ANTsX neuroimaging-derived structural phenotypes of UK Biobank

Nicholas J. Tustison[1,2], Michael A. Yassa[2], Batool Rizvi[2] Andrew J. Holbrook[3], Mithra T. Sathishkumar[2], James C. Gee[4], James R. Stone[1], and Brian B. Avants[1]

[1]Department of Radiology & Medical Imaging, University of Virginia, Charlottesville, VA
[2]Department of Neurobiology & Behavior, University of California, Irvine, CA
[3]Department of Biostatistics, University of California, Los Angeles, CA
[4]Department of Radiology, University of Pennsylvania, Philadelphia, PA

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

Abstract

UK Biobank is a large-scale epidemiological resource for investigating prospective correlations between various lifestyle, environmental, and genetic factors with health and disease progression. In addition to individual subject information obtained through surveys and physical examinations, a comprehensive neuroimaging battery consisting of multiple modalities provides imaging-derived phenotypes (IDPs) that can serve as biomarkers in neuroscience research. In this study, we augment the existing set of UK Biobank neuroimaging structural IDPs, obtained from well-established software libraries such as FSL and FreeSurfer, with related measurements acquired through the Advanced Normalization Tools Ecosystem. This includes previously established cortical and subcortical morphological measurements defined, in part, based on the Desikan-Killiany-Tourville atlas, as well as structural volumes from a recent extension of an automated medial temporal lobe parcellation of hippocampal and extra-hippocampal regions. Through predictive modeling, we assess the clinical utility of these IDP measurements, individually and in combination, using commonly studied phenotypic correlates including age, fluid intelligence, numeric memory, and several other sociodemographic variables. The predictive accuracy of these IDP-based models, in terms of root-mean-squared-error or area-under-the-curve for continuous and categorical variables, respectively, provides comparative insights between software libraries as well as potential clinical interpretability. Results demonstrate varied performance between package-based IDP sets and their combination, emphasizing the need for careful consideration in their selection and utilization.

Keywords: DeepFLASH, FreeSurfer, FSL, No Free Lunch Theorem, open science, TabNet, XGBoost

1 Introduction

UK Biobank (UKBB) is a unique epidemiological effort which aims to prospectively identify potential relationships between disease and associated risk factors through the leveraging of comprehensive individualized medical and sociodemographic data. Enrollment began in 2006 and continued for four years ultimately resulting in a cohort of approximately 500,000 individuals. Volunteer age criteria was limited to birth years between 1934 and 1971—an optimal range for observing the onset of certain diseases and their subsequent progression. Continued monitoring of a significant subset is expected to continue for at least 30 years facilitated, in part, by coordination with the National Health Services of the UK. This has resulted in several studies exploring a wide variety of research topics (e.g., the relationship between age and cognitive decline [Cornelis et al., 2019], the association of polygenic profiles and mental health [Hagenaars et al., 2016], and potential collider bias in COVID-19 assessment [Griffith et al., 2020]).

An integral component of UKBB is the subset of approximately 50,000 subjects who underwent comprehensive imaging batteries, including neuroimaging [Alfaro-Almagro et al., 2018; Miller et al., 2016], specifically structural T1-weighted MPRAGE and T2-FLAIR MRI; diffusion-weighted MRI; resting-state and task-based functional MRI; and susceptibility-weighted MRI. Employing specialized processing pipelines, these raw imaging data are used to generate various quantities, referred to as image-derived phenotypes (IDPs), for use as potential biomarkers. A sampling of resulting image-based research studies evinces insights into such topics as hippocampal volumetric nomograms across age [Nobis et al., 2019]; population modeling of age, fluid intelligence, and neuroticism [Dadi et al., 2021]; and brain structural changes associated with COVID-19 and the corresponding cognitive effects [Douaud et al., 2022].

Facilitating the majority of existing UKBB imaging-related research is the FMRIB Software Library (FSL) [Jenkinson et al., 2012] which has been specifically tailored to provide UKBB IDPs [Alfaro-Almagro et al., 2018; Miller et al., 2016]. For the structural data alone, this includes global and cortical IDPs from FMRIB's Automated Segmentation Tool (FAST) [Zhang et al., 2001], subcortical IDPs from FMRIB's Integrated Registration and Segmen-

tation Tool (FIRST) [Patenaude et al., 2011], and white matter hyperintensity (WMH) load using the Brain Intensity AbNormality Classification Algorithm (BIANCA) [Griffanti et al., 2016]. UKBB was subsequently augmented with FreeSurfer-based IDPs [Fischl, 2012] which include both the standard "aseg" segmentation, hippocampal subfield [Iglesias et al., 2015], and amygdala nuclei [Saygin et al., 2017] pipeline outputs.

Analogously, the Advanced Normalization Tools Ecosystem (ANTsX) is a collection of interrelated, open-source software libraries for biological and medical image processing and analysis [Tustison et al., 2021b] with developmental roots in high-performing medical image registration [Avants et al., 2011a; Avants et al., 2014] and built on the Insight Toolkit (ITK) [Yoo and Metaxas, 2005]. ANTsX-based IDPs have demonstrated utility in several studies spanning a variety of organ systems, species, and imaging modalities (e.g., [Diamond et al., 2022; Ding et al., 2022; Kini et al., 2021]). These IDPs include those which have been previously reported, such as global brain tissue volumes [Avants et al., 2011b] and more localized, FreeSurfer-analogous cortical thickness values [Tustison et al., 2014; Tustison et al., 2019; Tustison et al., 2021b] averaged over the Desikan-Killiany-Tourville (DKT) regions [Klein and Tourville, 2012]. In addition, recently developed ANTsX functionality includes a medial temporal lobe (MTL) parcellation framework known as "DeepFLASH," a neural network for segmenting hippocampal subfields and extra-hippocampal regions which extends previous work [Reagh et al., 2018].

Characterizing the respective sets of FSL, FreeSurfer, and ANTsX IDPs and their mutual relationships can guide researchers in their usage as there are both significant overlap and notable differences between these measures. And although comparison between sets is potentially insightful, a focused, package-wise comparison using UKBB is difficult due to 1) the absence of complete, individual IDP correspondence across packages and 2) the general purpose of UKBB data (in contrast, for example, to the ADNI data [Weiner et al., 2012] set which focuses on Alzheimer's disease). Regarding IDP differences, even between identically defined IDPs (e.g., hippocampal volume), *observer bias*[1] is a possible source of measurement

[1] "Observer bias" is considered in the context of casting computational measurement tools as "observers" with "observer bias" due to the specific set of choices that results in the final numerical measurement. These choices can include (but are certainly not limited to) modeling considerations, preferences with respect to anatomical definitional ambiguities, and the set of parameters used to run the corresponding software.

variance [Caliskan et al., 2017]. Note that this variance is not indicative of inaccuracy, per se, such as with *instrumentation bias* where sub-optimal calibration of software is used as a straw-man for comparative purposes [Tustison et al., 2013]. Rather, observer bias is supplemental to conventional signal noise considerations as a potential source of measurement discrepancy which can provide insight when considered within the appropriate context. For example, differing labeling protocols for specific anatomical structures, such as the hippocampal subfields and parahippocampal subregions, can reveal differences and those differences can motivate and facilitate harmonization [Yushkevich et al., 2015].

To this end, in addition to the core contribution of providing ANTsX-based UKBB IDPs, we explore the similarities and differences between the respective sets of structural IDPs and their combination. Given the usage and availability of several powerful machine learning (ML) techniques, we quantify performance using multiple predictive approaches for tabulated IDP data including gradient boosted decision trees (i.e., XGBoost [Chen and Guestrin, 2016]) and tailored deep learning networks (specifically, TabNet [Arik and Pfister, 2021]). Such frameworks potentially have the additional advantage of providing clinical interpretability of individual features (e.g., SHAP [Lundberg and Lee, 2017]). For example, one of the most well-studied neuroimaging structural correlative relationships is chronological and so-called "brain age" and their health-dependent divergence [Franke and Gaser, 2019]. Such subject-specific, single values are estimated using a variety of ML approaches and IDPs. Although establishing normative values over the human life span has clinical utility, as pointed out in Nyberg [Nyberg and Wåhlin, 2020], the single-valued brain age is at the extreme end of an "optimal balance between integration and diversification" required for neuroimaging studies. A single score or index most likely does not capture the extent of the non-linearity and heterogeneity of age and other effects on brain structure [Smith et al., 2020]. In contrast, the type of feature-based investigation performed here reveals insight into such questions as: "In what ways do the different IDP sets perform in terms of their predictive capabilities?," "How does this performance vary with different sociodemographic variables?," "In what ways are features complementary and can their combined effect improve prediction performance?," and "How does this performance vary with ML technique?"

2 Materials and Methods

2.1 UK Biobank data description

The study was conducted under UKBB Resource Application ID 63965. The total number of subjects at the time of download was 502,413 with 49,351 subjects having undergone the standard imaging battery. Of these imaging subjects, only 40,898 complete sets of downloaded IDPs were in common between those provided by FSL and FreeSurfer processing streams [Alfaro-Almagro et al., 2018]. Intersection with the final ANTs processed set resulted in a total study cohort size of 40,796.

2.2 FSL structural phenotypes

All structural FSL IDPs were included for consideration [UKBB,]. These included the following categories:

- FAST regional grey matter volumes (Category ID: 1101);
- FIRST subcortical volumes (Category ID: 1102);
- global brain tissue volumes and related quantities (Field ID: 25000–25010, 25025); and
- total volume of WMH load (Field ID: 25781)

for a total of $139_{FAST} + 14_{FIRST} + 12_{Global} + 1_{WMH} = 166$ IDPs.

2.3 FreeSurfer structural phenotypes

Several categories of IDPs are available for FreeSurfer comprising a total of 1242 measurements [UKBB,]. However, to make the study dataset more computationally tractable and reduce set size differences between packages, we selected the following popular IDP subsets:

- ASEG volumetric measurements (Category ID: 190);
- DKT volumes and mean thicknesses (Category ID: 196); and

- hippocampal subfields and amygdala nuclei (Category ID: 191)

totaling $56_{ASEG} + 124_{DKT} + 121_{hipp} = 301$ individual IDPs.

