

1

2 ANTsX: A dynamic ecosystem for 3 quantitative biological and medical imaging

4 Nicholas J. Tustison^{1,9}, Philip A. Cook², Andrew J. Holbrook³, Hans J. Johnson⁴, John
5 Muschelli⁵, Gabriel A. Devenyi⁶, Jeffrey T. Duda², Sandhitsu R. Das², Nicholas C. Cullen⁷,
6 Daniel L. Gillen⁸, Michael A. Yassa⁹, James R. Stone¹, James C. Gee², Brian B. Avants¹ for
7 the Alzheimer's Disease Neuroimaging Initiative

8 ¹Department of Radiology and Medical Imaging, University of Virginia, Charlottesville, VA

9 ²Department of Radiology, University of Pennsylvania, Philadelphia, PA

10 ³Department of Biostatistics, University of California, Los Angeles, CA

11 ⁴Department of Electrical and Computer Engineering, University of Iowa, Philadelphia, PA

12 ⁵School of Public Health, Johns Hopkins University, Baltimore, MD

13 ⁶Douglas Mental Health University Institute, Department of Psychiatry, McGill University, Montreal, QC

14 ⁷Lund University, Scania, SE

15 ⁸Department of Statistics, University of California, Irvine, CA

16 ⁹Department of Neurobiology and Behavior, University of California, Irvine, CA

17 Corresponding author:

18 Nicholas J. Tustison, DSc

19 Department of Radiology and Medical Imaging

20 University of Virginia

21 ntustison@virginia.edu

22 Abstract

23 The Advanced Normalizations Tools ecosystem, known as ANTsX, consists of multiple open-
24 source software libraries which house top-performing algorithms used worldwide by scientific
25 and research communities for processing and analyzing biological and medical imaging data.
26 The base software library, ANTs, is built upon, and contributes to, the NIH-sponsored
27 Insight Toolkit. Founded in 2008 with the highly regarded Symmetric Normalization image
28 registration framework, the ANTs library has since grown to include additional functionality.
29 Recent enhancements include statistical, visualization, and deep learning capabilities through
30 interfacing with both the R statistical project (ANTsR) and Python (ANTsPy). Additionally,
31 the corresponding deep learning extensions ANTsRNet and ANTsPyNet (built on the popular
32 TensorFlow/Keras libraries) contain several popular network architectures and trained models
33 for specific applications. One such comprehensive application is a deep learning analog
34 for generating cortical thickness data from structural T1-weighted brain MRI, both cross-
35 sectionally and longitudinally. These pipelines significantly improve computational efficiency
36 and provide comparable-to-superior accuracy over multiple criteria relative to the existing
37 ANTs workflows and simultaneously illustrate the importance of the comprehensive ANTsX
38 approach as a framework for medical image analysis.

³⁹ **The ANTsX ecosystem: A brief overview**

⁴⁰ **Image registration origins**

⁴¹ The Advanced Normalization Tools (ANTs) is a state-of-the-art, open-source software toolkit
⁴² for image registration, segmentation, and other functionality for comprehensive biological and
⁴³ medical image analysis. Historically, ANTs is rooted in advanced image registration techniques
⁴⁴ which have been at the forefront of the field due to seminal contributions that date back to
⁴⁵ the original elastic matching method of Bajcsy and co-investigators^{1–3}. Various independent
⁴⁶ platforms have been used to evaluate ANTs tools since their early development. In a landmark
⁴⁷ paper⁴, the authors reported an extensive evaluation using multiple neuroimaging datasets
⁴⁸ analyzed by fourteen different registration tools, including the Symmetric Normalization
⁴⁹ (SyN) algorithm⁵, and found that “ART, SyN, IRTK, and SPM’s DARTEL Toolbox gave
⁵⁰ the best results according to overlap and distance measures, with ART and SyN delivering
⁵¹ the most consistently high accuracy across subjects and label sets.” Participation in other
⁵² independent competitions^{6,7} provided additional evidence of the utility of ANTs registration
⁵³ and other tools^{8–10}. Despite the extremely significant potential of deep learning for image
⁵⁴ registration algorithmic development¹¹, ANTs registration tools continue to find application
⁵⁵ in the various biomedical imaging research communities.

⁵⁶ **Current developments**

⁵⁷ Since its inception, though, ANTs has expanded significantly beyond its image registration
⁵⁸ origins. Other core contributions include template building¹², segmentation¹³, image pre-
⁵⁹ processing (e.g., bias correction¹⁴ and denoising¹⁵), joint label fusion^{16,17}, and brain cortical
⁶⁰ thickness estimation^{18,19} (cf Table 1). Additionally, ANTs has been integrated into multiple,
⁶¹ publicly available workflows such as fMRIprep²⁰ and the Spinal Cord Toolbox²¹. Frequently
⁶² used ANTs pipelines, such as cortical thickness estimation¹⁹, have been integrated into Docker
⁶³ containers and packaged as Brain Imaging Data Structure (BIDS)²² and FlyWheel applica-
⁶⁴ tions (i.e., “gears”). It has also been independently ported for various platforms including
⁶⁵ Neurodebian²³ (Debian OS), Neuroconductor²⁴ (the R statistical project), and Nipype²⁵

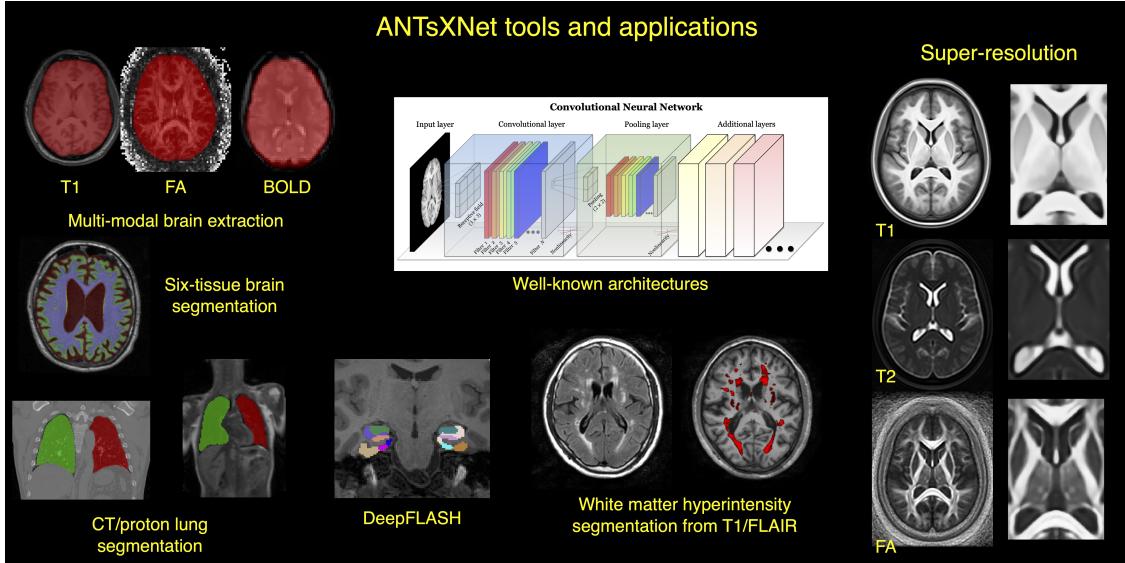


Figure 1: An illustration of the tools and applications available as part of the ANTsRNet and ANTsPyNet deep learning toolkits. Both libraries take advantage of ANTs functionality through their respective language interfaces—ANTsR (R) and ANTsPy (Python). Building on the Keras/TensorFlow language, both libraries standardize popular network architectures within the ANTs ecosystem and are cross-compatible. These networks are used to train models and weights for such applications as brain extraction which are then disseminated to the public.

⁶⁶ (Python). Additionally, other widely used software, such as FreeSurfer²⁶, have incorporated
⁶⁷ well-performing and complementary ANTs components^{14,15} into their own libraries. Finally,
⁶⁸ according to GitHub, recent unique “clones” have averaged 34 per day with the total number
⁶⁹ of clones being approximately twice that many. 50 unique contributors to the ANTs library
⁷⁰ have made a total of over 4500 commits. Additional insights into usage can be viewed at the
⁷¹ ANTs GitHub website.

