

The ANTsX ecosystem for quantitative biological and medical imaging

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

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

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

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

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

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

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

⁷Lund University, Scania, SE

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

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

Corresponding author:

Nicholas J. Tustison, DSc

Department of Radiology and Medical Imaging

University of Virginia

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
⁴³ and medical image analysis. Historically, ANTs is rooted in advanced image registration
⁴⁴ techniques which have been at the forefront of the field due to seminal contributions that
⁴⁵ date back to the original elastic matching method of Bajcsy and co-investigators.^{54,55,59}
⁴⁶ Various independent platforms have been used to evaluate ANTs tools since their early
⁴⁷ development. In a landmark paper,⁶⁵ the authors reported an extensive evaluation using
⁴⁸ multiple neuroimaging datasets analyzed by fourteen different registration tools, including
⁴⁹ the Symmetric Normalization (SyN) algorithm,⁶⁰ and found that “ART, SyN, IRTK, and
⁵⁰ SPM’s DARTEL Toolbox gave the best results according to overlap and distance measures,
⁵¹ with ART and SyN delivering the most consistently high accuracy across subjects and label
⁵² sets.” Participation in other independent competitions^{62,69} provided additional evidence of the
⁵³ utility of ANTs registration and other tools.^{13,14,42} Despite the extremely significant potential
⁵⁴ of deep learning for image registration algorithmic development,⁴¹ ANTs registration tools
⁵⁵ continue to find application 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 building,⁶⁴ segmentation,⁶⁸ image pre-
⁵⁹ processing (e.g., bias correction⁵² and denoising),⁵⁶ joint label fusion,^{53,63} and brain cortical
⁶⁰ thickness estimation^{57,66} (cf Table 1). Additionally, ANTs has been integrated into multiple,
⁶¹ publicly available workflows such as fMRIprep⁵⁰ and the Spinal Cord Toolbox.⁴⁹ Frequently
⁶² used ANTs pipelines, such as cortical thickness estimation,⁶⁶ have been integrated into Docker
⁶³ containers and packaged as Brain Imaging Data Structure (BIDS)⁴⁸ and FlyWheel applica-
⁶⁴ tions (i.e., “gears’ ’). It has also been independently ported for various platforms including
⁶⁵ Neurodebian⁴⁷ (Debian OS), Neuroconductor⁴⁶ (the R statistical project), and Nipype⁴⁵

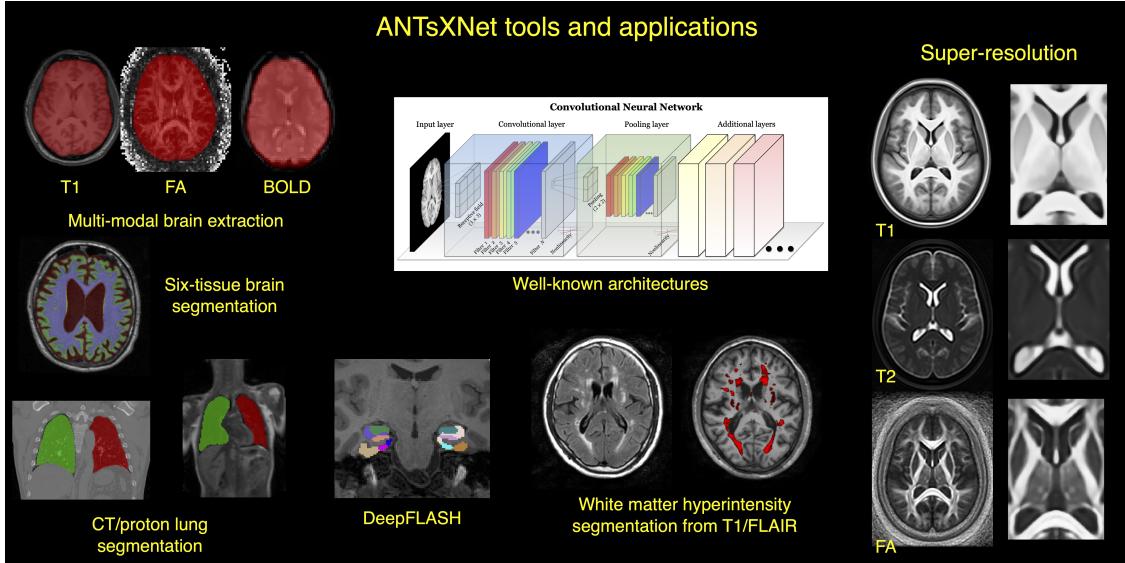


Figure 1: An illustration of the tools and applications available as part of the ANTsRNet and ANTsPyNet deep learning toolkits. Both libraries take advantage of ANTs functionality through their respective language interfaces—ANTsR (R) and ANTsPy (Python). Building on the Keras/TensorFlow language, both 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 FreeSurfer,⁶¹ have incorporated
 67 well-performing and complementary ANTs components^{52,56} into their own libraries. According
 68 to GitHub, recent unique “clones” have averaged 34 per day with the total number of clones
 69 being approximately twice that many. 50 unique contributors to the ANTs library have made
 70 a total of over 4500 commits. Additional insights into usage can be viewed at the ANTs
 71 GitHub website.

72 Over the course of its development, ANTs has been extended to complementary frameworks
 73 resulting in the Python- and R-based ANTsPy and ANTsR toolkits, respectively. These
 74 ANTs-based packages interface with extremely popular, high-level, open-source programming
 75 platforms which have significantly increased the user base of ANTs. The rapidly rising
 76 popularity of deep learning motivated further recent enhancement of ANTs and its extensions.
 77 Despite the existence of an abundance of online innovation and code for deep learning
 78 algorithms, much of it is disorganized and lacks a uniformity in structure and external data
 79 interfaces which would facilitate greater uptake. With this in mind, ANTsR spawned the deep

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).