2.4 ANTsX structural phenotypes

Both sociodemographic and bulk image data were downloaded to the high performance cluster at the University of Virginia for processing. Grad-warped distortion corrected [Glasser et al., 2013] T1-weighted and FLAIR image data were used to produce the following ANTsX IDPs:

- Deep Atropos brain tissue volumes (i.e., CSF, gray matter, white matter, deep gray matter, brain stem, and cerebellum);
- DKT DiReCT cortical thicknesses and volumes;
- DKT-based regional volumes;
- DeepFLASH regional volumes; and
- Regional WMH loads

totaling $7_{DeepAtropos} + 88_{DKTreg} + 128_{DKTDiReCT} + 20_{DeepFLASH} + 13_{WMH} = 254$ IDPs which are illustrated in Figure 1. We have reported previously on the first three categories of ANTsX IDPs [Tustison et al., 2021b] but provide a brief description below. We also provide further details concerning both DeepFLASH and the ANTsXNet-ported WMH algorithms.

2.4.1 Brain tissue volumes

The ANTsXNet deep learning libraries for Python and R (ANTsPyNet and ANTsRNet, respectively) were recently described in [Tustison et al., 2021b] where they were evaluated in terms of multi-site cortical thickness estimation. This extends previous work [Tustison et al., 2014; Tustison et al., 2019] in replacing key pipeline components with deep learning variants. For example, a trained network, denoted *Deep Atropos*, replaced the original Atropos algorithm [Avants et al., 2011b] for six-tissue segmentation (CSF, gray matter, white matter,

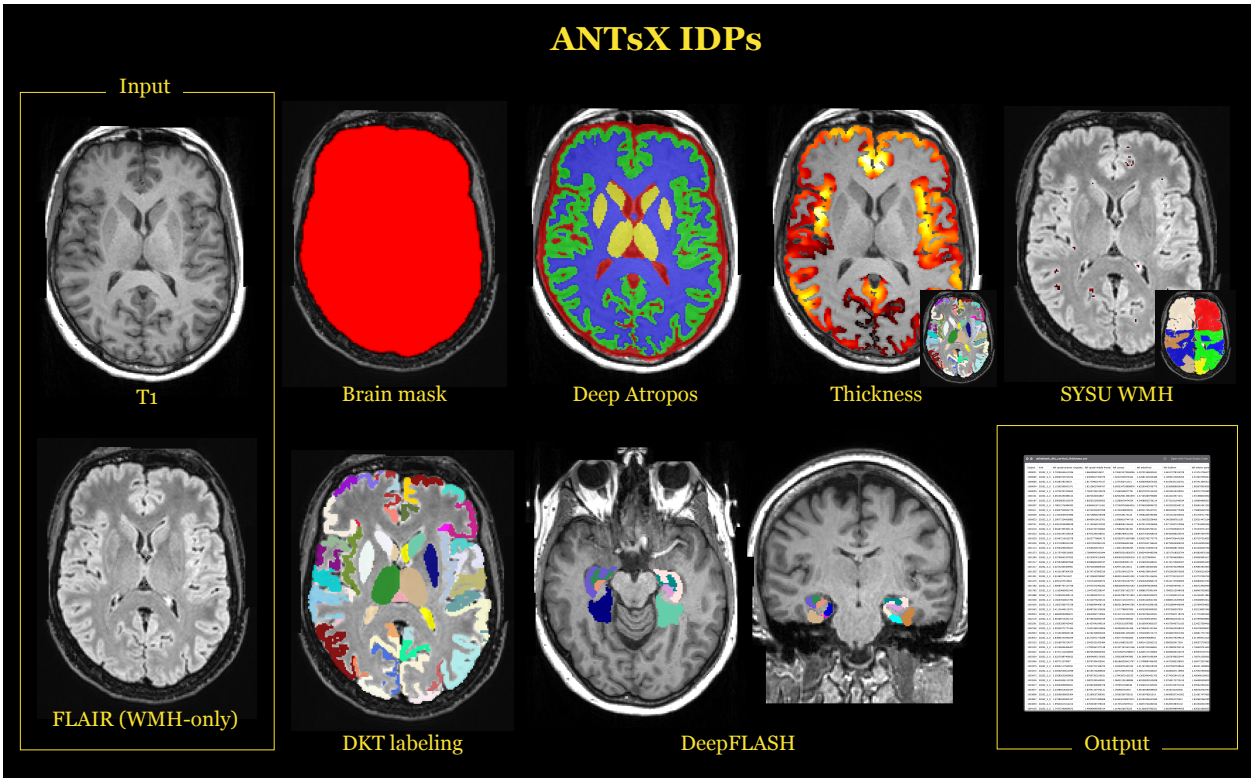

Figure 1: Illustration of the IDPs generated with ANTsX ecosystem tools. Using the gradient-distortion corrected versions of the T1 and FLAIR images, several categories of IDPs were tabulated. These include gobal brain and tissue volumes, cortical thicknesses averaged over the 62 DKT regions, WMH intensity load per lobe based on the SYSU algorithm, cortical and subcortical volumes from the DKT labeling, and MTL regional volumes using DeepFLASH.

deep gray matter, cerebellum, and brain stem) similar to functionality for whole brain deep learning-based brain extraction.

2.4.2 DKT cortical thickness, regional volumes, and lobar parcellation

As part of the deep learning refactoring of the cortical thickness pipeline mentioned in the previous section, a framework was developed to generate DKT cortical and subcortical regional labels from T1-weighted MRI [Tustison et al., 2021b]. This facilitates regional averaging of cortical thickness values over that atlas parcellation as well as being the source of other potentially useful geometry-based IDPs. In terms of network training and development, using multi-site data from [Tustison et al., 2014], two separate U-net [Falk et al., 2019] networks were trained for the "inner" (e.g., subcortical, cerebellar) labels and the "outer" cortical labels, respectively. Similar to Deep Atropos, preprocessing includes brain extraction and affine transformation to the space of the MNI152 template [Fonov et al., 2009] which includes corresponding prior probability maps. These maps are used as separate input channels for both training and prediction—a type of surrogate for network attention gating [Schlemper et al., 2019]. Using FreeSurfer's DKT atlas label-to-lobe mapping [Desikan et al., 2006], we use a fast marching approach [Sethian, 1996] to produce left/right parcellations of the frontal, temporal, parietal, and occipital lobes, as well as left/right divisions of the brain stem and cerebellum. Using the segmentation output from Deep Atropos, the DiReCT algorithm [Das et al., 2009] generates the subject-specific cortical thickness map which, as previously mentioned, is summarized in terms of IDPs by DKT regional definitions. Given the diffeomorphic and thickness constraints dictated by the DiReCT algorithm, we generate additional DKT regional labels (cortex only) from the non-zero cortical thickness regions to also be used as IDPs.

2.4.3 Fused labeling for automated segmentation of the hippocampus and extra-hippocampal regions (DeepFLASH)

An additional set of IDPs was generated from a deep learning-based framework for hippocampal and extra-hippocampal subfield parcellation also made publicly available within

ANTsXNet, which we refer to as "DeepFLASH." This work constitutes an extension of earlier work [Reagh et al., 2018], based on joint label fusion (JLF) [Wang et al., 2013], which has been used in a variety of studies [Brown et al., 2019b; Brown et al., 2019a; Holbrook et al., 2020; McMakin et al., 2021; Nguyen et al., 2019; Sinha et al., 2018]. DeepFLASH comprises both T1/T2 multi-modality and T1-only imaging networks for parcellating the following MTL regions:

- hippocampal subfields
  - Dentate gyrus/cornu ammonis 2–4 (DG/CA2/CA3/CA4)
  - cornu ammonis 1 (CA1)
  - subiculum
- extra-hippocampal regions
  - perirhinal
  - parahippocampal
  - antero-lateral entorhinal cortex (aLEC)
  - posteromedial entorhinal cortex (pMEC)

DeepFLASH employs a traditional 3-D U-net model [Falk et al., 2019] consisting of five layers with 32, 64, 96, 128, and 256 filters, respectively. In addition to the multi-region output, three additional binary outputs (the entire medial temporal lobe complex, the whole hippocampus, and the extra-hippocampal cortex) are incorporated as a hierarchical structural output set.

Training

Our earlier JLF-based work utilized a set of 17 hand-labeled atlases which were annotated using T2-weighted imaging (with accompanying T1 image data) based on a validated, well-established protocol [Reagh et al., 2018]. T1 MPRAGE [Mugler and Brookeman, 1990] scans were acquired in the sagittal orientation with an isotropic image resolution of $0.75 \times 0.75 \times 0.75$ mm^3. T2 image acquisition was performed perpendicular to the long axis of the hippocampus with resolution $0.47 \times 0.47 \times 2.0$ mm^3 consistent with previous recommendations [Yushkevich et al., 2010]. The optimal rigid transformation between corresponding T1 and T2 images

was determined using the Advanced Normalization Tools (ANTs) software package [Avants et al., 2014].

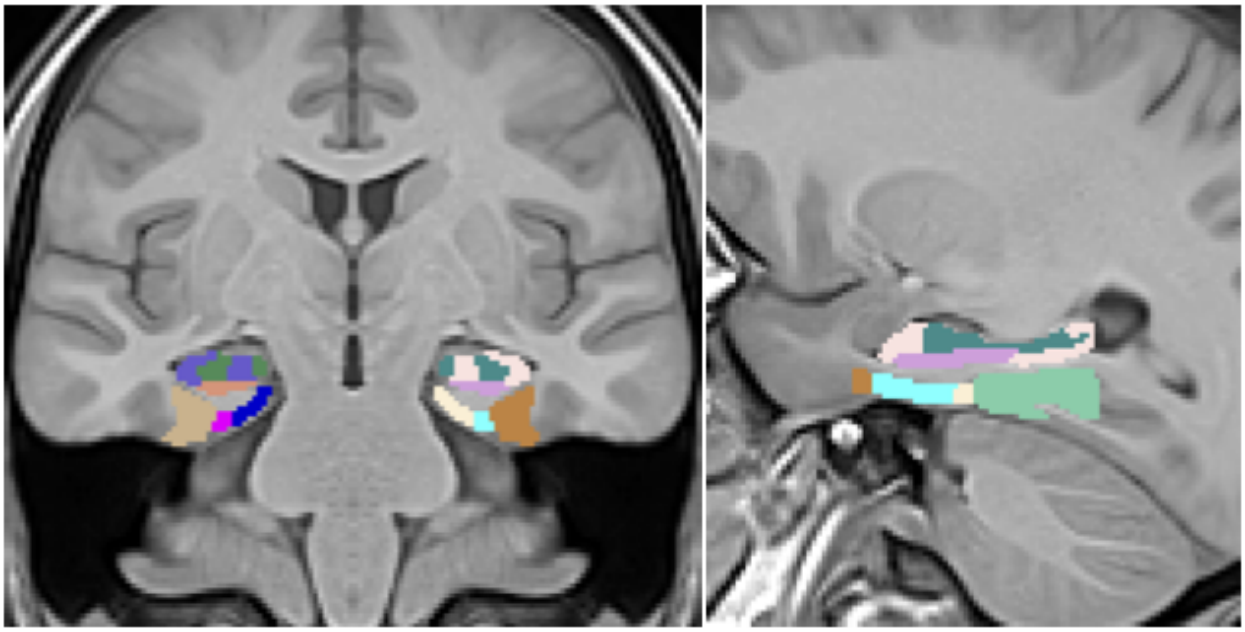

Figure 2: Coronal (left) and sagittal (right) views of the MTL parcellation generated from DeepFLASH superimposed on the T1 template component (T2 not shown). The template pose is oriented analogously to hippocampal specific MR acquisition protocols for a tailored segmentation domain.