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 packages interface with extremely popular, high-level, open-source programming platforms have significantly increased the user base of ANTs. 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

Functionality	Citations
SyN registration ⁵	2616
bias field correction ¹⁶	2188
ANTs registration evaluation ⁶	2013
joint label fusion ¹⁸	669
template generation ¹⁴	423
cortical thickness: implementation ²⁰	321
MAP-MRF segmentation ¹⁵	319
ITK integration ¹²	250
cortical thickness: theory ¹⁹	180

Table 1: The significance of core ANTs tools in terms of their number of citations (from October 17, 2020).

package²⁷ 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-net²⁸ 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 model²⁹, based on > 14,000 individuals; HippMapp3r³⁰, a hippocampal segmentation tool; the winning entry of the MICCAI 2017 white matter hyperintensity segmentation competition³¹; MRI super resolution using deep back-projection networks³²; and NoBrainer, a T1-weighted brain extraction approach based on FreeSurfer (see Figure 1).

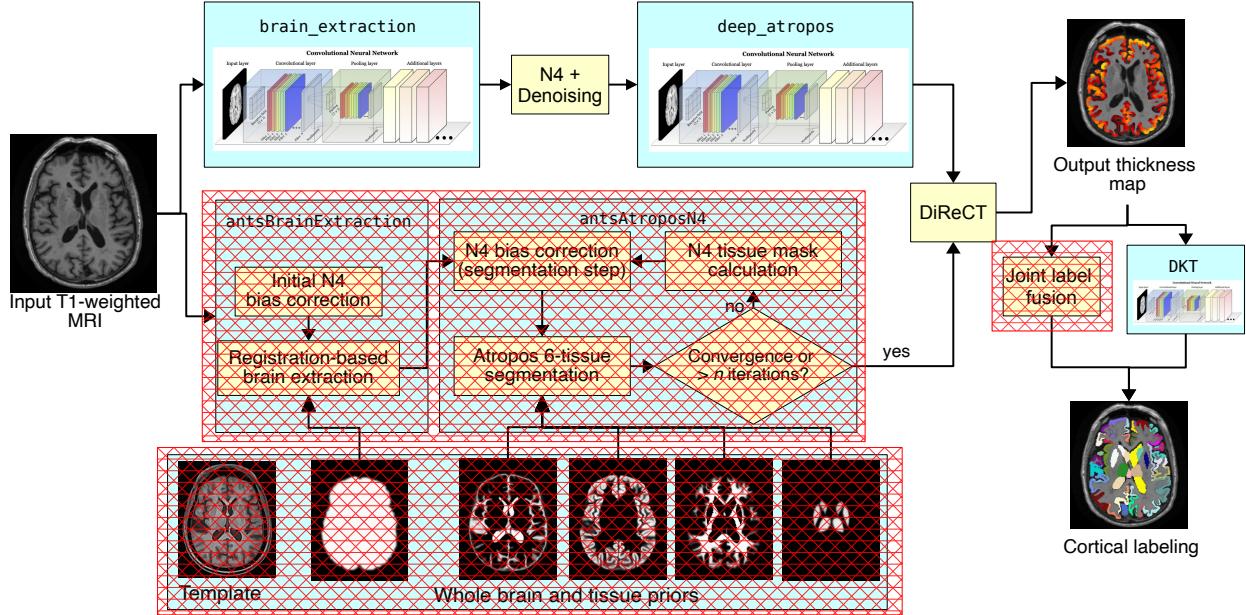


Figure 2: Illustration of the ANTsXNet cortical thickness pipeline and the relationship to its traditional ANTs analog. The hash-designated sections denote pipeline steps which have been obviated by the deep learning approach. These include template-based brain extraction, template-based n -tissue segmentation, and joint label fusion for cortical labeling. In our prior work, execution time of the thickness pipeline was dominated by registration. In the deep version of the pipeline, it is dominated by DiReCT. However, we note that registration and DiReCT execute much more quickly than in the past in part due to major improvements in the underlying ITK multi-threading strategy.

98 The ANTsXNet cortical thickness pipeline

99 The most recent ANTsX innovation involves the development of deep learning analogs of
100 our popular ANTs cortical thickness cross-sectional¹⁹ and longitudinal³³ pipelines within the
101 ANTsXNet framework. Figure 2, adapted from our previous work¹⁹, illustrates some of the
102 major changes associated with the single-subject pipeline. The resulting improvement in
103 efficiency derives primarily from eliminating deformable image registration from the pipeline—
104 a step which has historically been used to propagate prior, population-based information
105 (e.g., tissue maps) to individual subjects for such tasks as brain extraction³⁴ and tissue
106 segmentation¹³ which is now configured within the neural networks.

107 These structural MRI processing pipelines are currently available as open-source within the
108 ANTsXNet libraries. Evaluations using both cross-sectional and longitudinal data are de-

¹⁰⁹ scribed in subsequent sections and couched within the context of our previous publications^{19,33}.
¹¹⁰ Related work has been recently reported by external groups^{35,36} and provides a context for
¹¹¹ comparison to motivate the utility of the ANTsX ecosystem.

¹¹² Results

¹¹³ Cross-sectional performance evaluation

1) caudal anterior cingulate (cACC)	17) pars orbitalis (pORB)
2) caudal middle frontal (cMFG)	18) pars triangularis (pTRI)
3) cuneus (CUN)	19) pericalcarine (periCAL)
4) entorhinal (ENT)	20) postcentral (postC)
5) fusiform (FUS)	21) posterior cingulate (PCC)
6) inferior parietal (IPL)	22) precentral (preC)
7) inferior temporal (ITG)	23) precuneus (PCUN)
8) isthmus cingulate (iCC)	24) rostral anterior cingulate (rACC)
9) lateral occipital (LOG)	25) rostral middle frontal (rMFG)
10) lateral orbitofrontal (LOF)	26) superior frontal (SFG)
11) lingual (LING)	27) superior parietal (SPL)
12) medial orbitofrontal (MOF)	28) superior temporal (STG)
13) middle temporal (MTG)	29) supramarginal (SMAR)
14) parahippocampal (PARH)	30) transverse temporal (TT)
15) paracentral (paraC)	31) insula (INS)
16) pars opercularis (pOPER)	

Table 2: The 31 cortical labels (per hemisphere) of the Desikan-Killiany-Tourville atlas. The ROI abbreviations from the R `brainGraph` package are given in parentheses and used in later figures.

¹¹⁴ Due to the absence of ground-truth, we utilize the evaluation strategy from our previous
¹¹⁵ work¹⁹ where we used cross-validation to build and compare age prediction models from
¹¹⁶ data derived from both the proposed ANTsXNet pipeline and the established ANTs pipeline.
¹¹⁷ Specifically, we use “age” as a well-known and widely-available demographic correlate of
¹¹⁸ cortical thickness³⁷ and quantify the predictive capabilities of corresponding random forest
¹¹⁹ classifiers³⁸ of the form:

$$AGE \sim VOLUME + GENDER + \sum_{i=1}^{62} T(DKT_i) \quad (1)$$

120 with covariates *GENDER* and *VOLUME* (i.e., total intracranial volume). $T(DKT_i)$ is the
121 average thickness value in the i^{th} Desikan-Killiany-Tourville (DKT) region³⁹ (cf Table 2).
122 Root mean square error (RMSE) between the actual and predicted ages are the quantity
123 used for comparative evaluation. As we have explained previously¹⁹, we find these evaluation
124 measures to be much more useful than other commonly applied criteria as they are closer
125 to assessing the actual utility of these thickness measurements as biomarkers for disease⁴⁰
126 or growth. For example, in recent work³⁵ the authors employ correlation with FreeSurfer
127 thickness values as the primary evaluation for assessing relative performance with ANTs
128 cortical thickness¹⁹. This evaluation, unfortunately, is fundamentally flawed in that it is a
129 prime example of a type of circularity analysis⁴¹ whereby data selection is driven by the
130 same criteria used to evaluate performance. Specifically, the underlying DeepSCAN network
131 used for the tissue segmentation step employs training based on FreeSurfer results which
132 directly influences thickness values as thickness/segmentation are highly correlated and vary
133 characteristically between software packages. Relative performance with ANTs thickness
134 (which does not use FreeSurfer for training) is then assessed by determining correlations
135 with FreeSurfer thickness values. Almost as problematic is their use of repeatability, which
136 they confusingly label as “robustness,” as an additional ranking criterion. Repeatability
137 evaluations should be contextualized within considerations such as the bias-variance tradeoff
138 and quantified using relevant metrics, such as the intra-class correlation coefficient which
139 takes into account both inter- and intra-observer variability.