80 learning ANTsRNet package³² which is a growing Keras/TensorFlow-based library of popular
 81 deep learning architectures and applications specifically geared towards medical imaging.
 82 Analogously, ANTsPyNet is an additional ANTsX complement to ANTsPy. Both, which we
 83 collectively refer to as “ANTsXNet”, are co-developed so as to ensure cross-compatibility
 84 such that training performed in one library is readily accessible by the other library. In
 85 addition to a variety of popular network architectures (which are implemented in both 2-D
 86 and 3-D), ANTsXNet contains a host of functionality for medical image analysis that have
 87 been developed in-house and collected from other open-source projects. For example, an
 88 extremely popular ANTsXNet application is a multi-modal brain extraction tool that uses
 89 different variants of the popular U-net⁴⁴ architecture for segmenting the brain in multiple
 90 modalities. These modalities include conventional T1-weighted structural MRI as well as
 91 T2-weighted MRI, FLAIR, fractional anisotropy, and BOLD data. Demographic specialization
 92 also includes infant T1-weighted and/or T2-weighted MRI. Additionally, we have included
 93 other models and weights into our libraries such as a recent BrainAGE estimation model,²³
 94 based on > 14,000 individuals; HippMapp3r,⁴³ a hippocampal segmentation tool; the winning
 95 entry of the MICCAI 2017 white matter hyperintensity segmentation competition;⁴⁰ MRI
 96 super resolution using deep back-projection networks;²² and NoBrainer, a T1-weighted brain
 97 extraction approach based on FreeSurfer (see Figure 1).

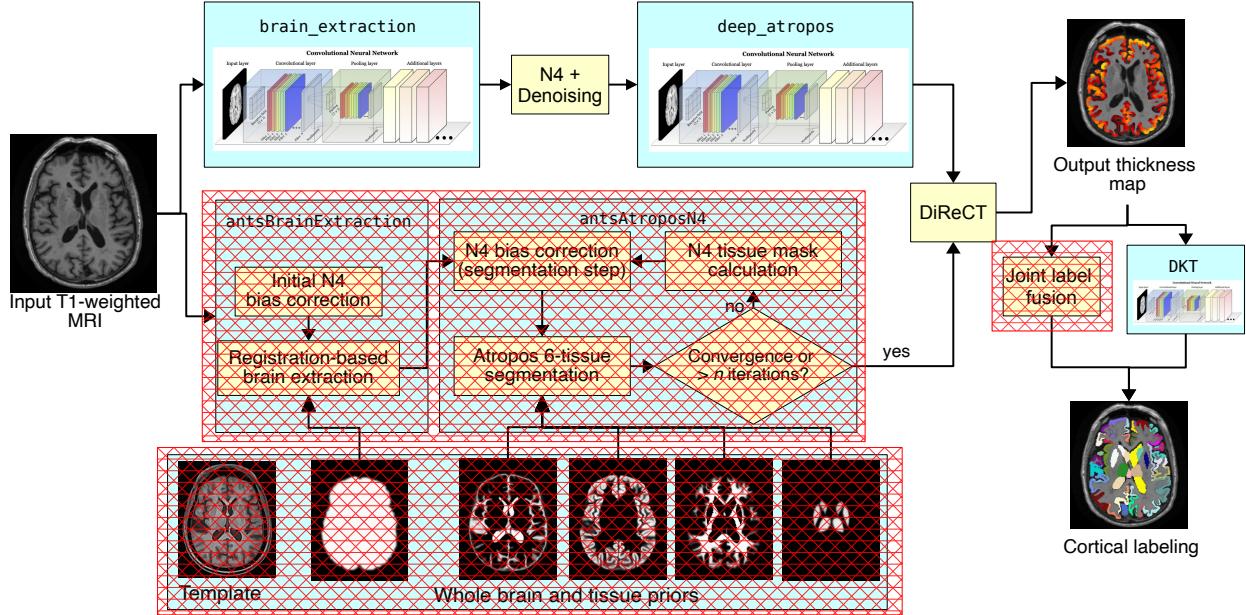


Figure 2: Illustration of the ANTsXNet cortical thickness pipeline and the relationship to its traditional ANTs analog. The hash-designated sections denote pipeline steps which have been obviated by the deep learning approach. These include template-based brain extraction, template-based n -tissue segmentation, 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-sectional⁶⁶ and longitudinal⁵¹ pipelines within
101 the ANTsXNet framework. Figure 2, adapted from our previous work,⁶⁶ illustrates some of
102 the major changes associated with the single-subject, cross-sectional pipeline. The resulting
103 improvement in efficiency derives primarily from eliminating deformable image registration
104 from the pipeline—a step which has historically been used to propagate prior, population-
105 based information (e.g., tissue maps) to individual subjects for such tasks as brain extraction⁵⁸
106 and tissue segmentation⁶⁸ which is now configured within the neural networks and trained
107 weights.

108 These structural MRI processing pipelines are currently available as open-source within the

¹⁰⁹ ANTsXNet libraries. Evaluations using both cross-sectional and longitudinal data are de-
¹¹⁰ scribed in subsequent sections and couched within the context of our previous publications.^{51,66}
¹¹¹ Related work has been recently reported by external groups^{38,39} and provides a context for
¹¹² comparison to motivate the utility of the ANTsX ecosystem.

¹¹³ **Results**

¹¹⁴ **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.

¹¹⁵ Due to the absence of ground-truth, we utilize the evaluation strategy from our previous
¹¹⁶ work⁶⁶ where we used cross-validation to build and compare age prediction models from
¹¹⁷ data derived from both the proposed ANTsXNet pipeline and the established ANTs pipeline.
¹¹⁸ Specifically, we use “age” as a well-known and widely-available demographic correlate of
¹¹⁹ cortical thickness³⁰ and quantify the predictive capabilities of corresponding random forest
¹²⁰ classifiers¹⁹ of the form:

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

121 with covariates *GENDER* and *VOLUME* (i.e., total intracranial volume). $T(DKT_i)$ is the
 122 average thickness value in the i^{th} Desikan-Killiany-Tourville (DKT) region³⁵ (cf Table 2).
 123 Root mean square error (RMSE) between the actual and predicted ages are the quantity
 124 used for comparative evaluation. As we have explained previously,⁶⁶ we find these evaluation
 125 measures to be much more useful than other commonly applied criteria as they are closer to
 126 assessing the actual utility of these thickness measurements as biomarkers for disease²¹ or
 127 growth. In recent work³⁹ the authors employ correlation with FreeSurfer thickness values as
 128 the primary evaluation for assessing relative performance with ANTs cortical thickness.⁶⁶
 129 This evaluation, unfortunately, is fundamentally flawed in that it is a prime example of a
 130 type of circularity analysis²⁹ whereby data selection is driven by the same criteria used to
 131 evaluate performance. Specifically, the underlying DeepSCAN network used for the tissue
 132 segmentation step employs training based on FreeSurfer results which directly influences
 133 thickness values as thickness/segmentation are highly correlated and vary characteristically
 134 between software packages. Relative performance with ANTs thickness (which does not use
 135 FreeSurfer for training) is then assessed by determining correlations with FreeSurfer thickness
 136 values. Almost as problematic is their use of repeatability, which they confusingly label
 137 as “robustness,” as an additional ranking criterion. Repeatability evaluations should be
 138 contextualized within considerations such as the bias-variance tradeoff and quantified using
 139 relevant metrics, such as the intra-class correlation coefficient which takes into account both
 140 inter- and intra-observer variability.