These 17 image pairs and their contralateral counterparts were used to create an optimal symmetric multivariate template [Avants et al., 2010; Tustison et al., 2015] in the orientation of the original T2 acquisition (i.e., perpendicular to the hippocampal axis) as illustrated in Figure 2. Each of the pairwise normalized images and corresponding labels were subsequently affinely registered to the symmetric template where the labels were used to create average prior probability images to be used as additional channel inputs for training/prediction. The ROIs for training/prediction for both left and right hippocampi are contralaterally symmetric within the template domain and are of size $64 \times 64 \times 96$ voxels (voxel size = $1mm^3$).

Intensity-based data augmentation consisted of randomly added noise based on ITK functionality, simulated bias fields based on N4 bias field modeling [Tustison et al., 2010], and histogram warping for mimicking well-known MRI intensity nonlinearities [Nyúl and Udupa, 1999; Tustison et al., 2021a]. These augmentation techniques are available in ANTsXNet (Python and R):

- simulated bias field
  - `simulate_bias_field.py`
  - `simulateBiasField.R`
- added noise
  - `add_noise_to_image.py`
  - `addNoiseToImage.R`
- MRI intensity nonlinearities
  - `histogram_warp_image_intensities.py`
  - `histogramWarpImageIntensities.R`

Shape-based data augmentation used both random linear and nonlinear deformations after whole-brain affine normalization to the template. This functionality is also instantiated within ANTsXNet:

- random spatial warping
  - `randomly_transform_image_data.py`
  - `randomlyTransformImageData.R`

Additionally, given the well-studied variation of hippocampal volume with age [Bussy et al., 2021; Driscoll et al., 2003] and disease (e.g., [Henneman et al., 2009]), the original 17 image sets were warped to additional age/gender templates created from the public IXI database (i.e., 20-30, 30-40, 50-60, 70-80, 80-90 and male/female) and ADNI templates made from diagnostic cohorts (i.e., normal, mild cognitive impairment, and Alzheimer's disease). This resulted in an extended training data set of an additional 288 subjects ($\times 2$ by including the contralateral versions). Prior to training, these data were skull stripped using ANTsXNet functionality and affinely warped to a brain extraction version of the DeepFLASH templates. From these training data, random intensity and shape data augmentations, as previously described, were generated on the fly over 200 epochs with a batch size of 16.

During the early training stages, a combined loss function incorporated the weighted categorical entropy for the multilabel first output and the binary dice coefficient for the remaining binary label regional outputs. Later stages replaced the weighted categorical entropy with a multilabel dice coefficient loss for the first output (with other output loss functions remaining the same). Additionally, earlier stages weighted more heavily the contribution of the binary regional outputs over the multilabel first output which was reversed in later stages. All training was performed on a DGX (GPUs: 4X Tesla V100, system memory: 256 GB LRDIMM DDR4) using ANTsXNet functionality with the training scripts available with the GitHub repository associated with this work [Tustison,].

Voxelwise Label Prediction

Given the bilateral symmetry in both the training data and training protocol, two sets of networks, left and right, were trained using both T1 and combined T1/T2 modalities. This symmetry prevents a prior volumetric bias from being encoded within the segmentation framework which limits biasing of hemispherical asymmetry in hippocampal volumes under conditions of aging and pathology [Rogers et al., 2012; Sarica et al., 2018; Thompson et al., 2009]. An additional advantage is that two trained networks are available for predicting both MTL regions by contralaterally flipping the results and averaging the estimated probability maps.

For predicting the MTL parcellation, preprocessing includes N4 bias correction [Tustison et al., 2010], brain extraction [Tustison et al., 2021a], affine warping [Avants et al., 2011a; Avants et al., 2014] to the brain extracted DeepFLASH template, and cropping to the hippocampal ROIs in the space of the template. For intensity matching within the respective ROIs, two different protocols were explored: rank intensity transform of the input images (e.g., intensity values of [0, 1, 2, 255] are converted to [0, 1, 2, 3]) and histogram matching [Nyúl and Udupa, 1999] to the MTL region of the corresponding modality template. Although producing similar results in our preliminary qualitative evaluations using a wide dataset sampling, we use the former for UKBB. However, both options are available to the user. During prediction, the U-net network is constructed internally and the requisite additional data (i.e., parameter weights and templates) are automatically downloaded from Figshare

and cached using TensorFlow/Keras utilities.

2.4.4 White matter hyperintensity segmentation

Although UKBB includes white matter hyperintensity segmentation masks [Alfaro-Almagro et al., 2018] derived from FMRIB's BIANCA tool [Griffanti et al., 2016], a recently developed WMH segmentation framework from the "SYSU" team [Li et al., 2018] was imported into the ANTsXNet libraries for WMH segmentation. As discussed in [Kuijf et al., 2019], this was the top performing algorithm at the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) held in 2017. Image data from five sites were used for both training and testing of segmentation algorithms from 20 different teams. Both the architecture and ensemble weights were made publicly available by the SYSU team which permitted a direct porting into ANTsXNet.

2.5 Predictive modeling for IDP characterization

Insight into the relationships between neurostructural and phenotypic measures is often possible through predictive modeling of sociodemographic targets and neuroimaging biomarkers. Many strategies for data exploration leverage standardized quantities derived from existing pipelines, which constitutes a form of dimensionality reduction or feature extraction based on clinically established relevance. Such tabulated data has several advantages over direct image use including being relatively easier to access, store, and manage. Analyses with off-the-shelf statistical packages is also greatly simplified. Additionally, using standardized features in predictive modeling, where feature importance is a component of the analysis, significantly facilitates the clinical interpretability of the modeling process.

Herein, we compare predictive modeling frameworks using tabular data. Baseline comparisons include standard linear regression where linear dependencies between covariates were resolved using `findLinearCombos` of the `caret` R package [Kuhn, 2008]. Although the number of observations relative to the number of covariates for the specified models is sufficiently large, given the constraints of traditionally-sized data sets, we also compared results based

on sparse linear regression, specifically, the Lasso method as found in the `glmnet` R package [Friedman et al., 2010], using recommended parameters.

We also evaluated two popular packages for use with tabular data, viz. XGBoost [Chen and Guestrin, 2016] and TabNet [Arik and Pfister, 2021]. The former is a popular implementation of gradient boosted decision trees known for superb performance and computational efficiency. Similar to random forests [Breiman, 2001], gradient boosted decision trees leverage ensembles of weak classifiers (i.e., individual decision trees) to enhance model accuracy. However, in contrast to random forests, which generate weak classifiers using random initialization, gradient boosted decision trees are constructed in stages based on the gradient of a specified cost function [Friedman, 2001]. The following common hyperparameters were used: maximum tree depth = 6, number of rounds = 1000 (with early stopping after 10 rounds of no improvement), squared error as the loss function for regression targets, and logistic and softmax losses for binary and multi-label classification problems, respectively.

TabNet is a deep learning framework specifically engineered for structured tabulated data which incorporates sparsity considerations as well as providing feature importance for interpretability. We used an established PyTorch implementation of TabNet [McTabberson,] with the default parameters which is reported to demonstrate good performance on a variety of predictive problem types. While deep learning methods have proven both highly popular and effective within neuroimaging research, much of this work has been restricted to convolutional neural networks for image-based analyses, as opposed to parallel research with tabular data, hence the motivation for the development of TabNet. Much more generally, this is a current topic of interest for the larger machine learning community (see, for example, [Gorishniy et al., 2021; Kadra et al., 2021; Shwartz-Ziv and Armon, 2022]). As a final baseline comparison, we used a basic dense neural network (DenseNet) consisting of two hidden layers with 512 units in the first layer and 256 units in the second layer using leaky ReLU activation where the output layer employed linear activation for regression and sigmoid/softmax for binary/multi classification.

For each of these predictive modeling frameworks, we selected several target variables for our comparative evaluation (cf. Table 1) and generated models of the form:

Table 1: Set of UKBB sociodemographic targets for evaluation.

Target	**Data ID**	**Brief Description**
Age	21003-2.0	Age (years) at imaging visit
Fluid intelligence score	20191-0.0	Number of correct answers (of 13)
Neuroticism score	20127-0.0	Summary of 12 behaviour domains
Numeric memory	20240-0.0	Maximum digits remembered correctly
Body mass index	21001-2.0	Impedance-based body composition
Townsend deprivation index	189-0.0	Material deprivation measure
Genetic sex	22001-0.0	Sex from genotyping
Hearing difficulty	2247-2.0	Any difficulty? Yes/No
Risk taking	2040-2.0	Do you take risks? Yes/No
Same sex intercourse	2159-2.0	Ever had? Yes/No
Smoking	1249-2.0	Daily, occasionally, or ≤ 2 times?
Alcohol	1558-2.0	Six frequency categories†

†: Daily, 3-4 times/week, 1-2 times/week, 1-3 times/month, special occasions only, and never.

$$Target \sim Age + Genetic\ Sex + \sum_{i=1}^{N} IDP_i \tag{1}$$

where i indexes over the set of N IDPs for a particular grouping. In the cases where *Age* or *Genetic Sex* is the target variable, it is omitted from the right side of the modeling equation.

Assessment of the model groups based on the three individual sets of IDPs and their combination employs standard quality measures: area under the curve (AUC) for classification targets and root-mean-square error (RMSE) for regression targets. We also explored individual IDP importance through the use of model-specific parameter assessment metrics such as the absolute value of the t-statistic for linear models and SHAP values [Lundberg and Lee, 2017] for neural networks.

3 Results

3.1 Package-wise Group IDP Comparison

To compare the groups of IDPs, we used the three IDP sets (FSL, FreeSurfer, ANTsX) and their combination ("All") to train predictive models using the preselected target sociodemographic variables from Table 1. We first revisit a previous evaluative framework of ANTsX cortical thickness values by comparing their ability to predict *Age* and *Genetic Sex* with corresponding FreeSurfer cortical thickness values [Tustison et al., 2021b].