140 In addition to the training data listed above, to ensure generalizability, we also compared
141 performance using the SRPB data set⁴² comprising over 1600 participants from 12 sites. Note
142 that we recognize that we are processing a portion of the evaluation data through certain
143 components of the proposed deep learning-based pipeline that were used to train the same
144 pipeline components. Although this does not provide evidence for generalizability (which is
145 why we include the much larger SRPB data set), it is still interesting to examine the results
146 since, in this case, the deep learning training can be considered a type of noise reduction on
147 the final results. It should be noted that training did not use age prediction (or any other
148 evaluation or related measure) as a criterion to be optimized during network model training

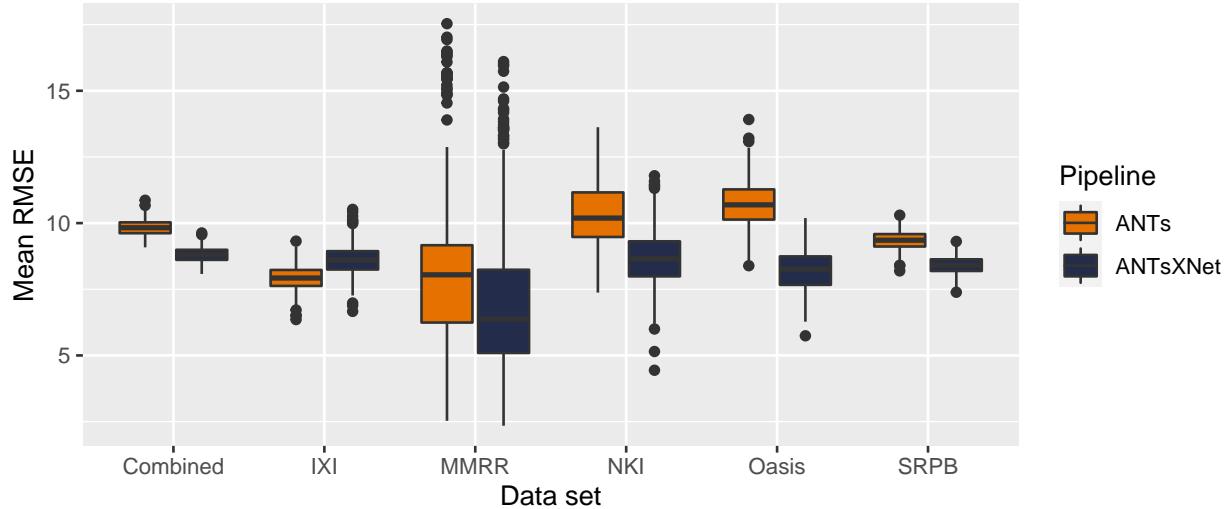


Figure 3: Distribution of mean RMSE values (500 permutations) for age prediction across the different data sets between the traditional ANTs and deep learning-based ANTsXNet pipelines. Total mean values are as follows: Combined—9.3 years (ANTs) and 8.2 years (ANTsXNet); IXI—7.9 years (ANTs) and 8.6 years (ANTsXNet); MMRR—7.9 years (ANTs) and 7.6 years (ANTsXNet); NKI—8.7 years (ANTs) and 7.9 years (ANTsXNet); OASIS—9.2 years (ANTs) and 8.0 years (ANTsXNet); and SRPB—9.2 years (ANTs) and 8.1 years (ANTsXNet).

¹⁴⁹ (i.e., circular analysis⁴¹).

¹⁵⁰ The results are shown in Figure 3 where we used cross-validation with 500 permutations
¹⁵¹ per model per data set (including a “combined” set) and an 80/20 training/testing split.
¹⁵² The ANTsXNet deep learning pipeline outperformed the classical pipeline¹⁹ in terms of age
¹⁵³ prediction in all data sets except for IXI. This also includes the cross-validation iteration
¹⁵⁴ where all data sets were combined. Additionally, repeatability assessment on [the regional](#)
¹⁵⁵ [cortical thickness values of the](#) MMRR data set yielded ICC values (“average random rater”)
¹⁵⁶ of 0.99 for both pipelines.

¹⁵⁷ [A comparative illustration of regional thickness measurements between the ANTs and](#)
¹⁵⁸ [ANTsXNet pipelines is provided in Figure 4 for three different ages spanning the lifespan.](#)
¹⁵⁹ [Linear models of the form](#)

$$T(DKT_i) \sim GENDER + AGE \quad (2)$$

were created for each of the 62 DKT regions for each pipeline. These models were then used to predict thickness values for each gender at ages of 25 years, 50 years, and 75 years and subsequently plotted relative to the absolute maximum predicted thickness value (ANTs: right entorhinal cortex at 25 years, male). Although there appear to be systematic differences between specific regional predicted thickness values (e.g., $T(ENT)_{ANTs} > T(ENT)_{ANTsXNet}$, $T(pORB)_{ANTs} < T(pORB)_{ANTsXNet}$), a pairwise t-test evidenced no statistically significant difference between the predicted thickness values of the two pipelines.

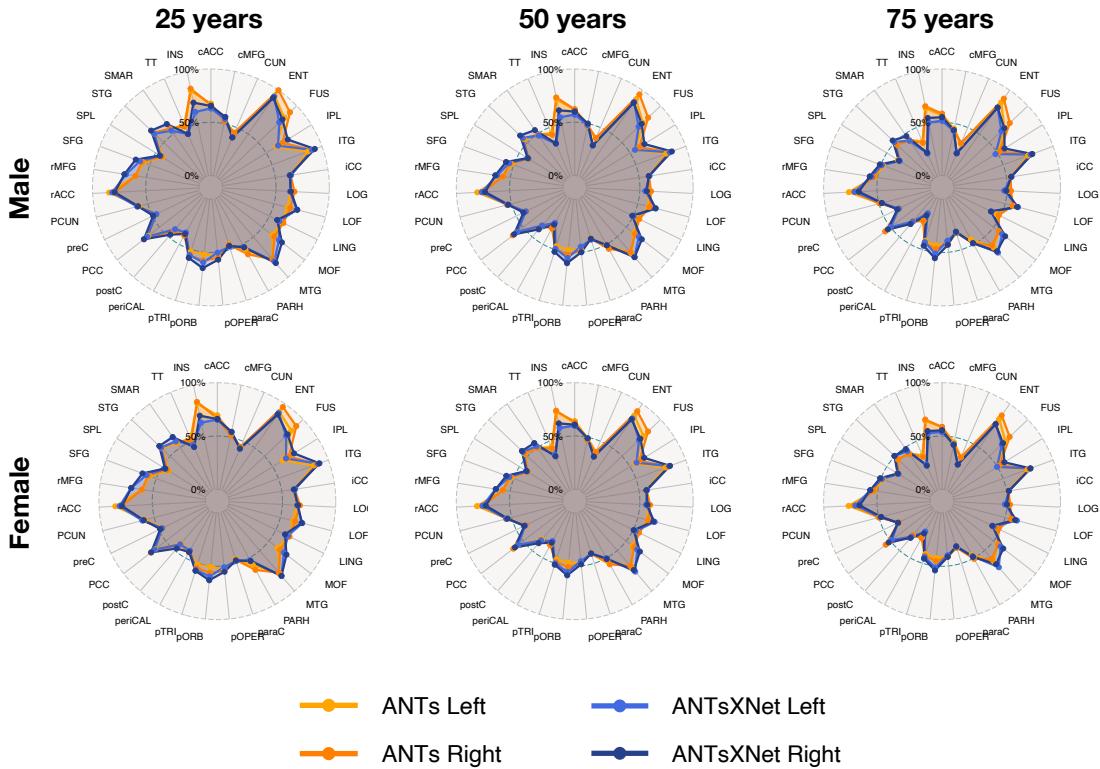
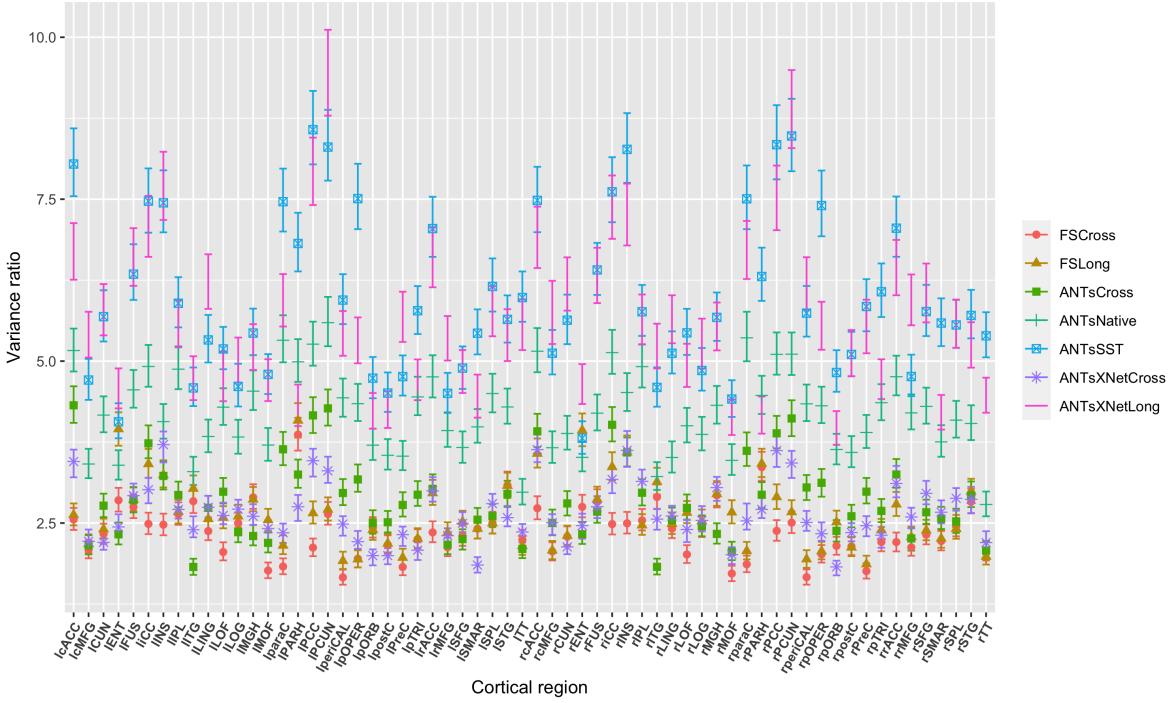