141 In addition to the training data listed above, to ensure generalizability, we also compared
 142 performance using the SRPB data set¹⁵ comprising over 1600 participants from 12 sites. Note
 143 that we recognize that we are processing a portion of the evaluation data through certain
 144 components of the proposed deep learning-based pipeline that were used to train the same
 145 pipeline components. Although this does not provide evidence for generalizability (which is
 146 why we include the much larger SRPB data set), it is still interesting to examine the results
 147 since, in this case, the deep learning training can be considered a type of noise reduction on

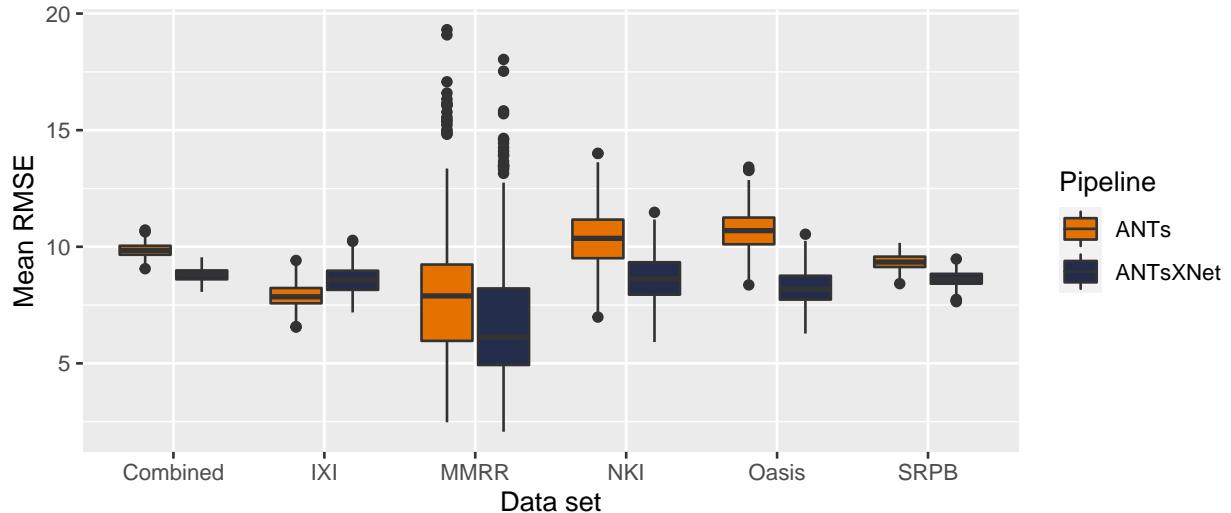


Figure 3: Distribution of mean RMSE values (500 permutations) for age prediction across the different data sets between the traditional ANTs and deep learning-based ANTsXNet pipelines. Total mean values are as follows: Combined—9.3 years (ANTs) and 8.2 years (ANTsXNet); IXI—7.9 years (ANTs) and 8.6 years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

148 the final results. It should be noted that training did not use age prediction (or any other
 149 evaluation or related measure) as a criterion to be optimized during network model training
 150 (i.e., circular analysis).²⁹

151 The results are shown in Figure 3 where we used cross-validation with 500 permutations
 152 per model per data set (including a “combined” set) and an 80/20 training/testing split.
 153 The ANTsXNet deep learning pipeline outperformed the classical pipeline⁶⁶ in terms of age
 154 prediction in all data sets except for IXI. This also includes the cross-validation iteration
 155 where all data sets were combined. Additionally, repeatability assessment on the regional
 156 cortical thickness values of the MMRR data set yielded ICC values (“average random rater”)
 157 of 0.99 for both pipelines.

158 A comparative illustration of regional thickness measurements between the ANTs and
 159 ANTsXNet pipelines is provided in Figure 4 for three different ages spanning the lifespan.
 160 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.