Following this initial comparative analysis, ten-fold cross validation, using random training/evaluation sampling sets (90% training/10% evaluation), per IDP set per target variable per machine learning technique (i.e., linear regression, Lasso, XGBoost, DenseNet, and TabNet) was used to train and evaluate the models described by Equation (1).

3.1.1 Revisiting ANTs and FreeSurfer cortical thickness comparison

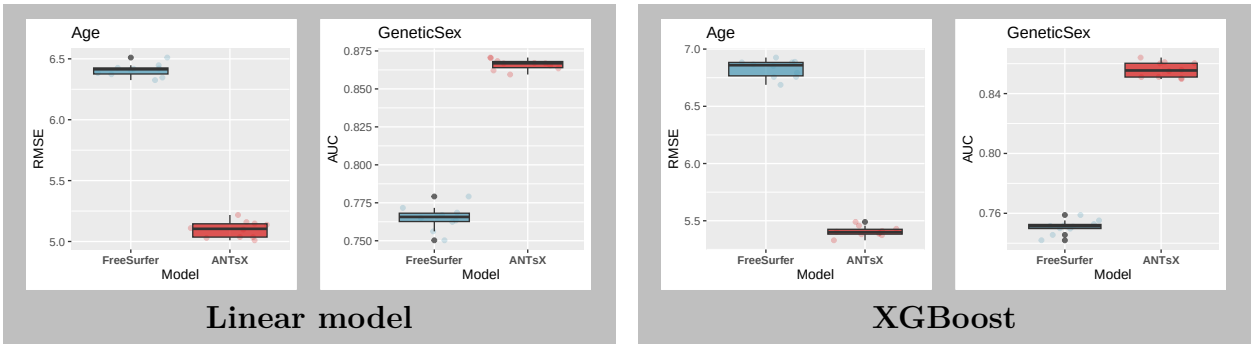

Figure 3: (Left) Linear and (right) XGBoost model results for predicting *Age* and *Genetic Sex* using both ANTsX and FreeSurfer cortical thickness data averaged over the 62 cortical regions of the DKT parcellation. RMSE and AUC were used to quantify the predictive accuracy of *Age* and *Genetic Sex*, respectively.

In [Tustison et al., 2014; Tustison et al., 2021b], IDPs under consideration were limited to ANTsX-based and FreeSurfer cortical thickness measurements averaged over the 62 regions of the DKT parcellation. These IDP sets were specifically compared in terms of the predictive capability vis-à-vis *Age* and *Genetic Sex*. With respect to UKBB-derived cortical thickness

Table 2: Summary statistics for the selected continuous UKBB sociodemographic target variables.

Target	**Mean ± SD**	**Range**
Age	63.97 ± 7.67	$[45, 82]$
Fluid intelligence score	6.63 ± 2.03	$[0, 13]$
Neuroticism score	3.81 ± 3.17	$[0, 12]$
Numeric memory	7.02 ± 1.44	$[2, 11]$
Body mass index	26.47 ± 4.36	$[13.4, 58.0]$
Townsend deprivation index	-1.89 ± 2.73	$[-6.26, 10.10]$

IDPs, similar analysis using both linear and XGBoost models demonstrates consistency with prior results (see Figure 3).

3.1.2 Package IDP comparison via continuous target variables

Predictive models for cohort *Age*, *Fluid Intelligence Score*, *Neuroticism Score*, *Numeric Memory*, *Body Mass Index*, and *Townsend Deprivation Index* were generated and evaluated as described previously. Summary statistics for these variables are provided in Table 2. The resulting accuracies, in terms of RMSE, are provided in Figure 4.

Linear and Lasso models provide the most consistently accurate results across the set of continuous target variables with the combined set of IDPs performing well for the majority of cases. All linear models demonstrate significant correlations across IDP sets (cf. Figure 5). Despite the large number of regressors, sparsity-based constraints did not significantly improve prediction performance. The deep learning models (both DenseNet and TabNet) performed similarly although were only competitive for selected subsets (e.g., *Neuroticism Score* and *Townsend Deprivation Index*). Although XGBoost performed well for the commonly studied *Age* target variable, performance measures were relatively much less accurate for the remaining categories. This could be the results of suboptimal hyperparameter choice with respect to these other categories but, as with the other techniques, this was not investigated further.

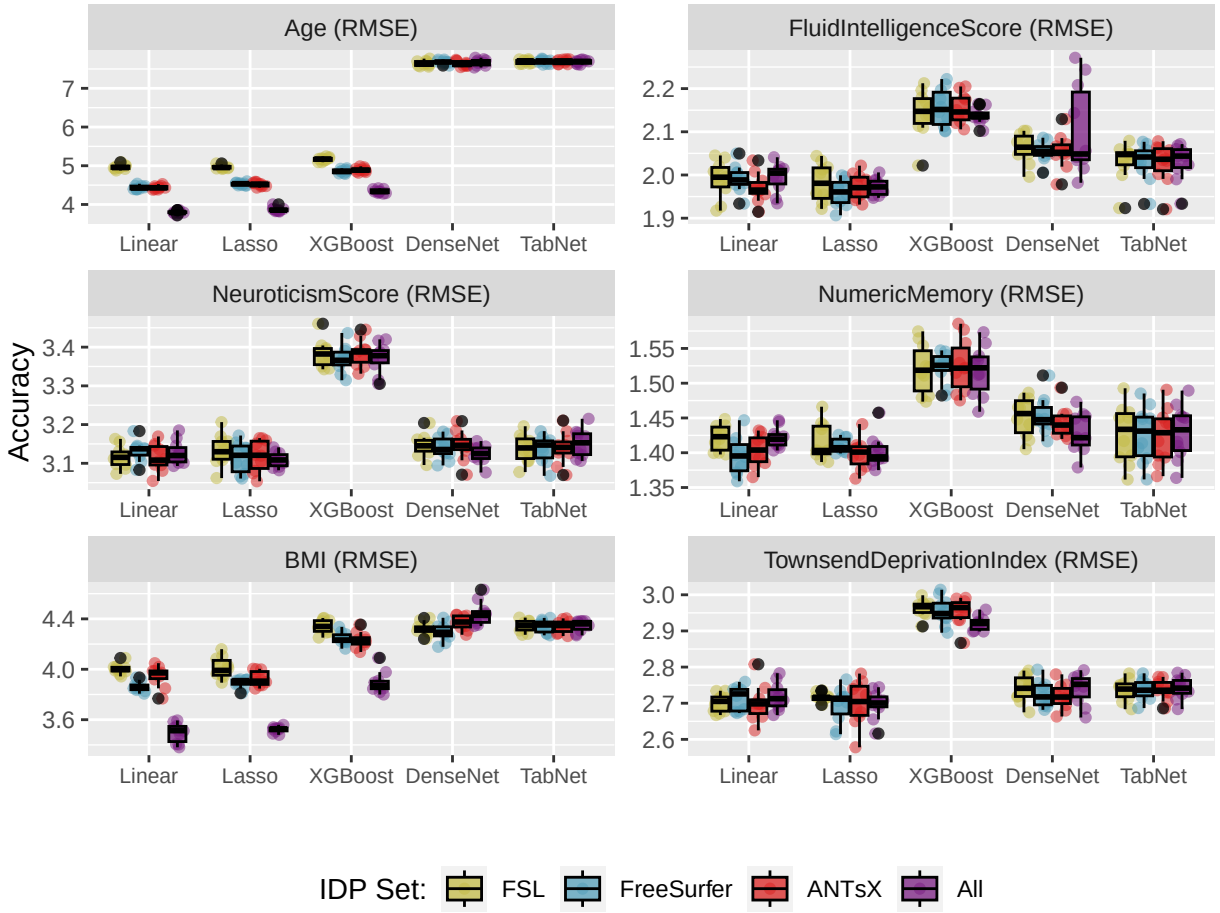

Figure 4: Comparison of machine learning frameworks for training and prediction of selected continuous UKBB sociodemographic continuous variables (cf. Table 1) with the different IDP sets and their combination (FSL, FreeSurfer, ANTsX, and All).

3.1.3 Package IDP comparison via categorical target variables

Predictive models for cohort categories associated with *Genetic Sex*, *Hearing Difficulty*, *Risk Taking*, *Same Sex Intercourse*, *Smoking Frequency*, and *Alcohol Frequency* were generated and evaluated as described previously. The resulting accuracies, in terms of binary or multi-class AUC, are provided in Figure 6.

Similar to the continuous variables, the Linear and Lasso models perform well for most of the target variables. Superior performance is seen for these models predicting *Genetic Sex* along with XGBoost. DenseNet performs similarly well as the Linear and Lasso models for *Hearing*, *Risk Taking*, *Same Sex Intercourse*, and *Alcohol*. DenseNet models are also superior