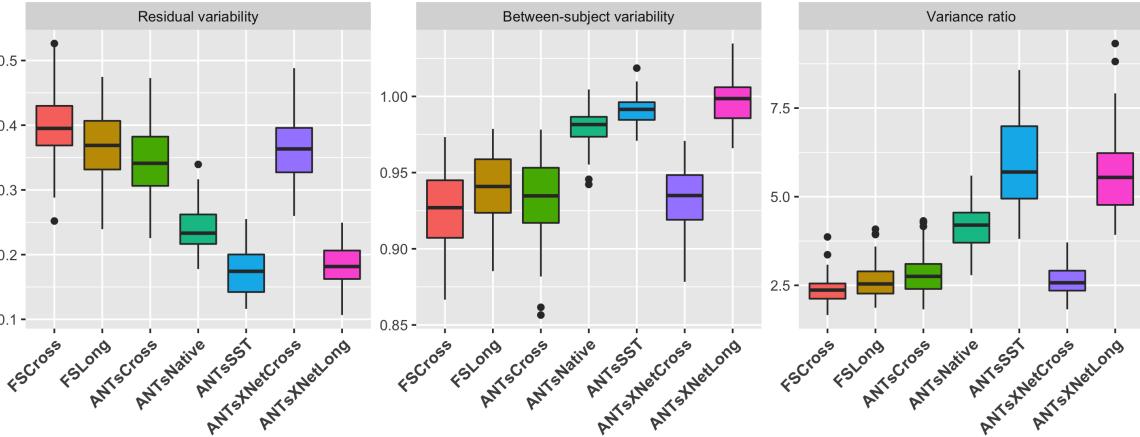
Figure 4: Radar plots enabling comparison of relative thickness values between the ANTs and ANTsXNet cortical thickness pipelines at three different ages sampling the life span.

167 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 pipeline³³. We also



(a)



(b)

Figure 5: Performance over longitudinal data as determined by the variance ratio. (a) Region-specific 95% confidence intervals of the variance ratio showing the superior performance of the longitudinally tailored ANTsX-based pipelines, including ANTsSST and ANTsXNetLong. (b) Residual variability, between subject, and variance ratio values per pipeline over all DKT regions.

evaluated an ANTsXNet-based pipeline tailored specifically for longitudinal data. In this variant, an SST is generated and processed using the previously described ANTsXNet cross-sectional pipeline which yields tissue spatial priors. These spatial priors are used in our

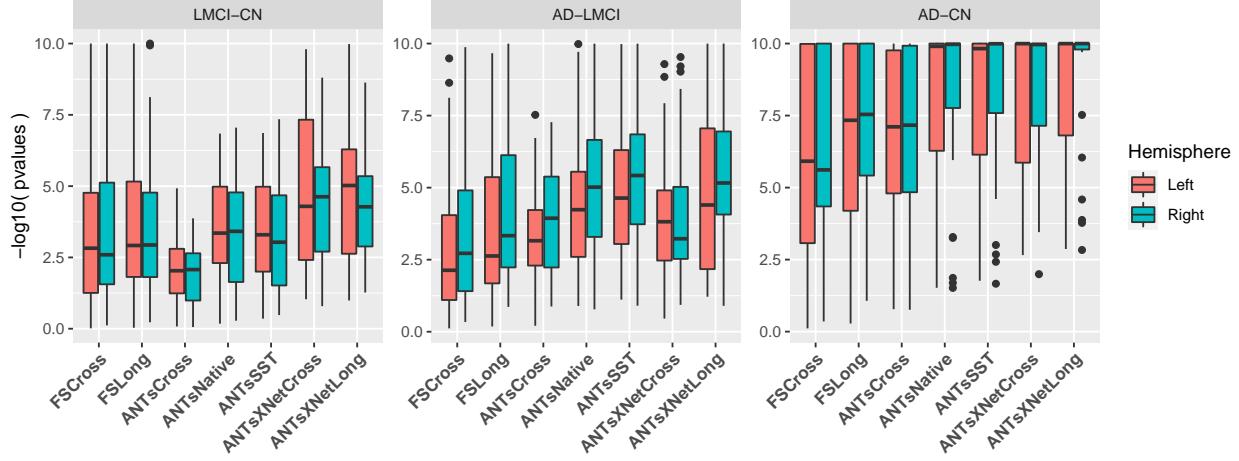


Figure 6: Measures for the supervised evaluation strategy where log p-values for diagnostic differentiation of LMCI-CN, AD-LMCI, and AD-CN subjects are plotted for all pipelines over all DKT regions.

¹⁷⁵ traditional brain segmentation approach¹³. The computational efficiency of this variant is
¹⁷⁶ also significantly improved due to the elimination of the costly SST prior generation which
¹⁷⁷ uses multiple registrations combined with joint label fusion¹⁷.

¹⁷⁸ The ADNI-1 data used for our longitudinal performance evaluation³³ consisted of over 600
¹⁷⁹ subjects (197 cognitive normals, 324 LMCI subjects, and 142 AD subjects) with one or
¹⁸⁰ more follow-up image acquisition sessions every 6 months (up to 36 months) for a total
¹⁸¹ of over 2500 images. In addition to the ANTsXNet pipelines (“ANTsXNetCross” and
¹⁸² “ANTsXNetLong”) for the current evaluation, our previous work included the FreeSurfer²⁶
¹⁸³ cross-sectional (“FSCross”) and longitudinal (“FSLong”) streams, the ANTs cross-sectional
¹⁸⁴ pipeline (“ANTsCross”) in addition to two longitudinal ANTs-based variants (“ANTsNative”
¹⁸⁵ and “ANTsSST”). Two evaluation measurements, one unsupervised and one supervised, were
¹⁸⁶ used to assess comparative performance between all seven pipelines. We add the results of
¹⁸⁷ the ANTsXNet pipeline cross-sectional and longitudinal evaluations in relation to these other
¹⁸⁸ pipelines to provide a comprehensive overview of relative performance.