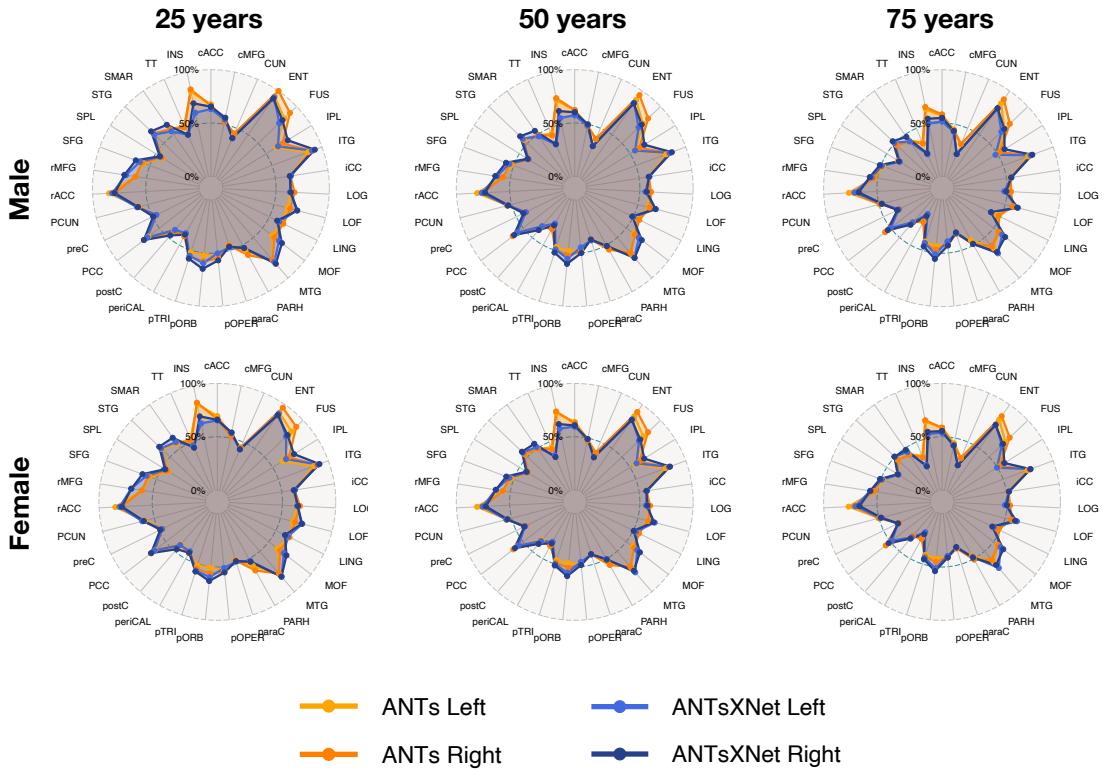
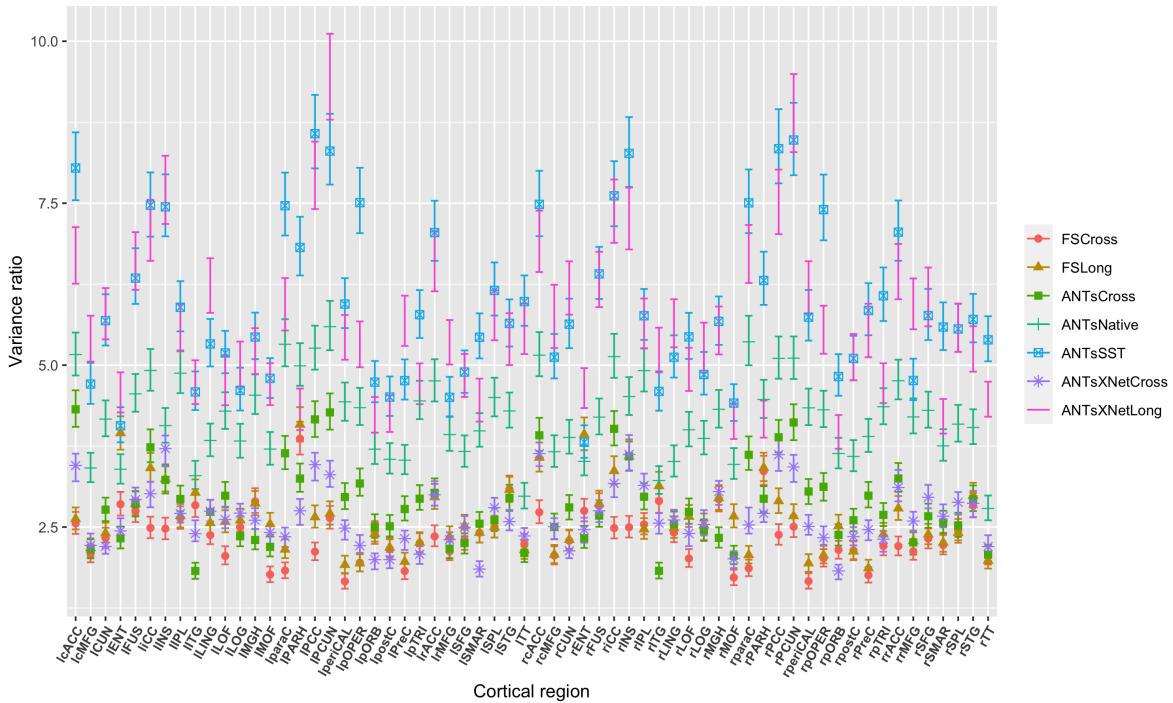
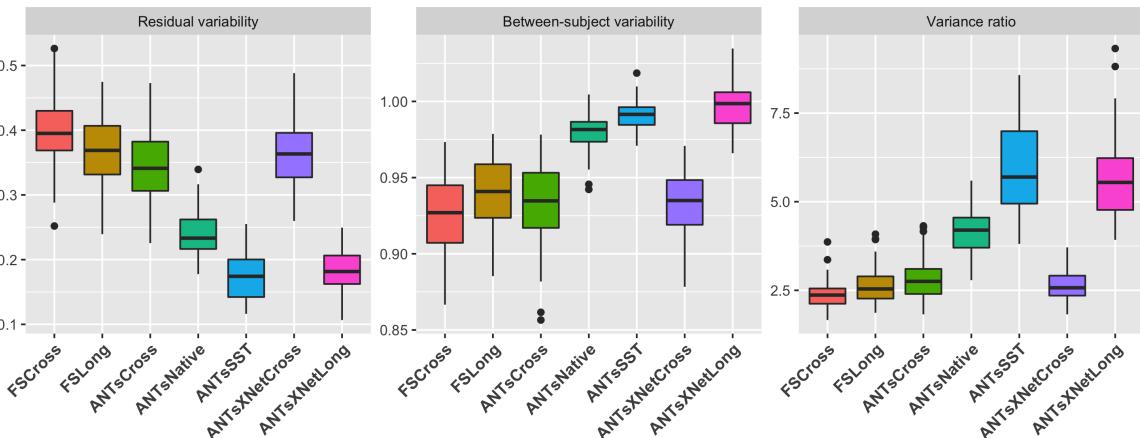


Figure 4: Radar plots enabling comparison of relative thickness values between the ANTs and ANTsXNet cortical thickness pipelines at three different ages sampling the life span. See Table 2 for region abbreviations.



(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.

168 Longitudinal performance evaluation

¹⁶⁹ Given the excellent performance and superior computational efficiency of the proposed
¹⁷⁰ ANTsXNet pipeline for cross-sectional data, we evaluated its performance on longitudinal

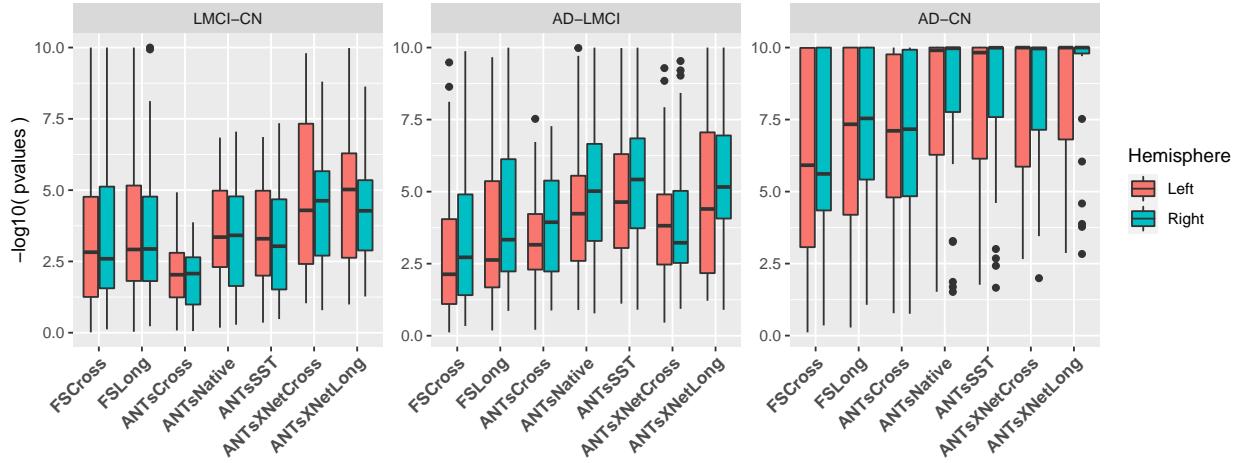


Figure 6: Measures for the supervised evaluation strategy where log p-values for diagnostic differentiation of LMCI-CN, AD-LMCI, and AD-CN subjects are plotted for all pipelines over all DKT regions.