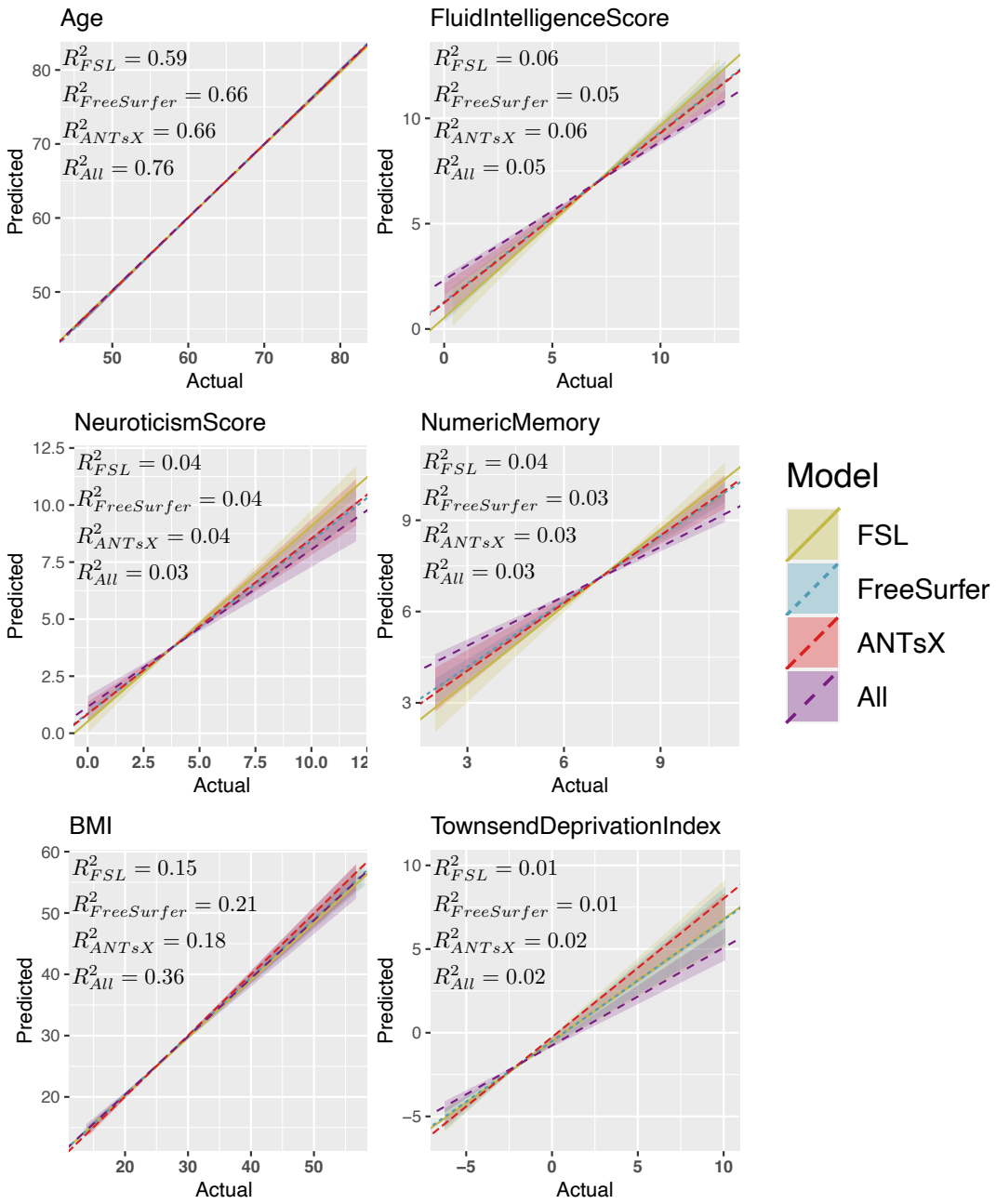

Figure 5: Regression regions defined by the linear models represented in Figure 4 showing the relationship between the predicted and actual target values. We also plot the median line for each model-based grouping as defined by the slope and list the average R^2 values for each IDP set.

performers for predicting *Smoking* categories.

Table 3: Top 10 features for *Age*, *Fluid Intelligence Score*, and *Neuroticism Score* target variables specified based on the specific, top-performing machine learning techniques for the combined (i.e., All) IDP set.

Package	**IDP Subset**	**Individual IDP**
	Age	
ANTsX	Atropos	Volume of deep GM
FreeSurfer	ASEG	Volume of right ventral DC
FreeSurfer	ASEG	Volume of right putamen
FreeSurfer	ASEG	Volume of left thalamus-proper
FSL	Other	Volume of ventricular CSF
FreeSurfer	ASEG	Volume of left ventral DC
FreeSurfer	ASEG	Volume of right thalamus-proper
ANTsX	Atropos	Volume of white matter
ANTsX	Region	Volume of left ventral DC
ANTsX	Region	Volume of right ventral DC
	Fluid Intelligence	
FreeSurfer	ASEG	Volume of CC-Anterior (whole brain)
–	–	Genetic sex
ANTsX	Thickness	Thickness of left middle temporal
FSL	FAST	Volume of GM in right insular cortex
ANTsX	Thickness	Thickness of left rostral anterior cingulate
FSL	Other	Volumetric scaling from T1 head image to standard space
FSL	FAST	Volume of GM in left middle temporal gyrus (anterior)
ANTsX	Thickness	Thickness of left fusiform
FreeSurfer	Thickness	Thickness of right entorhinal
FSL	FAST	Volume of GM in left ventral striatum
	Neuroticism	
–	–	Age
–	–	Genetic sex
FSL	FAST	Volume of GM in left subcallosal cortex
ANTsX	Cortical	Volume of right lingual
ANTsX	Thickness	Thickness of right pars orbitalis
FSL	FAST	Volume of GM in left frontal medial cortex
ANTsX	Region	Volume of right lingual
FreeSurfer	ASEG	Volume of CC-Mid-Anterior
FSL	FAST	Volume of GM in left temporal occipital fusiform
ANTsX	Cortical	Volume of right inferior parietal

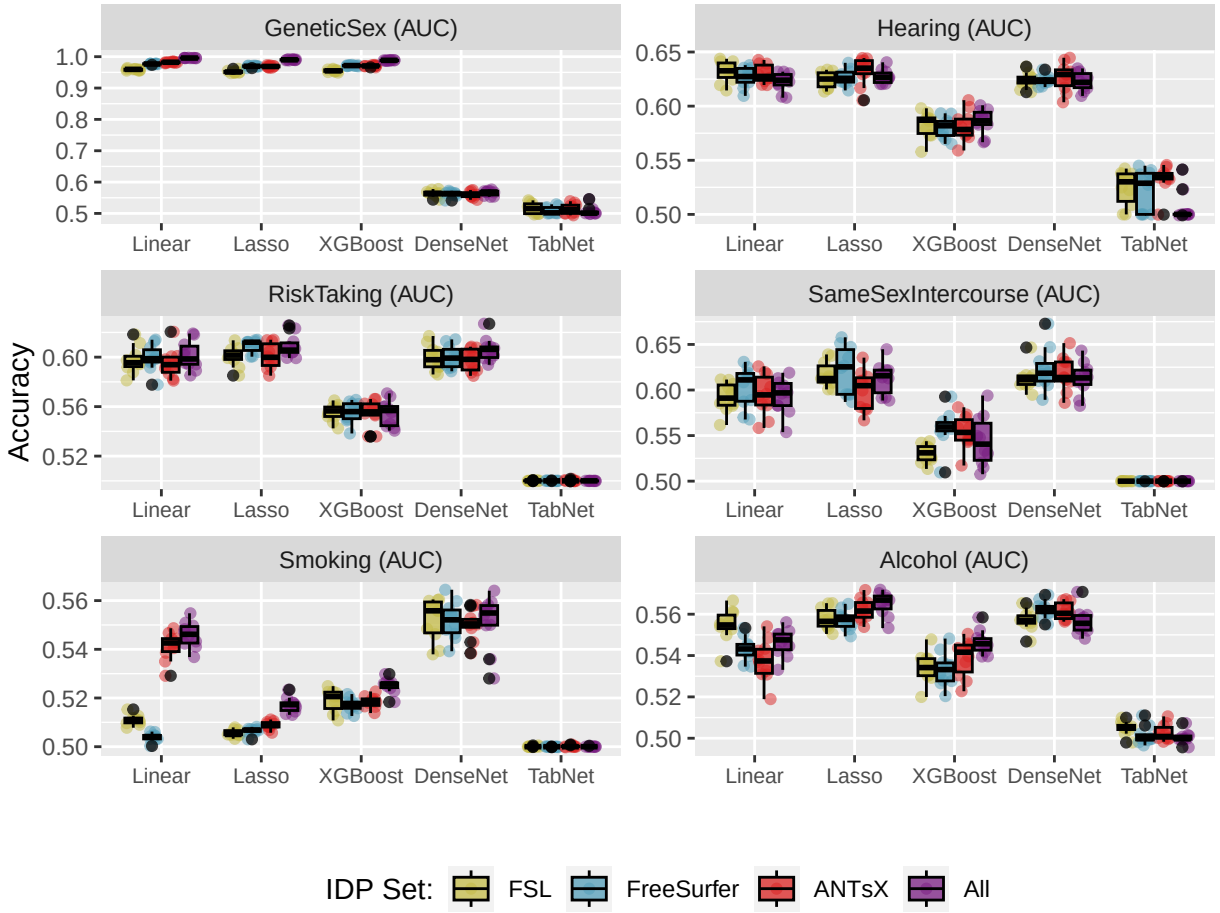

Figure 6: Comparison of machine learning frameworks for training and prediction of selected binary and multilabel categorical UKBB sociodemographic variables (cf. Table 1) with the different IDP sets and their combination (FSL, FreeSurfer, ANTsX, and All). *Smoking* and *Alcohol* target variables have more than two labels.

3.2 Individual IDP comparison

To compare individual IDPs, for each target variable, we selected the set of results corresponding to the machine learning technique which demonstrated superior performance, in terms of median predictive accuracy, for the combined (All) IDP grouping. The top ten features for the principle continuous variables of *Age*, *Fluid Intelligence Score*, and *Neuroticism Score* are listed in Table 3 and ranked according to variable importance score (specifically, absolute t-statistic value for linear models). The ranked lists are also color-coded by IDP package. For additional insight into individual IDPs, full feature lists with feature importance rankings are available for all target variables in the supplementary material hosted at

the corresponding GitHub repository [Tustison,].

4 Discussion

Much UKBB research is made possible through the availability of its characteristic large-scale, subject-specific epidemiological data, including IDPs and enhanced by the stringent data acquisition protocols to ensure consistency across sites. In this work, we complement the existing FSL- and FreeSurfer-based UKBB IDPs with the generation and potential distribution of corresponding ANTsX-derived IDPs. These latter IDPs were generated from well-vetted pipelines that have been used in previous research and are publicly available through the ANTsX ecosystem. By providing these IDP-producing utilities within high-level languages, such as Python and R, in a comprehensive, open-source package, we are able to leverage the computational efficiency of deep learning libraries while also leveraging the numerous packages available for the curation, analysis, and visualization of tabulated data.

In addition to the availability of these ANTsX UKBB IDPs, we explored their utility with respect to other package-specific groupings and their combinations. For exploration of these IDP group permutations, we used popular machine learning algorithms to predict commonly studied sociodemographic variables of current research interest (Table 1). In addition to research presentation in traditional venues, at least two of these target variables, specifically *Age* and *Fluid Intelligence*, have been the focus of two recent competitions.

Regarding the former, research concerning brain age estimation from neuroimaging is extensive and growing (cf. recent reviews [Baecker et al., 2021; Franke and Gaser, 2019; Mishra et al., 2021]). It was also the subject of the recent Predictive Analytics Competition held in 2019 (PAC2019). This competition featured 79 teams leveraging T1-weighted MRI with a variety of quantitative approaches from convolutional neural networks (CNNs) to common machine learning frameworks based on morphological descriptors (i.e., structural IDPs) derived from FreeSurfer [Lombardi et al., 2020]. The winning team [Gong et al., 2021], using an ensemble of CNNs and pretrained on a UKBB cohort of $N = 14,503$ subjects, had a

mean absolute error (MAE) of 2.90 years. Related CNN-based deep learning approaches achieved comparable performance levels and simultaneously outperformed more traditional machine learning approaches. For example, the FreeSurfer IDP approach using a dense neural network [Lombardi et al., 2020] yielded an overall MAE accuracy of 4.6 years. Alternative strategies based on Lasso, Random Forests, and support vector regression techniques were attempted but did not achieve similar accuracy levels.