First, linear mixed-effects (LME)⁴³ 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:

$$\begin{aligned} Y_{ij}^k &\sim N(\alpha_i^k + \beta_i^k t_{ij}, \sigma_k^2) \\ \alpha_i^k &\sim N(\alpha_0^k, \tau_k^2) \quad \beta_i^k \sim N(\beta_0^k, \rho_k^2) \\ \alpha_0^k, \beta_0^k &\sim N(0, 10) \quad \sigma_k, \tau_k, \rho_k \sim \text{Cauchy}^+(0, 5) \end{aligned} \quad (3)$$

where Y_{ij}^k denotes the i^{th} individual's cortical thickness measurement corresponding to the k^{th} region of interest at the time point indexed by j and specification of variance priors to half-Cauchy distributions reflects commonly accepted best practice in the context of hierarchical models⁴⁴. The ratio of interest, r^k , per region of the between-subject variability, τ_k , and residual variability, σ_k is

$$r^k = \frac{\tau_k}{\sigma_k}, k = 1, \dots, 62 \quad (4)$$

¹⁸⁹ where the posterior distribution of r_k was summarized via the posterior median.

Second, 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

$$\begin{aligned} \Delta Y &\sim Y_{bl} + AGE_{bl} + ICV_{bl} + APOE_{bl} + GENDER + DIAGNOSIS_{bl} \\ &+ VISIT : DIAGNOSIS_{bl} + (1|ID) + (1|SITE). \end{aligned} \quad (5)$$

¹⁹⁰ Here, ΔY is the change in thickness of the k^{th} DKT region from baseline (bl) thickness
¹⁹¹ Y_{bl} with random intercepts for both the individual subject (ID) and the acquisition site.
¹⁹² The subject-specific covariates AGE , $APOE$ status, $GENDER$, $DIAGNOSIS$, ICV , and
¹⁹³ $VISIT$ were taken directly from the ADNIMERGE package.

¹⁹⁴ Results for all pipelines with respect to the longitudinal evaluation criteria are shown in
¹⁹⁵ Figures 5 and 6. Figure 5(a) provides the 95% confidence intervals of the variance ratio for

196 all 64 regions of the DKT cortical labeling where ANTsSST consistently performs best with
197 ANTsXNetLong also performing well. These quantities are summarized in Figure 5(b). The
198 second evaluation criteria compares diagnostic differentiation via LMEs. Log p-values are
199 provided in Figure 6 which demonstrate excellent LMCI-CN and AD-CN differentiation for
200 both deep learning pipelines.

201 Discussion

202 The ANTsX software ecosystem provides a comprehensive framework for quantitative biological
203 and medical imaging. Although ANTs, the original core of ANTsX, is still at the forefront
204 of image registration technology, it has moved significantly beyond its image registration
205 origins. This expansion is not confined to technical contributions (of which there are many)
206 but also consists of facilitating access to a wide range of users who can use ANTsX tools
207 (whether through bash, Python, or R scripting) to construct tailored pipelines for their own
208 studies or to take advantage of our pre-fabricated pipelines. And given the open-source
209 nature of the ANTsX software, usage is not limited, for example, to non-commercial use—a
210 common constraint characteristic of other packages such as the FMRIB Software Library
211 (<https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Licence>).

212 One of our most widely used pipelines is the estimation of cortical thickness from neuroimaging.
213 This is understandable given the widespread usage of regional cortical thickness as a
214 biomarker for developmental or pathological trajectories of the brain. In this work, we used
215 this well-vetted ANTs tool to provide training data for producing alternative variants which
216 leverage deep learning for improved computational efficiency and also provides superior perfor-
217 mance with respect to previously proposed evaluation measures for both cross-sectional¹⁹ and
218 longitudinal scenarios³³. In addition to providing the tools which generated the original train-
219 ing data for the proposed ANTsXNet pipeline, the ANTsX ecosystem provides a full-featured
220 platform for the additional steps such as preprocessing (ANTsR/ANTsPy); data augmenta-
221 tion (ANTsR/ANTsPy); network construction and training (ANTsRNet/ANTsPyNet); and
222 visualization and statistical analysis of the results (ANTsR/ANTsPy).

223 It is the comprehensiveness of ANTsX that provides significant advantages over much of the
224 deep learning work that is currently taking place in medical imaging. In other words, various
225 steps in the deep learning training processing (e.g., data augmentation, preprocessing) can all
226 be performed within the same ecosystem where such important details as header information
227 for image geometry are treated the same. In contrast, related work³⁵ described and evaluated
228 a similar thickness measurement pipeline. However, due to the lack of a complete processing
229 and analysis framework, training data was generated using the FreeSurfer stream, deep
230 learning-based brain segmentation employed DeepSCAN⁴⁵ (in-house software), and cortical
231 thickness estimation¹⁸ was generated using the ANTs toolkit. The interested reader must
232 also ensure the consistency of the input/output interface between packages (a task for which
233 the Nipype development team is quite familiar.)}

234 Although potentially advantageous in terms of such issues as computational efficiency and
235 other performance measures, there are a number of limitations associated with the ANTsXNet
236 pipeline that should be mentioned both to guide potential users and possibly motivate future
237 related research. As is the case with many deep learning models, usage is restricted based on
238 training data. For example, much of the publicly available brain data has been anonymized
239 through various defacing protocols. That is certainly the case with the training data used for
240 the ANTsXNet pipeline which has consequences specific to the brain extraction step which
241 could lead to poor performance. We are currently aware of this issue and have provided
242 a temporary workaround while simultaneously resuming training on whole head data to
243 mitigate this issue. Related, although the ANTsXNet pipeline performs relatively well as
244 assessed across lifespan data, performance might be hampered for specific age ranges (e.g.,
245 neonates), whereas the traditional ANTs cortical thickness pipeline is more flexible and might
246 provide better age-targeted performance. This is the subject of ongoing research. Additionally,
247 application of the ANTsXNet pipeline would be limited with high-resolution acquisitions.
248 Due to the heavy memory requirements associated with deep learning training, the utility of
249 any resolution greater than 1 mm isotropic would not be leveraged by the existing pipeline.
250 However, there is a potential pipeline variation (akin to the longitudinal variant) that would
251 be worth exploring where Deep Atropos is used only to provide the priors for a subsequent

252 traditional Atropos segmentation on high-resolution data.

253 In terms of additional future work, the recent surge and utility of deep learning in medical
254 image analysis has significantly guided the areas of active ANTsX development. As demon-
255 strated in this work with our widely used cortical thickness pipelines, there are many potential
256 benefits of deep learning analogs to existing ANTs tools as well as the development of new
257 ones. Performance is mostly comparable-to-superior relative to existing pipelines depending
258 on the evaluation metric. Specifically, the ANTsXNet cross-sectional pipeline does well for
259 the age prediction performance framework and in terms of the ICC. Additionally, this pipeline
260 performs relatively well for longitudinal ADNI data for disease differentiation but not so
261 much in terms of the generic variance ratio criterion. However, for such longitudinal-specific
262 studies, the ANTsXNet longitudinal variant performs well for both performance measures.
263 We see possible additional longitudinal extensions incorporating subject ID and months as
264 additional network inputs.

265 Methods

266 The original ANTs cortical thickness pipeline

267 The original ANTs cortical thickness pipeline¹⁹ consists of the following steps:

- 268 • preprocessing: denoising¹⁵ and bias correction⁴⁶;
- 269 • brain extraction³⁴;
- 270 • brain segmentation with spatial tissue priors¹³ comprising the
 - 271 – cerebrospinal fluid (CSF),
 - 272 – gray matter (GM),
 - 273 – white matter (WM),
 - 274 – deep gray matter,
 - 275 – cerebellum, and
 - 276 – brain stem; and
- 277 • cortical thickness estimation¹⁸.

²⁷⁸ Our recent longitudinal variant³³ incorporates an additional step involving the construction
²⁷⁹ of a single subject template (SST)¹² coupled with the generation of tissue spatial priors of
²⁸⁰ the SST for use with the processing of the individual time points as described above.

²⁸¹ Although the resulting thickness maps are conducive to voxel-based⁴⁷ and related analyses⁴⁸,
²⁸² here we employ the well-known Desikan-Killiany-Tourville (DKT)³⁹ labeling protocol (31
²⁸³ labels per hemisphere) to parcellate the cortex for averaging thickness values regionally (cf
²⁸⁴ Table 2). This allows us to 1) be consistent in our evaluation strategy for comparison with
²⁸⁵ our previous work^{19,33} and 2) leverage an additional deep learning-based substitution within
²⁸⁶ the proposed pipeline.