171 data using the longitudinally-specific evaluation strategy and data we employed with the
 172 introduction of the longitudinal version of the ANTs cortical thickness pipeline.⁵¹ We also
 173 evaluated an ANTsXNet-based pipeline tailored specifically for longitudinal data. In this
 174 variant, an SST is generated and processed using the previously described ANTsXNet cross-
 175 sectional pipeline which yields tissue spatial priors. These spatial priors are used in our
 176 traditional brain segmentation approach⁶⁸. The computational efficiency of this variant is
 177 also significantly improved, in part, due to the elimination of the costly SST prior generation
 178 which uses multiple registrations combined with joint label fusion.⁵³

179 The ADNI-1 data used for our longitudinal performance evaluation⁵¹ consists of over 600
 180 subjects (197 cognitive normals, 324 LMCI subjects, and 142 AD subjects) with one or
 181 more follow-up image acquisition sessions every 6 months (up to 36 months) for a total
 182 of over 2500 images. In addition to the ANTsXNet pipelines (“ANTsXNetCross” and
 183 “ANTsXNetLong”) for the current evaluation, our previous work included the FreeSurfer⁶¹
 184 cross-sectional (“FSCross”) and longitudinal (“FSLong”) streams, the ANTs cross-sectional
 185 pipeline (“ANTsCross”) in addition to two longitudinal ANTs-based variants (“ANTsNative”
 186 and “ANTsSST”). Two evaluation measurements, one unsupervised and one supervised, were
 187 used to assess comparative performance between all seven pipelines. We add the results of
 188 the ANTsXNet pipeline cross-sectional and longitudinal evaluations in relation to these other

¹⁸⁹ pipelines to provide a comprehensive overview of relative performance.

First, linear mixed-effects (LME)²⁰ 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:

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

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 models.²⁸ The ratio of interest, r^k , per region of the between-subject variability, τ_k , and residual variability, σ_k is

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

¹⁹⁰ 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

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

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

¹⁹⁴ *VISIT* were taken directly from the ADNIMERGE package.

¹⁹⁵ Results for all pipelines with respect to the longitudinal evaluation criteria are shown in
¹⁹⁶ Figures 5 and 6. Figure 5(a) provides the 95% confidence intervals of the variance ratio for
¹⁹⁷ all 62 regions of the DKT cortical labeling where ANTsSST consistently performs best with
¹⁹⁸ ANTsXNetLong also performing well. These quantities are summarized in Figure 5(b). The
¹⁹⁹ second evaluation criteria compares diagnostic differentiation via LMEs. Log p-values are
²⁰⁰ provided in Figure 6 which demonstrate excellent LMCI-CN and AD-CN differentiation for
²⁰¹ both deep learning pipelines.

²⁰² Discussion

²⁰³ The ANTsX software ecosystem provides a comprehensive framework for quantitative biological
²⁰⁴ and medical imaging. Although ANTs, the original core of ANTsX, is still at the forefront
²⁰⁵ of image registration technology, it has moved significantly beyond its image registration
²⁰⁶ origins. This expansion is not confined to technical contributions (of which there are many)
²⁰⁷ but also consists of facilitating access to a wide range of users who can use ANTsX tools
²⁰⁸ (whether through bash, Python, or R scripting) to construct tailored pipelines for their own
²⁰⁹ studies or to take advantage of our pre-fabricated pipelines. And given the open-source
²¹⁰ nature of the ANTsX software, usage is not limited, for example, to non-commercial use—a
²¹¹ common constraint characteristic of other packages such as the FMRIB Software Library
²¹² (<https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Licence>).

²¹³ One of our most widely used pipelines is the estimation of cortical thickness from neuroimaging.
²¹⁴ This is understandable given the widespread usage of regional cortical thickness as a
²¹⁵ biomarker for developmental or pathological trajectories of the brain. In this work, we used
²¹⁶ this well-vetted ANTs tool to provide training data for producing alternative variants which
²¹⁷ leverage deep learning for improved computational efficiency and also provides superior performance
²¹⁸ with respect to previously proposed evaluation measures for both cross-sectional⁶⁶ and
²¹⁹ longitudinal scenarios.⁵¹ In addition to providing the tools which generated the original training
²²⁰ data for the proposed ANTsXNet pipeline, the ANTsX ecosystem provides a full-featured

221 platform for the additional steps such as preprocessing (ANTsR/ANTsPy); data augmentation
222 (ANTsR/ANTsPy); network construction and training (ANTsRNet/ANTsPyNet); and
223 visualization and statistical analysis of the results (ANTsR/ANTsPy).

224 Using ANTsX, various steps in the deep learning training processing (e.g., data augmentation,
225 preprocessing) can all be performed within the same ecosystem where such important details
226 as header information for image geometry are treated the same. In contrast, related work³⁹
227 described and evaluated a similar thickness measurement pipeline. However, due to the lack
228 of a complete processing and analysis framework, training data was generated using the
229 FreeSurfer stream, deep learning-based brain segmentation employed DeepSCAN²⁷ (in-house
230 software), and cortical thickness estimation⁵⁷ was generated using the ANTs toolkit. The
231 interested researcher must ensure the consistency of the input/output interface between
232 packages (a task for which the Nipype development team is quite familiar.)

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

250 be worth exploring where Deep Atropos is used only to provide the priors for a subsequent
251 traditional Atropos segmentation on high-resolution data. Although direct evaluation by
252 the principal co-authors of the ANTs toolkit, the similarity in resulting cortical thickness
253 values, as indicated by Figure 4, and considerations of the training data origins all strongly
254 suggest similarity between Atropos and Deep Atropos output, further evaluation is certainly
255 warranted and would benefit other potential applications.

