, K. J. Voxel-based morphometry–the methods. *Neuroimage* **11**, 805–21 (2000).

35. Avants, B. *et al.* Eigenanatomy improves detection power for longitudinal cortical change. *Med Image Comput Comput Assist Interv* **15**, 206–13 (2012).

36. Klein, A. & Tourville, J. 101 labeled brain images and a consistent human cortical labeling protocol. *Front Neurosci* **6**, 171 (2012).

37. https://brain-development.org/ixi-dataset/.

38. Landman, B. A. *et al.* Multi-parametric neuroimaging reproducibility: A 3-T resource study. *Neuroimage* **54**, 2854–66 (2011).

39. http://fcon\_1000.projects.nitrc.org/indi/pro/nki.html.

40. https://www.oasis-brains.org.

41. Tustison, N. J. *et al.* Convolutional neural networks with template-based data augmentation for functional lung image quantification. *Acad Radiol* **26**, 412–423 (2019).

42. Schlemper, J. *et al.* Attention gated networks: Learning to leverage salient regions in medical images. *Med Image Anal* **53**, 197–207 (2019).

43. Lemaitre, H. *et al.* Normal age-related brain morphometric changes: Nonuniformity across cortical thickness, surface area and gray matter volume? *Neurobiol Aging* **33**, 617.e1–9 (2012).

44. Breiman, L. Random forests. *Machine Learning* **45**, 5–32 (2001).

45. Holbrook, A. J. *et al.* Anterolateral entorhinal cortex thickness as a new biomarker for early detection of Alzheimer’s disease. *Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring* **12**, e12068 (2020).

46. Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F. & Baker, C. I. Circular analysis in systems neuroscience: The dangers of double dipping. *Nat Neurosci* **12**, 535–40 (2009).

47. https://bicr-resource.atr.jp/srpbs1600/.

48. Verbeke, G. Linear mixed models for longitudinal data. in *Linear mixed models in practice* 63–153 (Springer, 1997).

49. Gelman, A. & others. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis* **1**, 515–534 (2006).

50. McKinley, R. *et al.* Few-shot brain segmentation from weakly labeled data with deep heteroscedastic multi-task networks. *CoRR* **abs/1904.02436**, (2019).

51. Fonov, V. S., Evans, A. C., McKinstry, R. C., Almli, C. & Collins, D. L. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage* **S102**, (2009).

52. Nyúl, L. G. & Udupa, J. K. On standardizing the MR image intensity scale. *Magn Reson Med* **42**, 1072–81 (1999).

53. Clarkson, M. J. *et al.* A comparison of voxel and surface based cortical thickness estimation methods. *Neuroimage* **57**, 856–65 (2011).

54. Schwarz, C. G. *et al.* A large-scale comparison of cortical thickness and volume methods for measuring alzheimer’s disease severity. *Neuroimage Clin* **11**, 802–812 (2016).

55. Tustison, N. J. *et al.* Instrumentation bias in the use and evaluation of scientific software: Recommendations for reproducible practices in the computational sciences. *Front Neurosci* **7**, 162 (2013).