²⁸⁷ Overview of cortical thickness via ANTsXNet

²⁸⁸ The entire analysis/evaluation framework, from preprocessing to statistical analysis, is made
²⁸⁹ possible through the ANTsX ecosystem and simplified through the open-source R and
²⁹⁰ Python platforms. Preprocessing, image registration, and cortical thickness estimation are
²⁹¹ all available through the ANTsPy and ANTsR libraries whereas the deep learning steps are
²⁹² performed through networks constructed and trained via ANTsRNet/ANTsPyNet with data
²⁹³ augmentation strategies and other utilities built from ANTsR/ANTsPy functionality.

²⁹⁴ The brain extraction, brain segmentation, and DKT parcellation deep learning components
²⁹⁵ were trained using data derived from our previous work¹⁹. Specifically, the IXI⁴⁹, MMRR⁵⁰,
²⁹⁶ NKI⁵¹, and OASIS⁵² data sets, and the corresponding derived data, comprising over 1200
²⁹⁷ subjects from age 4 to 94, were used for network training. Brain extraction employs a
²⁹⁸ traditional 3-D U-net network²⁸ with whole brain, template-based data augmentation²⁷
²⁹⁹ whereas brain segmentation and DKT parcellation are processed via 3-D U-net networks with
³⁰⁰ attention gating⁵³ on image octant-based batches. [Additional network architecture details](#)
³⁰¹ [are given below](#). We emphasize that a single model ([as opposed to ensemble approaches](#)
³⁰² [where multiple models are used to produce the final solution³¹](#)) was created for each of these
³⁰³ steps and was used for all the experiments described below.

304 **Implementation**

305 Software, average DKT regional thickness values for all data sets, and the scripts to perform
306 both the analysis and obtain thickness values for a single subject (cross-sectionally or
307 longitudinally) are provided as open-source. Specifically, all the ANTsX libraries are hosted
308 on GitHub (<https://github.com/ANTsX>). The cross-sectional data and analysis code are
309 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
310 are organized with the repository associated with our previous work³³ (<https://github.com/ntustison/CrossLong>).
312

```
313
314 import ants
315 import antsPyNet
316
317 # ANTsPy/ANTsPyNet processing for subject IXI002-Guys-0828-T1
318 t1_file = "IXI002-Guys-0828-T1.nii.gz"
319 t1 = ants.image_read(t1_file)
320
321 # Atropos six-tissue segmentation
322 atropos = antsPyNet.deep_atropos(t1, do_preprocessing=True, verbose=True)
323
324 # Kelly Kapowski cortical thickness (combine Atropos WM and deep GM)
325 kk_segmentation = atropos['segmentation_image']
326 kk_segmentation[kk_segmentation == 4] = 3
327 kk_gray_matter = atropos['probability_images'][2]
328 kk_white_matter = atropos['probability_images'][3] + atropos['probability_images'][4]
329 kk = ants.kelly_kapowski(s=kk_segmentation, g=kk_gray_matter, w=kk_white_matter,
330                         its=45, r=0.025, m=1.5, x=0, verbose=1)
331
332 # Desikan-Killiany-Tourville labeling
333 dkt = antsPyNet.desikan_killiany_tourville_labeling(t1, do_preprocessing=True, verbose=True)
334
335 # DKT label propagation throughout the cortex
336 dkt_cortical_mask = ants.threshold_image(dkt, 1000, 3000, 1, 0)
337 dkt = dkt_cortical_mask * dkt
338 kk_mask = ants.threshold_image(kk, 0, 0, 0, 1)
339 dkt_propagated = ants.iMath(kk_mask, "PropagateLabelsThroughMask", kk_mask * dkt)
340
341 # Get average regional thickness values
342 kkRegional_stats = ants.label_stats(kk, dkt_propagated)
```

Listing 1: ANTsPy/ANTsPyNet command calls for a single IXI subject in the evaluation study for the cross-sectional pipeline.

344 In Listing 1, we show the ANTsPy/ANTsPyNet code snippet for cross-sectional processing
345 a single subject which starts with reading the T1-weighted MRI input image, through the
346 generation of the Atropos-style six-tissue segmentation and probability images, applica-
347 tion of `ants.kelly_kapowski` (i.e., DiReCT), DKT cortical parcellation, subsequent label

348 propagation through the cortex, and, finally, regional cortical thickness tabulation. The
349 cross-sectional and longitudinal pipelines are encapsulated in the ANTsPyNet functions
350 `antspynet.cortical_thickness` and `antspynet.longitudinal_cortical_thickness`, re-
351 spectively. Note that there are precise, line-by-line R-based analogs available through
352 ANTsR/ANTsRNet.

353 Both the `ants.deep_atropos` and `antspynet.desikan_killiany_tourville_labeling`
354 functions perform brain extraction using the `antspynet.brain_extraction` function. Inter-
355 nally, `antspynet.brain_extraction` contains the requisite code to build the network and
356 assign the appropriate hyperparameters. The model weights are automatically downloaded
357 from the online hosting site <https://figshare.com> (see the function `get_pretrained_network`
358 in ANTsPyNet or `getPretrainedNetwork` in ANTsRNet for links to all models and weights)
359 and loaded to the constructed network. `antspynet.brain_extraction` performs a quick
360 translation transformation to a specific template (also downloaded automatically) using the
361 centers of intensity mass, a common alignment initialization strategy. This is to ensure
362 proper gross orientation. Following brain extraction, preprocessing for the other two deep
363 learning components includes `ants.denoise_image` and `ants.n4_bias_correction` and an
364 affine-based reorientation to a version of the MNI template⁵⁴.

365 We recognize the presence of some redundancy due to the repeated application of certain
366 preprocessing steps. Thus, each function has a `do_preprocessing` option to eliminate this
367 redundancy for knowledgeable users but, for simplicity in presentation purposes, we do not
368 provide this modified pipeline here. Although it should be noted that the time difference is
369 minimal considering the longer time required by `ants.kelly_kapowski.ants.deep_atropos`
370 returns the segmentation image as well as the posterior probability maps for each tissue
371 type listed previously. `antspynet.desikan_killiany_tourville_labeling` returns only
372 the segmentation label image which includes not only the 62 cortical labels but the remaining
373 labels as well. The label numbers and corresponding structure names are given in the program
374 description/help. Because the DKT parcellation will, in general, not exactly coincide with
375 the non-zero voxels of the resulting cortical thickness maps, we perform a label propagation
376 step to ensure the entire cortex, and only the non-zero thickness values in the cortex, are

377 included in the tabulated regional values.

378 As mentioned previously, the longitudinal version, `antspynet.longitudinal_cortical_thickness`,
379 adds an SST generation step which can either be provided as a program input or it can
380 be constructed from spatial normalization of all time points to a specified template.
381 `ants.deep_atropos` is applied to the SST yielding spatial tissues priors which are then used
382 as input to `ants.atropos` for each time point. `ants.kelly_kapowski` is applied to the
383 result to generate the desired cortical thickness maps.

384 Computational time on a CPU-only platform is approximately 1 hour primarily due to
385 `ants.kelly_kapowski` processing. Other preprocessing steps, i.e., bias correction and de-
386 noising, are on the order of a couple minutes. This total time should be compared with 4 – 5
387 hours using the traditional pipeline employing the `quick` registration option or 10 – 15 hours
388 with the more comprehensive registration parameters employed). As mentioned previously,
389 elimination of the registration-based propagation of prior probability images to individual
390 subjects is the principal source of reduced computational time. For ROI-based analyses, this
391 is in addition to the elimination of the optional generation of a population-specific template.
392 Additionally, the use of `antspynet.desikan_killiany_tourville_labeling`, for cortical
393 labeling (which completes in less than five minutes) eliminates the need for joint label fusion
394 which requires multiple pairwise registrations for each subject in addition to the fusion
395 algorithm itself.