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

268 Methods

269 The original ANTs cortical thickness pipeline

270 The original ANTs cortical thickness pipeline⁶⁶ consists of the following steps:

- 271 • preprocessing: denoising⁵⁶ and bias correction,⁶⁷
- 272 • brain extraction,⁵⁸
- 273 • brain segmentation with spatial tissue priors⁶⁸ comprising the
 - 274 – cerebrospinal fluid (CSF),
 - 275 – gray matter (GM),
 - 276 – white matter (WM),

- 277 – deep gray matter,
278 – cerebellum, and
279 – brain stem; and
280 • cortical thickness estimation.⁵⁷

281 Our recent longitudinal variant⁵¹ incorporates an additional step involving the construction
282 of a single subject template (SST)⁶⁴ coupled with the generation of tissue spatial priors of
283 the SST for use with the processing of the individual time points as described above.

284 Although the resulting thickness maps are conducive to voxel-based³⁶ and related analyses³⁷,
285 here we employ the well-known Desikan-Killiany-Tourville (DKT)³⁵ labeling protocol (31
286 labels per hemisphere) to parcellate the cortex for averaging thickness values regionally (cf
287 Table 2). This allows us to 1) be consistent in our evaluation strategy for comparison with
288 our previous work^{51,66} and 2) leverage an additional deep learning-based substitution within
289 the proposed pipeline.

290 Overview of cortical thickness via ANTsXNet

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

297 The brain extraction, brain segmentation, and DKT parcellation deep learning components
298 were trained using data derived from our previous work.⁶⁶ Specifically, the IXI,¹⁸ MMRR,³¹
299 NKI,¹⁷ and OASIS¹⁶ data sets, and the corresponding derived data, comprising over 1200
300 subjects from age 4 to 94, were used for network training. Brain extraction employs a
301 traditional 3-D U-net network⁴⁴ with whole brain, template-based data augmentation³²
302 whereas brain segmentation and DKT parcellation are processed via 3-D U-net networks with
303 attention gating³³ on image octant-based batches. Additional network architecture details
304 are given below. We emphasize that a single model (as opposed to ensemble approaches

305 where multiple models are used to produce the final solution)⁴⁰ was created for each of these
306 steps and was used for all the experiments described below.

307 Implementation

308 Software, average DKT regional thickness values for all data sets, and the scripts to perform
309 both the analysis and obtain thickness values for a single subject (cross-sectionally or
310 longitudinally) are provided as open-source. Specifically, all the ANTsX libraries are hosted
311 on GitHub (<https://github.com/ANTsX>). The cross-sectional data and analysis code
312 are available as .csv files and R scripts at the GitHub repository dedicated to this paper
313 (<https://github.com/ntustison/PaperANTsX>) whereas the longitudinal data and evaluation
314 scripts are organized with the repository associated with our previous work⁵¹ (<https://github.com/ntustison/CrossLong>).
315

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

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

347 In Listing 1, we show the ANTsPy/ANTsPyNet code snippet for cross-sectional processing

348 a single subject which starts with reading the T1-weighted MRI input image, through the
349 generation of the Atropos-style six-tissue segmentation and probability images, applica-
350 tion of `ants.kelly_kapowski` (i.e., DiReCT), DKT cortical parcellation, subsequent label
351 propagation through the cortex, and, finally, regional cortical thickness tabulation. The
352 cross-sectional and longitudinal pipelines are encapsulated in the ANTsPyNet functions
353 `antspynet.cortical_thickness` and `antspynet.longitudinal_cortical_thickness`, re-
354 spectively. Note that there are precise, line-by-line R-based analogs available through
355 ANTsR/ANTsRNet.

356 Both the `ants.deep_atropos` and `antspynet.desikan_killiany_tourville_labeling`
357 functions perform brain extraction using the `antspynet.brain_extraction` function. Inter-
358 nally, `antspynet.brain_extraction` contains the requisite code to build the network and
359 assign the appropriate hyperparameters. The model weights are automatically downloaded
360 from the online hosting site <https://figshare.com> (see the function `get_pretrained_network`
361 in ANTsPyNet or `getPretrainedNetwork` in ANTsRNet for links to all models and weights)
362 and loaded to the constructed network. `antspynet.brain_extraction` performs a quick
363 translation transformation to a specific template (also downloaded automatically) using the
364 centers of intensity mass, a common alignment initialization strategy. This is to ensure
365 proper gross orientation. Following brain extraction, preprocessing for the other two deep
366 learning components includes `ants.denoise_image` and `ants.n4_bias_correction` and an
367 affine-based reorientation to a version of the MNI template.³⁴

368 We recognize the presence of some redundancy due to the repeated application of certain
369 preprocessing steps. Thus, each function has a `do_preprocessing` option to eliminate this
370 redundancy for knowledgeable users but, for simplicity in presentation purposes, we do not
371 provide this modified pipeline here. Although it should be noted that the time difference is
372 minimal considering the longer time required by `ants.kelly_kapowski`. `ants.deep_atropos`
373 returns the segmentation image as well as the posterior probability maps for each tissue
374 type listed previously. `antspynet.desikan_killiany_tourville_labeling` returns only
375 the segmentation label image which includes not only the 62 cortical labels but the remaining
376 labels as well. The label numbers and corresponding structure names are given in the program

377 description/help. Because the DKT parcellation will, in general, not exactly coincide with
378 the non-zero voxels of the resulting cortical thickness maps, we perform a label propagation
379 step to ensure the entire cortex, and only the non-zero thickness values in the cortex, are
380 included in the tabulated regional values.

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

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

399 Training details

400 Training differed slightly between models and so we provide details for each of these com-
401 ponents below. For all training, we used ANTsRNet scripts and custom batch generators.
402 Although the network construction and other functionality is available in both ANTsPyNet
403 and ANTsRNet (as is model weights compatibility), we have not written such custom batch
404 generators for the former (although this is on our to-do list). In terms of hardware, all

405 training was done on a DGX (GPUs: 4X Tesla V100, system memory: 256 GB LRDIMM
406 DDR4).

407 **T1-weighted brain extraction.** A whole-image 3-D U-net model⁴⁴ was used in conjunction
408 with multiple training sessions employing a Dice loss function followed by categorical cross
409 entropy. Training data was derived from the same multi-site data described previously
410 processed through our registration-based approach.⁵⁸ A center-of-mass-based transformation
411 to a standard template was used to standardize such parameters as orientation and voxel size.
412 However, to account for possible different header orientations of input data, a template-based
413 data augmentation scheme was used³² whereby forward and inverse transforms are used
414 to randomly warp batch images between members of the training population (followed by
415 reorientation to the standard template). A digital random coin flipping for possible histogram
416 matching²⁶ between source and target images further increased data augmentation. The
417 output of the network is a probabilistic mask of the brain. The architecture consists of
418 four encoding/decoding layers with eight filters at the base layer which doubled every layer.
419 Although not detailed here, training for brain extraction in other modalities was performed
420 similarly.