Given that RMSE provides a general upper bound on MAE (i.e., MAE $\leq$ RMSE), the accuracy levels yielded by our FSL, FreeSurfer, ANTsX, and combined linear, lasso, and XGBoost models can be seen from Figure 4 to perform comparatively well. The FreeSurfer and ANTsX linear models performed similarly with RMSE prediction values of approximately 4.4 years whereas FSL was a little higher at 4.96 years. However, combining all IDPs resulted in an average RMSE value of 3.8 years. When looking at the top 10 overall linear model features (Table 3) ranked in terms of absolute t-statistic value, all three packages are represented and appear to reflect both global structures (white matter and CSF volumes) and general subcortical structural volumes (ANTsX "deep GM" and both FreeSurfer and ANTsX bi-hemispherical ventral dienchephalon volumes). Increases in CSF volume and ventricular spaces is well-known to be associated with brain shrinkage and aging [Matsumae et al., 1996; Murphy et al., 1992; Scahill et al., 2003].

Similarly, the association between brain structure and fluid intelligence has been well-studied [Vieira et al., 2022] despite potentially problematic philosophical and ethical issues [Eickhoff and Langner, 2019]. With intentions of furthering this research, the ABCD Neurocognitive Prediction Challenge (ABCD-NP-Challenge) was held in 2019 which concerned predicting fluid intelligence scores (using the NIH Toolbox Cognition Battery [Weintraub et al., 2013]) in a population of 9-10 year pediatric subjects using T1-weighted MRI. Fluid intelligence scores were residualized from brain volume, acquisition site, age, ethnicity, genetic sex, and parental attributes of income, education, and marriage (additional data processing details are provided in the Data Supplement [Pfefferbaum et al., 2018]).

Of the 29 submitting teams, the first place team of the final leaderboard employed kernel ridge regression with voxelwise features based on the T1-weighted-based probabilistic tissue

segmentations specifically, CSF, gray matter, and white matter— both modulated and unmodulated versions for a total of six features per subject. In contrast to the winning set of predictive sparse and global features, the second place team used 332 total cortical, subcortical, white matter, cerebellar, and CSF volumetric features. Although exploring several machine learning modeling techniques, the authors ultimately used an ensemble of models for prediction which showed improvement over gradient boosted decision trees. From Table 3, most predictive features from our study, regardless of package, are localized measures of gray matter.

Although the stated, primary objective of these competitions is related to superior performance in terms of algorithmic prediction of quantitative sociodemographics, similar to the evaluation strategy used in this work, outside of the clinical research into brain age estimation, none of these performance metric reach the level of individual-level prediction. Consequently, these may be more informative as an interpretation of the systems- level relationship between brain structure and behavior. An obvious secondary benefit is the insight gained into the quality and relevance of measurements and modeling techniques used. In this way, these considerations touch on fundamental implications of the No Free Lunch Theorems for search and optimization [Wolpert and Macready, 1997] where prior distributions (i.e., correspondence of measurements and clinical domain for algorithmic modeling) differentiate general performance. Relatedly, although all packages are represented amongst the top-performing IDPs, their relative utility is dependent, expectedly so, on the specific target variable, and, to a lesser extent, on the chosen machine learning technique. Such considerations should be made along with other relevant factors (e.g., computational requirements, open-source availability) for tailored usage.

5 Conclusion

The UK Biobank is an invaluable resource for large-scale epidemiological research which includes a thorough neuroimaging battery for a significant subset of the study volunteers. For quantitative exploration and inference of population trends from leveraging imaging

data, well-vetted measurement tools are essential. The primary contribution that we have described is the generation and public availability of the set of UK Biobank neuroimaging structural IDPs generated using the ANTsX ecosystem. These ANTsX IDPs, which includes DeepFLASH for hippocampal and extra-hippocampal parcellation, complement the existing sets of FSL and FreeSurfer IDPs. A predictive modeling strategy using a variety of sociodemographic target variables was used to explore IDP viability, importance, and utility via the modeling capabilities of several machine learning techniques with linear regression demonstrating overall good performance.

Supplementary material

Deep FLASH IDPs of UKBB

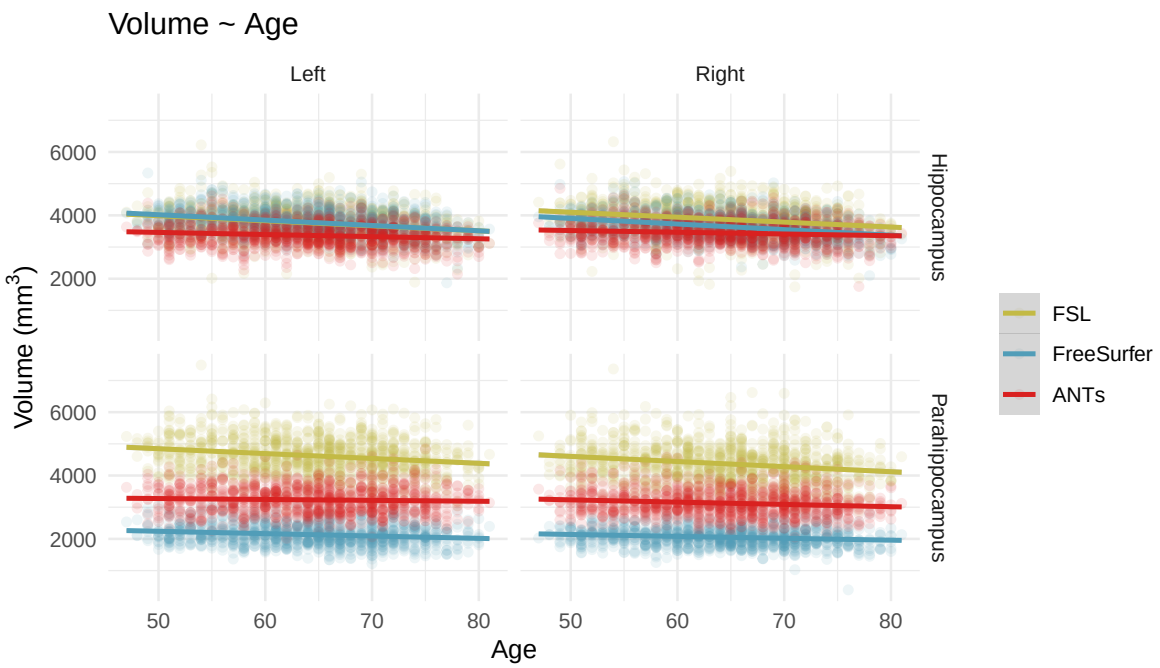

Figure 7: A random sample of $n = 1000$ UKBB imaging subjects showing the cross-sectional age trends of hemispherical hippocampal (top) and parahippocampal (bottom) volumes from the three sets of IDPs. Although hippocampal volumes appear to be relatively consistent across FSL, FreeSurfer, and ANTs; parahippocampal volumes, in contrast, appear to exhibit systematic discrepancies.

Although earlier versions of DeepFLASH have been evaluated and used in multiple studies (as previously cited), UKBB provides a unique platform for evaluating the recently developed deep learning variant that is unique in terms of available demographics, inter-package IDP definition and availability, as well as scale. Such data permits (and necessitates)[2] alternative evaluation strategies different from a conventional approaches involving segmentation overlap measures (e.g., Dice, Hausdorff). For example, the only volumetric IDPs in common between

[2]Aside from the usual constraints of available gold- or silver-standard data of sufficient scale for common segmentation voxelwise assessment strategies, the unavailability of FreeSurfer processed data makes such comparisons difficult. Although we could, in theory, process the data ourselves through the corresponding FreeSurfer pipeline (in addition to the previous joint label fusion version of DeepFLASH), such intensive large-scale processing for such a limited purpose, where alternative strategies are viable, would be excessively time consuming and environmentally inconsiderate [Lannelongue et al., 2021].

all three sets are the hippocampus and the parahippocampus (see Figure 7) with there appearing to be much less discrepancy between the former versus the latter. As mentioned previously, these differences are not indicative of inaccuracy, per se, but likely more a result of varying anatomical definitions.

To accommodate these issues within the context of UKBB, we first explore how DeepFLASH hippocampal and parahippocampal volumetric IDPs compare to the corresponding FSL and FreeSurfer quantities in the context of three principal demographic variables available in UKBB (i.e., age [Bussy et al., 2021; Echávarri et al., 2011], fluid intelligence [Reuben et al., 2011], and neuroticism [Liu et al., 2021]) [Dadi et al., 2021] with appropriate confound modeling [Alfaro-Almagro et al., 2021]. We then demonstrate their significance with respect to the unique hierarchical parcellation of DeepFLASH IDPs.

Interset IDP evaluation

Table 4 summarizes the package-specific statistical relationships between hippocampal and parahippocampal volumes (with the prescribed confounds) and the well-studied demographic target variables of age, fluid intelligence score, and neuoroticism score [Dadi et al., 2021]. Similar trends are seen across all IDPs although age demonstrates the greatest mutual disparity in the hippocampal and parahippocampal regions.

DeepFLASH Hierarchical IDP comparison

As mentioned previously, DeepFLASH produces three separate outputs structured based on the specified anatomical hierarchy. Similar to the inter-package analysis found in the previous section, we used regression models to compare the different DeepFLASH output regions for the demographic target variables of age, fluid intelligence score, and neuoroticism score [Dadi et al., 2021]. Specifically, we model

$$Age \sim \{Region\} + Sex + Brain + Site \tag{2}$$

Table 4: Brain structural and demographic modeling of hippocampal and parahippocampal volumes across IDP sets. Coefficients (and standard errors) along with statistical significance designation for the regressors of interest are provided.