396 Training details

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

⁴⁰⁴ **T1-weighted brain extraction.** A whole-image 3-D U-net model²⁸ was used in conjunction
⁴⁰⁵ with multiple training sessions employing a Dice loss function followed by categorical cross
⁴⁰⁶ entropy. Training data was derived from the same multi-site data described previously
⁴⁰⁷ processed through our registration-based approach³⁴. A center-of-mass-based transformation
⁴⁰⁸ to a standard template was used to standardize such parameters as orientation and voxel size.
⁴⁰⁹ However, to account for possible different header orientations of input data, a template-based
⁴¹⁰ data augmentation scheme was used²⁷ whereby forward and inverse transforms are used
⁴¹¹ to randomly warp batch images between members of the training population (followed by
⁴¹² reorientation to the standard template). A digital random coin flipping for possible histogram
⁴¹³ matching⁵⁵ between source and target images further increased data augmentation. The
⁴¹⁴ output of the network is a probabilistic mask of the brain. **The architecture consists of**
⁴¹⁵ **four encoding/decoding layers with eight filters at the base layer which doubled every layer.**
⁴¹⁶ Although not detailed here, training for brain extraction in other modalities was performed
⁴¹⁷ similarly.

⁴¹⁸ **Deep Atropos.** Dealing with 3-D data presents unique barriers for training that are often
⁴¹⁹ unique to medical imaging. Various strategies are employed such as minimizing the number
⁴²⁰ of layers and/or the number of filters at the base layer of the U-net architecture (as we
⁴²¹ do for brain extraction). However, we found this to be too limiting for capturing certain
⁴²² brain structures such as the cortex. 2-D and 2.5-D approaches are often used with varying
⁴²³ levels of success but we also found better performance using full 3-D information. This led
⁴²⁴ us to try randomly selected 3-D patches of various sizes. However, for both the six-tissue
⁴²⁵ segmentations and DKT parcellations, we found that an octant-based patch strategy yielded
⁴²⁶ the desired results. Specifically, after a brain extracted affine normalization to the MNI
⁴²⁷ template, the normalized image is cropped to a size of [160, 190, 160]. Overlapping octant
⁴²⁸ patches of size [112, 112, 112] were extracted from each image and trained using a batch size
⁴²⁹ of 12 such octant patches with weighted categorical cross entropy as the loss function. **The**
⁴³⁰ **architecture consists of four encoding/decoding layers with 16 filters at the base layer which**
⁴³¹ **doubled every layer.**

⁴³² As we point out in our earlier work¹⁹, 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 development¹⁹ 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 groups^{35,56,57} use components of our cortical thickness pipeline which can be potentially problematic⁵⁸. 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. Specifically, the weights of each tissue type were altered in order to produce segmentations which most resemble the traditional Atropos segmentations. Ultimately, we settled on a weight vector of (0.05, 1.5, 1, 3, 4, 3, 3) 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 segmentation¹³ during the course of our previous work¹⁹. 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 subsequently 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 subsequently refined using a Dice loss function. Other hyperparameters can be directly

462 inferred from explicit specification in the actual code. Training data was derived from
463 application of joint label fusion¹⁶ during the course of our previous work¹⁹. When call-
464 ing `antspynet.desikan_killiany_tourville_labeling`, inner labels are estimated first
465 followed by the outer, cortical labels.

⁴⁶⁶ **Acknowledgments**

⁴⁶⁷ Support for the research reported in this work includes funding from the National Heart, Lung,
⁴⁶⁸ and Blood Institute of the National Institutes of Health (R01HL133889) and a combined
⁴⁶⁹ grant from Cohen Veterans Bioscience (CVB-461) and the Office of Naval Research (N00014-
⁴⁷⁰ 18-1-2440).

⁴⁷¹ Data used in preparation of this article were obtained from the Alzheimer's Disease Neu-
⁴⁷² roimaging Initiative (ADNI) database (<http://adni.loni.usc.edu>). As such, the investigators
⁴⁷³ within the ADNI contributed to the design and implementation of ADNI and/or provided
⁴⁷⁴ data but did not participate in analysis or writing of this report. A complete listing of
⁴⁷⁵ ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/AD_NI_Acknowledgement_List.pdf

⁴⁷⁷ Data collection and sharing for this project was funded by the Alzheimer's Disease Neu-
⁴⁷⁸ roimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD
⁴⁷⁹ ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the
⁴⁸⁰ National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering,
⁴⁸¹ and through generous contributions from the following: AbbVie, Alzheimer's Association;
⁴⁸² Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-
⁴⁸³ Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.;
⁴⁸⁴ Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company
⁴⁸⁵ Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy
⁴⁸⁶ Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development
⁴⁸⁷ LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx
⁴⁸⁸ Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Pira-
⁴⁸⁹ mal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The
⁴⁹⁰ Canadian Institutes of Health Research is providing funds to support ADNI clinical sites
⁴⁹¹ in Canada. Private sector contributions are facilitated by the Foundation for the National
⁴⁹² Institutes of Health (www.fnih.org). The grantee organization is the Northern California
⁴⁹³ Institute for Research and Education, and the study is coordinated by the Alzheimer's

⁴⁹⁴ Therapeutic Research Institute at the University of Southern California. ADNI data are
⁴⁹⁵ disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

496 **References**

- 497 1. Bajcsy, R. & Broit, C. Matching of deformed images. in *Sixth International Conference on*
498 *Pattern Recognition (ICPR'82)* 351–353 (1982).
- 499 2. Bajcsy, R. & Kovacic, S. Multiresolution elastic matching. *Computer Vision, Graphics,*
500 *and Image Processing* **46**, 1–21 (1989).
- 501 3. Gee, J., Sundaram, T., Hasegawa, I., Uematsu, H. & Hatabu, H. Characterization of
502 regional pulmonary mechanics from serial magnetic resonance imaging data. *Acad Radiol* **10**,
503 1147–52 (2003).
- 504 4. Klein, A. *et al.* Evaluation of 14 nonlinear deformation algorithms applied to human brain
505 MRI registration. *Neuroimage* **46**, 786–802 (2009).
- 506 5. Avants, B. B., Epstein, C. L., Grossman, M. & Gee, J. C. Symmetric diffeomorphic
507 image registration with cross-correlation: Evaluating automated labeling of elderly and
508 neurodegenerative brain. *Med Image Anal* **12**, 26–41 (2008).
- 509 6. Murphy, K. *et al.* Evaluation of registration methods on thoracic CT: The EMPIRE10
510 challenge. *IEEE Trans Med Imaging* **30**, 1901–20 (2011).
- 511 7. Menze, B., Reyes, M. & Van Leemput, K. The multimodal brain tumor image segmentation
512 benchmark (BRATS). *IEEE Trans Med Imaging* (2014) doi:[10.1109/TMI.2014.2377694](https://doi.org/10.1109/TMI.2014.2377694).
- 513 8. Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J. & Dalca, A. V. VoxelMorph: A
514 learning framework for deformable medical image registration. *IEEE Trans Med Imaging*
515 (2019) doi:[10.1109/TMI.2019.2897538](https://doi.org/10.1109/TMI.2019.2897538).
- 516 9. Vos, B. D. de *et al.* A deep learning framework for unsupervised affine and deformable
517 image registration. *Med Image Anal* **52**, 128–143 (2019).
- 518 10. Fu, Y. *et al.* DeepReg: A deep learning toolkit for medical image registration. *Journal of*
519 *Open Source Software* **5**, 2705 (2020).

- 520 11. Tustison, N. J., Avants, B. B. & Gee, J. C. Learning image-based spatial transformations
521 via convolutional neural networks: A review. *Magn Reson Imaging* **64**, 142–153 (2019).
- 522 12. Avants, B. B. *et al.* The optimal template effect in hippocampus studies of diseased
523 populations. *Neuroimage* **49**, 2457–66 (2010).
- 524 13. Avants, B. B., Tustison, N. J., Wu, J., Cook, P. A. & Gee, J. C. An open source multivariate
525 framework for n -tissue segmentation with evaluation on public data. *Neuroinformatics* **9**,
526 381–400 (2011).
- 527 14. Tustison, N. J. & Gee, J. C. N4ITK: Nick’s N3 ITK implementation for MRI bias field
528 correction. *The Insight Journal* (2009).
- 529 15. Manjón, J. V., Coupé, P., Martí-Bonmatí, L., Collins, D. L. & Robles, M. Adaptive
530 non-local means denoising of MR images with spatially varying noise levels. *J Magn Reson*
531 *Imaging* **31**, 192–203 (2010).
- 532 16. Wang, H. & Yushkevich, P. A. Multi-atlas segmentation with joint label fusion and
533 corrective learning—an open source implementation. *Front Neuroinform* **7**, 27 (2013).
- 534 17. Wang, H. *et al.* Multi-atlas segmentation with joint label fusion. *IEEE Trans Pattern*
535 *Anal Mach Intell* **35**, 611–23 (2013).
- 536 18. Das, S. R., Avants, B. B., Grossman, M. & Gee, J. C. Registration based cortical thickness
537 measurement. *Neuroimage* **45**, 867–79 (2009).
- 538 19. Tustison, N. J. *et al.* Large-scale evaluation of ANTs and FreeSurfer cortical thickness
539 measurements. *Neuroimage* **99**, 166–79 (2014).
- 540 20. Esteban, O. *et al.* FMRIprep: A robust preprocessing pipeline for functional MRI. *Nat*
541 *Methods* **16**, 111–116 (2019).
- 542 21. De Leener, B. *et al.* SCT: Spinal cord toolbox, an open-source software for processing
543 spinal cord MRI data. *Neuroimage* **145**, 24–43 (2017).