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

434 doubled every layer.

435 As we point out in our earlier work,⁶⁶ obtaining proper brain segmentation is perhaps the
436 most critical step to estimating thickness values that have the greatest utility as a potential
437 biomarker. In fact, the first and last authors (NT and BA, respectively) spent much time
438 during the original ANTs pipeline development⁶⁶ trying to get the segmentation correct which
439 required manually looking at many images and adjusting parameters where necessary. This
440 fine-tuning is often omitted or not considered when other groups^{24,25,39} use components of our
441 cortical thickness pipeline which can be potentially problematic⁷⁰. Fine-tuning for this partic-
442 ular workflow was also performed between the first and last authors using manual variation of
443 the weights in the weighted categorical cross entropy. Specifically, the weights of each tissue
444 type were altered in order to produce segmentations which most resemble the traditional
445 Atropos segmentations. Ultimately, we settled on a weight vector of (0.05, 1.5, 1, 3, 4, 3, 3) for
446 the CSF, GM, WM, Deep GM, brain stem, and cerebellum, respectively. Other hyperparam-
447 eters can be directly inferred from explicit specification in the actual code. As mentioned
448 previously, training data was derived from application of the ANTs Atropos segmentation⁶⁸
449 during the course of our previous work.⁶⁶ Data augmentation included small affine and
450 deformable perturbations using `antspynet.randomly_transform_image_data` and random
451 contralateral flips.

452 **Desikan-Killiany-Tourville parcellation.** Preprocessing for the DKT parcellation train-
453 ing was similar to the Deep Atropos training. However, the number of labels and the
454 complexity of the parcellation required deviation from other training steps. First, labeling
455 was split into an inner set and an outer set. Subsequent training was performed separately
456 for both of these sets. For the cortical labels, a set of corresponding input prior probability
457 maps were constructed from the training data (and are also available and automatically
458 downloaded, when needed, from <https://figshare.com>). Training occurred over multiple
459 sessions where, initially, categorical cross entropy was used and then subsequently refined
460 using a Dice loss function. Whole-brain training was performed on a brain-cropped template
461 size of [96, 112, 96]. Inner label training was performed similarly to our brain extraction
462 training where the number of layers at the base layer was reduced to eight. Training also

⁴⁶³ occurred over multiple sessions where, initially, categorical cross entropy was used and then
⁴⁶⁴ subsequently refined using a Dice loss function. Other hyperparameters can be directly
⁴⁶⁵ inferred from explicit specification in the actual code. Training data was derived from
⁴⁶⁶ application of joint label fusion⁶³ during the course of our previous work.⁶⁶ When call-
⁴⁶⁷ ing `antspynet.desikan_killiany_tourville_labeling`, inner labels are estimated first
⁴⁶⁸ followed by the outer cortical labels.

⁴⁶⁹ **Other softwares**

⁴⁷⁰ Several R¹ packages were used in preparation of this manuscript including R Markdown,^{10–12}
⁴⁷¹ lme4,⁷ RStan,⁶ ggplot2,⁹ and ggradar2.⁸ Other packages used include Apple Pages,³ ITK-
⁴⁷² SNAP,² LibreOffice,⁴ and diagrams.net.⁵

473 **Acknowledgments**

474 Support for the research reported in this work includes funding from the National Heart, Lung,
475 and Blood Institute of the National Institutes of Health (R01HL133889) and a combined
476 grant from Cohen Veterans Bioscience (CVB-461) and the Office of Naval Research (N00014-
477 18-1-2440).

478 Data used in preparation of this article were obtained from the Alzheimer's Disease Neu-
479 roimaging Initiative (ADNI) database (<http://adni.loni.usc.edu>). As such, the investigators
480 within the ADNI contributed to the design and implementation of ADNI and/or provided
481 data but did not participate in analysis or writing of this report. A complete listing of
482 ADNI investigators can be found at: <http://adni.loni.usc.edu/wp-content/uploads/how> to
483 apply/AD NI Acknowledgement List.pdf

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

502 disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

503 **References**

- 504 1. R Core Team. *R: A language and environment for statistical computing*. (R Foundation for Statistical Computing, 2020).
- 505 2. Yushkevich, P. A. *et al.* User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage* **31**, 1116–1128 (2006).
- 506 3. <https://www.apple.com/pages/>.
- 507 4. <https://www.libreoffice.org/>.
- 508 5. <https://app.diagrams.net>.
- 509 6. Stan Development Team. RStan: The R interface to Stan. (2020).
- 510 7. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67**, 1–48 (2015).
- 511 8. <https://github.com/xl0418/ggradar2>.
- 512 9. Wickham, H. *ggplot2: Elegant graphics for data analysis*. (Springer-Verlag New York, 2016).
- 513 10. Allaire, J. *et al.* Rmarkdown: Dynamic documents for r. (2021).
- 514 11. Xie, Y., Allaire, J. J. & Grolemund, G. *R markdown: The definitive guide*. (Chapman; Hall/CRC, 2018).
- 515 12. Xie, Y., Dervieux, C. & Riederer, E. *R markdown cookbook*. (Chapman; Hall/CRC, 2020).
- 516 13. Fu, Y. *et al.* DeepReg: A deep learning toolkit for medical image registration. *Journal of Open Source Software* **5**, 2705 (2020).
- 517 14. Vos, B. D. de *et al.* A deep learning framework for unsupervised affine and deformable image registration. *Med Image Anal* **52**, 128–143 (2019).
- 518 15. <https://bicr-resource.atr.jp/srpbs1600/>.
- 519 16. <https://www.oasis-brains.org>.
- 520 17. http://fcon_1000.projects.nitrc.org/indi/pro/nki.html.
- 521 18. <https://brain-development.org/ixi-dataset/>.
- 522 19. Breiman, L. Random forests. *Machine Learning* **45**, 5–32 (2001).