	FSL	**FreeSurfer**	**DeepFLASH**
$Age \sim Hippocampal + Parahippocampal + Sex + Brain + Site$			
Hippocampal	-0.002^{***} (0.00005)	-0.004^{***} (0.0001)	-0.002^{***} (0.0001)
Parahippocampal	-0.001^{***} (0.00005)	-0.001^{***} (0.0001)	-0.001^{***} (0.0001)
Sex	3.748^{***} (0.088)	2.908^{***} (0.086)	3.810^{***} (0.090)
Brain	-0.00000^{***} (0.00000)	0.00001^{***} (0.00000)	-0.00001^{***} (0.00000)
R^2	0.151	0.199	0.096
F Statistic	$1{,}047.043^{***}$	$1{,}463.211^{***}$	624.199^{***}
$FluidIntelligence \sim Hippocampal + Parahippocampal + Age + Age^2 + Sex + Age * Sex + Brain + Site$			
Hippocampal	0.0001^{***} (0.00002)	0.0001^{***} (0.00003)	0.0001^{***} (0.00003)
Parahippocampal	0.00004^{**} (0.00002)	0.00004 (0.00003)	0.00001 (0.00002)
Age	-0.010^{***} (0.003)	-0.010^{***} (0.003)	-0.011^{***} (0.003)
Sex	0.273 (0.231)	0.311 (0.231)	0.332 (0.230)
Brain	0.00000^{***} (0.00000)	0.00000^{***} (0.00000)	0.00000^{***} (0.00000)
Age:Sex	-0.009^{**} (0.004)	-0.009^{**} (0.004)	-0.010^{***} (0.004)
R^2	0.046	0.045	0.045
F Statistic	117.017^{***}	115.656^{***}	115.108^{***}
$Neuroticism \sim Hippocampal + Parahippocampal + Age + Age^2 + Sex + Age * Sex + Brain + Site$			
Hippocampal	0.00002 (0.00002)	0.0001^{*} (0.00003)	0.0001^{***} (0.00003)
Parahippocampal	-0.00005^{**} (0.00002)	0.00003 (0.00003)	0.00001 (0.00003)
Age	-0.047^{***} (0.003)	-0.044^{***} (0.003)	-0.045^{***} (0.003)
Sex	-0.499^{*} (0.285)	-0.522^{*} (0.284)	-0.519^{*} (0.284)
Brain	-0.00000^{***} (0.00000)	-0.00000^{***} (0.00000)	-0.00000^{***} (0.00000)
Age:Sex	-0.003 (0.004)	-0.003 (0.004)	-0.003 (0.004)
R^2	0.034	0.034	0.034
F Statistic	136.059^{***}	136.031^{***}	136.368^{***}

Note: Site,constant effects not listed. *p<0.1; **p<0.05; ***p<0.01

Table 5: Hierarchical regional regression coefficients (and their standard errors) are illustrated with respect to the target variables of age, neuroticism score, and fluid intelligence score.

	Age	**FluidIntelligence**	**Neuroticism**
MTL	-0.00052 (0.00003)***	0.00001 (0.00001)	0.00003 (0.00001)
hippocampus	-0.00181 (0.00007)***	0.0001 (0.00003)***	0.00009 (0.00003)*
extra.hippocampal	-0.00036 (0.00003)***	0 (0.00001)	0.00002 (0.00001)
DG.CA2.CA3.CA4	-0.00425 (0.00022)***	0.00026 (0.00008)**	-0.00007 (0.0001)
CA1	0.00214 (0.00012)***	0.00008 (0.00004)	0.00016 (0.00006)**
subiculum	-0.00818 (0.00027)***	0.00026 (0.0001)*	0.00055 (0.00013)***
aLEC	0.00436 (0.00015)***	-0.00009 (0.00005)	0.00016 (0.00007)
pMEC	0.01244 (0.0004)***	0.00023 (0.00015)	0.0004 (0.00019)
perirhinal	0.00004 (0.00004)	-0.00002 (0.00002)	0.00002 (0.00002)
parahippocampal	-0.00129 (0.00006)***	0.00003 (0.00002)	0.00002 (0.00003)

*p<0.1; **p<0.05; ***p<0.01

and

$$\begin{aligned} FluidIntelligence \text{ or } Neuroticism \sim & \{Region\} + Age + Age^2 \\ & + Sex + Age * Sex + Brain + Site \end{aligned} \tag{3}$$

where $\{Region\}$ is one of the hierarchical DeepFlash outputs listed in the left column of Table 5 along with the regression coefficients, standard errors, and statistical significance associated with each region.

Acknowledgments

References

Alfaro-Almagro F, Jenkinson M, Bangerter NK, Andersson JLR, Griffanti L, Douaud G, Sotiropoulos SN, Jbabdi S, Hernandez-Fernandez M, Vallee E, Vidaurre D, Webster M, McCarthy P, Rorden C, Daducci A, Alexander DC, Zhang H, Dragonu I, Matthews PM, Miller KL, Smith SM (2018): Image processing and quality control for the first 10,000 brain imaging datasets from UK biobank. Neuroimage 166:400–424.

Alfaro-Almagro F, McCarthy P, Afyouni S, Andersson JLR, Bastiani M, Miller KL, Nichols TE, Smith SM (2021): Confound modelling in UK biobank brain imaging. Neuroimage 224:117002.

Arik SÖ, Pfister T (2021): TabNet: Attentive interpretable tabular learning. Proceedings of the AAAI Conference on Artificial Intelligence 35:6679–6687. https://ojs.aaai.org/index.php/AAAI/article/view/16826.

Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC (2011a): A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage 54:2033–44.

Avants BB, Tustison NJ, Stauffer M, Song G, Wu B, Gee JC (2014): The Insight ToolKit image registration framework. Front Neuroinform 8:44.

Avants BB, Tustison NJ, Wu J, Cook PA, Gee JC (2011b): An open source multivariate framework for n-tissue segmentation with evaluation on public data. Neuroinformatics 9:381–400.

Avants BB, Yushkevich P, Pluta J, Minkoff D, Korczykowski M, Detre J, Gee JC (2010): The optimal template effect in hippocampus studies of diseased populations. Neuroimage 49:2457–66.

Baecker L, Garcia-Dias R, Vieira S, Scarpazza C, Mechelli A (2021): Machine learning for brain age prediction: Introduction to methods and clinical applications. EBioMedicine 72:103600.

Breiman L (2001): Random forests. Machine Learning 45:5–32.

Brown ES, Kulikova A, Van Enkevort E, Nakamura A, Ivleva EI, Tustison NJ, Roberts J, Yassa MA, Choi C, Frol A, Khan DA, Vazquez M, Holmes T, Malone K (2019a): A randomized trial of an NMDA receptor antagonist for reversing corticosteroid effects on the human hippocampus. Neuropsychopharmacology 44:2263–2267.

Brown ES, Sayed N, Choi C, Tustison N, Roberts J, Yassa MA, Van Enkevort E, Nakamura A, Ivleva EI, Sunderajan P, Khan DA, Vazquez M, McEwen B, Kulikova A, Frol AB, Holmes T (2019b): A randomized, double-blind, placebo-controlled trial of lamotrigine for prescription corticosteroid effects on the human hippocampus. Eur Neuropsychopharmacol 29:376–383.

Bussy A, Patel R, Plitman E, Tullo S, Salaciak A, Bedford SA, Farzin S, Béland M-L, Valiquette V, Kazazian C, Tardif CL, Devenyi GA, Chakravarty MM (2021): Hippocampal shape across the healthy lifespan and its relationship with cognition. Neurobiol Aging 106:153–168.

Caliskan A, Bryson JJ, Narayanan A (2017): Semantics derived automatically from language corpora contain human-like biases. Science 356:183–186.

Chen T, Guestrin C (2016): XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. New York, NY, USA: ACM. KDD '16 pp 785–794. http://doi.acm.org/10.1145/2939672.2939785.

Cornelis MC, Wang Y, Holland T, Agarwal P, Weintraub S, Morris MC (2019): Age and cognitive decline in the UK biobank. PLoS One 14:e0213948.

Dadi K, Varoquaux G, Houenou J, Bzdok D, Thirion B, Engemann D (2021): Population modeling with machine learning can enhance measures of mental health. Gigascience 10.

Das SR, Avants BB, Grossman M, Gee JC (2009): Registration based cortical thickness measurement. Neuroimage 45:867–79.

Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, Albert MS, Killiany RJ (2006): An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions

of interest. Neuroimage 31:968–80.

Diamond KM, Rolfe SM, Kwon RY, Maga AM (2022): Computational anatomy and geometric shape analysis enables analysis of complex craniofacial phenotypes in zebrafish. Biol Open 11.

Ding AS, Lu A, Li Z, Galaiya D, Ishii M, Siewerdsen JH, Taylor RH, Creighton FX (2022): Automated extraction of anatomical measurements from temporal bone CT imaging. Otolaryngol Head Neck Surg:1945998221076801.

Douaud G, Lee S, Alfaro-Almagro F, Arthofer C, Wang C, McCarthy P, Lange F, Andersson JLR, Griffanti L, Duff E, Jbabdi S, Taschler B, Keating P, Winkler AM, Collins R, Matthews PM, Allen N, Miller KL, Nichols TE, Smith SM (2022): SARS-CoV-2 is associated with changes in brain structure in UK biobank. Nature.

Driscoll I, Hamilton DA, Petropoulos H, Yeo RA, Brooks WM, Baumgartner RN, Sutherland RJ (2003): The aging hippocampus: Cognitive, biochemical and structural findings. Cereb Cortex 13:1344–51.

Echávarri C, Aalten P, Uylings HBM, Jacobs HIL, Visser PJ, Gronenschild EHBM, Verhey FRJ, Burgmans S (2011): Atrophy in the parahippocampal gyrus as an early biomarker of alzheimer's disease. Brain Struct Funct 215:265–71.

Eickhoff SB, Langner R (2019): Neuroimaging-based prediction of mental traits: Road to utopia or orwell? PLoS Biol 17:e3000497.

Falk T, Mai D, Bensch R, Çiçek Ö, Abdulkadir A, Marrakchi Y, Böhm A, Deubner J, Jäckel Z, Seiwald K, Dovzhenko A, Tietz O, Dal Bosco C, Walsh S, Saltukoglu D, Tay TL, Prinz M, Palme K, Simons M, Diester I, Brox T, Ronneberger O (2019): U-net: Deep learning for cell counting, detection, and morphometry. Nat Methods 16:67–70.

Fischl B (2012): FreeSurfer. Neuroimage 62:774–81.

Fonov VS, Evans AC, McKinstry RC, Almli C, Collins DL (2009): Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. NeuroImage S102.

Franke K, Gaser C (2019): Ten years of BrainAGE as a neuroimaging biomarker of brain aging: What insights have we gained? Front Neurol 10:789.

Friedman JH (2001): Greedy function approximation: A gradient boosting machine. The Annals of Statistics 29:1189–1232. https://doi.org/10.1214/aos/1013203451.

Friedman J, Hastie T, Tibshirani R (2010): Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33:1–22. https://www.jstatsoft.org/v33/i01/.

Glasser MF, Sotiropoulos SN, Wilson JA, Coalson TS, Fischl B, Andersson JL, Xu J, Jbabdi S, Webster M, Polimeni JR, Van Essen DC, Jenkinson M, WU-Minn HCP Consortium (2013): The minimal preprocessing pipelines for the human connectome project. Neuroimage 80:105–24.