- 544 22. Gorgolewski, K. J. *et al.* The brain imaging data structure, a format for organizing and
545 describing outputs of neuroimaging experiments. *Sci Data* **3**, 160044 (2016).
- 546 23. Halchenko, Y. O. & Hanke, M. Open is not enough. Let's take the next step: An
547 integrated, community-driven computing platform for neuroscience. *Front Neuroinform* **6**, 22
548 (2012).
- 549 24. Muschelli, J. *et al.* Neuroconductor: An R platform for medical imaging analysis.
550 *Biostatistics* **20**, 218–239 (2019).
- 551 25. Gorgolewski, K. *et al.* Nipype: A flexible, lightweight and extensible neuroimaging data
552 processing framework in python. *Front Neuroinform* **5**, 13 (2011).
- 553 26. Fischl, B. FreeSurfer. *Neuroimage* **62**, 774–81 (2012).
- 554 27. Tustison, N. J. *et al.* Convolutional neural networks with template-based data augmenta-
555 tion for functional lung image quantification. *Acad Radiol* **26**, 412–423 (2019).
- 556 28. Falk, T. *et al.* U-net: Deep learning for cell counting, detection, and morphometry. *Nat
557 Methods* **16**, 67–70 (2019).
- 558 29. Bashyam, V. M. *et al.* MRI signatures of brain age and disease over the lifespan based
559 on a deep brain network and 14,468 individuals worldwide. *Brain* **143**, 2312–2324 (2020).
- 560 30. Goubran, M. *et al.* Hippocampal segmentation for brains with extensive atrophy using
561 three-dimensional convolutional neural networks. *Hum Brain Mapp* **41**, 291–308 (2020).
- 562 31. Li, H. *et al.* Fully convolutional network ensembles for white matter hyperintensities
563 segmentation in mr images. *Neuroimage* **183**, 650–665 (2018).
- 564 32. Haris, M., Shakhnarovich, G. & Ukita, N. Deep back-projection networks for super-
565 resolution. in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*
566 1664–1673 (2018). doi:[10.1109/CVPR.2018.00179](https://doi.org/10.1109/CVPR.2018.00179).
- 567 33. Tustison, N. J. *et al.* Longitudinal mapping of cortical thickness measurements: An

- 568 Alzheimer's Disease Neuroimaging Initiative-based evaluation study. *J Alzheimers Dis* (2019)
- 569 doi:[10.3233/JAD-190283](https://doi.org/10.3233/JAD-190283).
- 570 34. Avants, B. B., Klein, A., Tustison, N. J., Woo, J. & Gee, J. C. Evaluation of open-access,
- 571 automated brain extraction methods on multi-site multi-disorder data. in *16th annual meeting*
- 572 *for the organization of human brain mapping* (2010).
- 573 35. Rebsamen, M., Rummel, C., Reyes, M., Wiest, R. & McKinley, R. Direct cortical
- 574 thickness estimation using deep learning-based anatomy segmentation and cortex parcellation.
- 575 *Hum Brain Mapp* (2020) doi:[10.1002/hbm.25159](https://doi.org/10.1002/hbm.25159).
- 576 36. Henschel, L. *et al.* FastSurfer - a fast and accurate deep learning based neuroimaging
- 577 pipeline. *Neuroimage* **219**, 117012 (2020).
- 578 37. Lemaitre, H. *et al.* Normal age-related brain morphometric changes: Nonuniformity
- 579 across cortical thickness, surface area and gray matter volume? *Neurobiol Aging* **33**, 617.e1–9
- 580 (2012).
- 581 38. Breiman, L. Random forests. *Machine Learning* **45**, 5–32 (2001).
- 582 39. Klein, A. & Tourville, J. 101 labeled brain images and a consistent human cortical
- 583 labeling protocol. *Front Neurosci* **6**, 171 (2012).
- 584 40. Holbrook, A. J. *et al.* Anterolateral entorhinal cortex thickness as a new biomarker for
- 585 early detection of Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment &*
- 586 *Disease Monitoring* **12**, e12068 (2020).
- 587 41. Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F. & Baker, C. I. Circular analysis
- 588 in systems neuroscience: The dangers of double dipping. *Nat Neurosci* **12**, 535–40 (2009).
- 589 42. <https://bigr-resource.atr.jp/srpbs1600/>.
- 590 43. Verbeke, G. Linear mixed models for longitudinal data. in *Linear mixed models in practice*
- 591 63–153 (Springer, 1997).

- 592 44. Gelman, A. & others. Prior distributions for variance parameters in hierarchical models
593 (comment on article by Browne and Draper). *Bayesian analysis* **1**, 515–534 (2006).
- 594 45. McKinley, R. *et al.* Few-shot brain segmentation from weakly labeled data with deep
595 heteroscedastic multi-task networks. *CoRR* **abs/1904.02436**, (2019).
- 596 46. Tustison, N. J. *et al.* N4ITK: Improved N3 bias correction. *IEEE Trans Med Imaging*
597 **29**, 1310–20 (2010).
- 598 47. Ashburner, J. & Friston, K. J. Voxel-based morphometry—the methods. *Neuroimage* **11**,
599 805–21 (2000).
- 600 48. Avants, B. *et al.* Eigenanatomy improves detection power for longitudinal cortical change.
601 *Med Image Comput Comput Assist Interv* **15**, 206–13 (2012).
- 602 49. <https://brain-development.org/ixi-dataset/>.
- 603 50. Landman, B. A. *et al.* Multi-parametric neuroimaging reproducibility: A 3-T resource
604 study. *Neuroimage* **54**, 2854–66 (2011).
- 605 51. http://fcon_1000.projects.nitrc.org/indi/pro/nki.html.
- 606 52. <https://www.oasis-brains.org>.
- 607 53. Schlemper, J. *et al.* Attention gated networks: Learning to leverage salient regions in
608 medical images. *Med Image Anal* **53**, 197–207 (2019).
- 609 54. Fonov, V. S., Evans, A. C., McKinstry, R. C., Almlí, C. & Collins, D. L. Unbiased
610 nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*
611 **S102**, (2009).
- 612 55. Nyúl, L. G. & Udupa, J. K. On standardizing the MR image intensity scale. *Magn Reson*
613 *Med* **42**, 1072–81 (1999).
- 614 56. Clarkson, M. J. *et al.* A comparison of voxel and surface based cortical thickness
615 estimation methods. *Neuroimage* **57**, 856–65 (2011).

- 616 57. Schwarz, C. G. *et al.* A large-scale comparison of cortical thickness and volume methods
617 for measuring alzheimer's disease severity. *Neuroimage Clin* **11**, 802–812 (2016).
- 618 58. Tustison, N. J. *et al.* Instrumentation bias in the use and evaluation of scientific software:
619 Recommendations for reproducible practices in the computational sciences. *Front Neurosci*
620 **7**, 162 (2013).

621 **Author contributions**

- 622 • Conception and design N.T., A.H., M.Y., J.S., B.A.
- 623 • Analysis and interpretation N.T., A.H., D.G., M.Y., J.S. B.A.
- 624 • Creation of new software N.T., P.C., H.J., J.M., G.D., J.D., S.D., N.C., J.G., B.A.
- 625 • Drafting of manuscript N.T., A.H., P.C., H.J., J.M., G.D., J.G., B.A.

626 **Competing interests**

627 The authors declare no competing interests.