- 523 20. Verbeke, G. Linear mixed models for longitudinal data. in *Linear mixed models in practice* 63–153 (Springer, 1997).
- 524 21. 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).
- 525 22. 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).
- 526 23. 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).
- 527 24. Clarkson, M. J. *et al.* A comparison of voxel and surface based cortical thickness estimation methods. *Neuroimage* **57**, 856–65 (2011).
- 528 25. 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).
- 529 26. Nyúl, L. G. & Udupa, J. K. On standardizing the MR image intensity scale. *Magn Reson Med* **42**, 1072–81 (1999).
- 530 27. McKinley, R. *et al.* Few-shot brain segmentation from weakly labeled data with deep heteroscedastic multi-task networks. *CoRR* **abs/1904.02436**, (2019).
- 531 28. Gelman, A. & others. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis* **1**, 515–534 (2006).
- 532 29. 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).
- 533 30. 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).
- 534 31. Landman, B. A. *et al.* Multi-parametric neuroimaging reproducibility: A 3-T resource study. *Neuroimage* **54**, 2854–66 (2011).
- 535 32. Tustison, N. J. *et al.* Convolutional neural networks with template-based data augmentation for functional lung image quantification. *Acad Radiol* **26**, 412–423 (2019).
- 536 33. Schlemper, J. *et al.* Attention gated networks: Learning to leverage salient regions in medical images. *Med Image Anal* **53**, 197–207 (2019).

- 537 34. Fonov, V. S., Evans, A. C., McKinstry, R. C., Almlie, C. & Collins, D. L. Unbiased nonlinear
average age-appropriate brain templates from birth to adulthood. *NeuroImage* **S102**, (2009).
- 538 35. Klein, A. & Tourville, J. 101 labeled brain images and a consistent human cortical labeling
protocol. *Front Neurosci* **6**, 171 (2012).
- 539 36. Ashburner, J. & Friston, K. J. Voxel-based morphometry—the methods. *Neuroimage* **11**,
805–21 (2000).
- 540 37. Avants, B. *et al.* Eigenanatomy improves detection power for longitudinal cortical change.
Med Image Comput Comput Assist Interv **15**, 206–13 (2012).
- 541 38. Henschel, L. *et al.* FastSurfer - a fast and accurate deep learning based neuroimaging pipeline.
Neuroimage **219**, 117012 (2020).
- 542 39. 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).
- 543 40. Li, H. *et al.* Fully convolutional network ensembles for white matter hyperintensities segmen-
tation in MR images. *Neuroimage* **183**, 650–665 (2018).
- 544 41. 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).
- 545 42. Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J. & Dalca, A. V. VoxelMorph: A
learning framework for deformable medical image registration. *IEEE Trans Med Imaging*
(2019) doi:[10.1109/TMI.2019.2897538](https://doi.org/10.1109/TMI.2019.2897538).
- 546 43. Goubran, M. *et al.* Hippocampal segmentation for brains with extensive atrophy using
three-dimensional convolutional neural networks. *Hum Brain Mapp* **41**, 291–308 (2020).
- 547 44. Falk, T. *et al.* U-net: Deep learning for cell counting, detection, and morphometry. *Nat
Methods* **16**, 67–70 (2019).
- 548 45. Gorgolewski, K. *et al.* Nipype: A flexible, lightweight and extensible neuroimaging data
processing framework in python. *Front Neuroinform* **5**, 13 (2011).
- 549 46. Muschelli, J. *et al.* Neuroconductor: An R platform for medical imaging analysis. *Biostatistics*
20, 218–239 (2019).
- 550 47. 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).

- 551 48. 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).
- 552 49. De Leener, B. *et al.* SCT: Spinal cord toolbox, an open-source software for processing spinal
cord MRI data. *Neuroimage* **145**, 24–43 (2017).
- 553 50. Esteban, O. *et al.* fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nat
Methods* **16**, 111–116 (2019).
- 554 51. 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).
- 555 52. Tustison, N. J. & Gee, J. C. N4ITK: Nick’s N3 ITK implementation for MRI bias field
correction. *The Insight Journal* (2009).
- 556 53. Wang, H. *et al.* Multi-atlas segmentation with joint label fusion. *IEEE Trans Pattern Anal
Mach Intell* **35**, 611–23 (2013).
- 557 54. Bajcsy, R. & Kovacic, S. Multiresolution elastic matching. *Computer Vision, Graphics, and
Image Processing* **46**, 1–21 (1989).
- 558 55. Bajcsy, R. & Broit, C. Matching of deformed images. in *Sixth International Conference on
Pattern Recognition (ICPR’82)* 351–353 (1982).
- 559 56. 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).
- 560 57. Das, S. R., Avants, B. B., Grossman, M. & Gee, J. C. Registration based cortical thickness
measurement. *Neuroimage* **45**, 867–79 (2009).
- 561 58. 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).
- 562 59. 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).
- 563 60. Avants, B. B., Epstein, C. L., Grossman, M. & Gee, J. C. Symmetric diffeomorphic image
registration with cross-correlation: Evaluating automated labeling of elderly and neurodegen-
erative brain. *Med Image Anal* **12**, 26–41 (2008).

- 564 61. Fischl, B. FreeSurfer. *Neuroimage* **62**, 774–81 (2012).
- 565 62. 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).
- 566 63. Wang, H. & Yushkevich, P. A. Multi-atlas segmentation with joint label fusion and corrective
learning—an open source implementation. *Front Neuroinform* **7**, 27 (2013).
- 567 64. Avants, B. B. *et al.* The optimal template effect in hippocampus studies of diseased populations.
Neuroimage **49**, 2457–66 (2010).
- 568 65. Klein, A. *et al.* Evaluation of 14 nonlinear deformation algorithms applied to human brain
MRI registration. *Neuroimage* **46**, 786–802 (2009).
- 569 66. Tustison, N. J. *et al.* Large-scale evaluation of ANTs and FreeSurfer cortical thickness
measurements. *Neuroimage* **99**, 166–79 (2014).
- 570 67. Tustison, N. J. *et al.* N4ITK: Improved N3 bias correction. *IEEE Trans Med Imaging* **29**,
1310–20 (2010).
- 571 68. Avants, B. B., Tustison, N. J., Wu, J., Cook, P. A. & Gee, J. C. An open source multivariate
framework for n -tissue segmentation with evaluation on public data. *Neuroinformatics* **9**,
381–400 (2011).
- 572 69. Murphy, K. *et al.* Evaluation of registration methods on thoracic CT: The EMPIRE10
challenge. *IEEE Trans Med Imaging* **30**, 1901–20 (2011).
- 573 70. 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).

574 **Author contributions**

- 575 • Conception and design N.T., A.H., M.Y., J.S., B.A.
- 576 • Analysis and interpretation N.T., A.H., D.G., M.Y., J.S. B.A.
- 577 • Creation of new software N.T., P.C., H.J., J.M., G.D., J.D., S.D., N.C., J.G., B.A.
- 578 • Drafting of manuscript N.T., A.H., P.C., H.J., J.M., G.D., J.G., B.A.

579 **Competing interests**

580 The authors declare no competing interests.