Gong W, Beckmann CF, Vedaldi A, Smith SM, Peng H (2021): Optimising a simple fully convolutional network for accurate brain age prediction in the PAC 2019 challenge. Front Psychiatry 12:627996.

Gorishniy Y, Rubachev I, Khrulkov V, Babenko A (2021): Revisiting deep learning models for tabular data. CoRR abs/2106.11959. https://arxiv.org/abs/2106.11959.

Griffanti L, Zamboni G, Khan A, Li L, Bonifacio G, Sundaresan V, Schulz UG, Kuker W, Battaglini M, Rothwell PM, Jenkinson M (2016): BIANCA (brain intensity AbNormality classification algorithm): A new tool for automated segmentation of white matter hyperintensities. Neuroimage 141:191–205.

Griffith GJ, Morris TT, Tudball MJ, Herbert A, Mancano G, Pike L, Sharp GC, Sterne J, Palmer TM, Davey Smith G, Tilling K, Zuccolo L, Davies NM, Hemani G (2020): Collider bias undermines our understanding of COVID-19 disease risk and severity. Nat Commun 11:5749.

Hagenaars SP, Harris SE, Davies G, Hill WD, Liewald DCM, Ritchie SJ, Marioni RE, Fawns-Ritchie C, Cullen B, Malik R, METASTROKE Consortium, International Consortium for Blood Pressure GWAS, SpiroMeta Consortium, CHARGE Consortium Pulmonary Group, CHARGE Consortium Aging and Longevity Group, Worrall BB, Sudlow CLM, Wardlaw JM, Gallacher J, Pell J, McIntosh AM, Smith DJ, Gale CR, Deary IJ (2016): Shared genetic aetiology between cognitive functions and physical and mental health in

UK biobank (n=112,151) and 24 GWAS consortia. Mol Psychiatry 21:1624–1632.

Henneman WJP, Sluimer JD, Barnes J, Flier WM van der, Sluimer IC, Fox NC, Scheltens P, Vrenken H, Barkhof F (2009): Hippocampal atrophy rates in alzheimer disease: Added value over whole brain volume measures. Neurology 72:999–1007.

Holbrook AJ, Tustison NJ, Marquez F, Roberts J, Yassa MA, Gillen DL, Alzheimer's Disease Neuroimaging Initiative (2020): Anterolateral entorhinal cortex thickness as a new biomarker for early detection of alzheimer's disease. Alzheimers Dement (Amst) 12:e12068.

Iglesias JE, Augustinack JC, Nguyen K, Player CM, Player A, Wright M, Roy N, Frosch MP, McKee AC, Wald LL, Fischl B, Van Leemput K, Alzheimer's Disease Neuroimaging Initiative (2015): A computational atlas of the hippocampal formation using ex vivo, ultra-high resolution MRI: Application to adaptive segmentation of in vivo MRI. Neuroimage 115:117–37.

Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM (2012): FSL. Neuroimage 62:782–90.

Kadra A, Lindauer M, Hutter F, Grabocka J (2021): Regularization is all you need: Simple neural nets can excel on tabular data. CoRR abs/2106.11189. https://arxiv.org/abs/2106.11189.

Kini LG, Thaker AA, Hadar PN, Shinohara RT, Brown M-G, Dubroff JG, Davis KA (2021): Quantitative [18]FDG PET asymmetry features predict long-term seizure recurrence in refractory epilepsy. Epilepsy Behav 116:107714.

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

Kuhn M (2008): Building predictive models in r using the caret package. Journal of Statistical Software 28:1–26. https://www.jstatsoft.org/index.php/jss/article/view/v028i05.

Kuijf HJ, Biesbroek JM, De Bresser J, Heinen R, Andermatt S, Bento M, Berseth M, Belyaev M, Cardoso MJ, Casamitjana A, Collins DL, Dadar M, Georgiou A, Ghafoorian M, Jin D, Khademi A, Knight J, Li H, Llado X, Luna M, Mahmood Q, McKinley R, Mehrtash

A, Ourselin S, Park B-Y, Park H, Park SH, Pezold S, Puybareau E, Rittner L, Sudre CH, Valverde S, Vilaplana V, Wiest R, Xu Y, Xu Z, Zeng G, Zhang J, Zheng G, Chen C, Flier W van der, Barkhof F, Viergever MA, Biessels GJ (2019): Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge. IEEE Trans Med Imaging 38:2556–2568.

Lannelongue L, Grealey J, Bateman A, Inouye M (2021): Ten simple rules to make your computing more environmentally sustainable. PLoS Comput Biol 17:e1009324.

Li H, Jiang G, Zhang J, Wang R, Wang Z, Zheng W-S, Menze B (2018): Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images. Neuroimage 183:650–665.

Liu X, Lai H, Li J, Becker B, Zhao Y, Cheng B, Wang S (2021): Gray matter structures associated with neuroticism: A meta-analysis of whole-brain voxel-based morphometry studies. Hum Brain Mapp 42:2706–2721.

Lombardi A, Monaco A, Donvito G, Amoroso N, Bellotti R, Tangaro S (2020): Brain age prediction with morphological features using deep neural networks: Results from predictive analytic competition 2019. Front Psychiatry 11:619629.

Lundberg SM, Lee S-I (2017): A unified approach to interpreting model predictions. In: Guyon, I, Luxburg, UV, Bengio, S, Wallach, H, Fergus, R, Vishwanathan, S, Garnett, R, editors. Advances in neural information processing systems 30. Curran Associates, Inc. pp 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Matsumae M, Kikinis R, Mórocz IA, Lorenzo AV, Sándor T, Albert MS, Black PM, Jolesz FA (1996): Age-related changes in intracranial compartment volumes in normal adults assessed by magnetic resonance imaging. J Neurosurg 84:982–91.

McMakin DL, Kimbler A, Tustison NJ, Pettit JW, Mattfeld AT (2021): Negative overgeneralization is associated with anxiety and mechanisms of pattern completion in peripubertal youth. Soc Cogn Affect Neurosci.

McTabberson T. https://github.com/dreamquark-ai/tabnet.

Miller KL, Alfaro-Almagro F, Bangerter NK, Thomas DL, Yacoub E, Xu J, Bartsch AJ, Jbabdi S, Sotiropoulos SN, Andersson JLR, Griffanti L, Douaud G, Okell TW, Weale P, Dragonu I, Garratt S, Hudson S, Collins R, Jenkinson M, Matthews PM, Smith SM (2016): Multimodal population brain imaging in the UK biobank prospective epidemiological study. Nat Neurosci 19:1523–1536.

Mishra S, Beheshti I, Khanna P (2021): A review of neuroimaging-driven brain age estimation for identification of brain disorders and health conditions. IEEE Rev Biomed Eng PP.

Mugler JP 3rd, Brookeman JR (1990): Three-dimensional magnetization-prepared rapid gradient-echo imaging (3D MP RAGE). Magn Reson Med 15:152–7.

Murphy DG, DeCarli C, Schapiro MB, Rapoport SI, Horwitz B (1992): Age-related differences in volumes of subcortical nuclei, brain matter, and cerebrospinal fluid in healthy men as measured with magnetic resonance imaging. Arch Neurol 49:839–45.

Nguyen DM, Yassa MA, Tustison NJ, Roberts JM, Kulikova A, Nakamura A, Ivleva EI, Van Enkevort E, Brown ES (2019): The relationship between cumulative exogenous corticosteroid exposure and volumes of hippocampal subfields and surrounding structures. J Clin Psychopharmacol 39:653–657.

Nobis L, Manohar SG, Smith SM, Alfaro-Almagro F, Jenkinson M, Mackay CE, Husain M (2019): Hippocampal volume across age: Nomograms derived from over 19,700 people in UK biobank. Neuroimage Clin 23:101904.

Nyberg L, Wåhlin A (2020): The many facets of brain aging. Elife 9.

Nyúl LG, Udupa JK (1999): On standardizing the MR image intensity scale. Magn Reson Med 42:1072–81.

Patenaude B, Smith SM, Kennedy DN, Jenkinson M (2011): A bayesian model of shape and appearance for subcortical brain segmentation. Neuroimage 56:907–22.

Pfefferbaum A, Kwon D, Brumback T, Thompson WK, Cummins K, Tapert SF, Brown SA, Colrain IM, Baker FC, Prouty D, De Bellis MD, Clark DB, Nagel BJ, Chu W, Park SH,

Pohl KM, Sullivan EV (2018): Altered brain developmental trajectories in adolescents after initiating drinking. Am J Psychiatry 175:370–380.

Reagh ZM, Noche JA, Tustison NJ, Delisle D, Murray EA, Yassa MA (2018): Functional imbalance of anterolateral entorhinal cortex and hippocampal dentate/CA3 underlies age-related object pattern separation deficits. Neuron 97:1187–1198.e4.

Reuben A, Brickman AM, Muraskin J, Steffener J, Stern Y (2011): Hippocampal atrophy relates to fluid intelligence decline in the elderly. J Int Neuropsychol Soc 17:56–61.

Rogers BP, Sheffield JM, Luksik AS, Heckers S (2012): Systematic error in hippocampal volume asymmetry measurement is minimal with a manual segmentation protocol. Front Neurosci 6:179.

Sarica A, Vasta R, Novellino F, Vaccaro MG, Cerasa A, Quattrone A, Alzheimer's Disease Neuroimaging Initiative (2018): MRI asymmetry index of hippocampal subfields increases through the continuum from the mild cognitive impairment to the alzheimer's disease. Front Neurosci 12:576.

Saygin ZM, Kliemann D, Iglesias JE, Kouwe AJW van der, Boyd E, Reuter M, Stevens A, Van Leemput K, McKee A, Frosch MP, Fischl B, Augustinack JC, Alzheimer's Disease Neuroimaging Initiative (2017): High-resolution magnetic resonance imaging reveals nuclei of the human amygdala: Manual segmentation to automatic atlas. Neuroimage 155:370–382.

Scahill RI, Frost C, Jenkins R, Whitwell JL, Rossor MN, Fox NC (2003): A longitudinal study of brain volume changes in normal aging using serial registered magnetic resonance imaging. Arch Neurol 60:989–94.

Schlemper J, Oktay O, Schaap M, Heinrich M, Kainz B, Glocker B, Rueckert D (2019): Attention gated networks: Learning to leverage salient regions in medical images. Med Image Anal 53:197–207.

Sethian JA (1996): A fast marching level set method for monotonically advancing fronts. Proc Natl Acad Sci U S A 93:1591–5.

Shwartz-Ziv R, Armon A (2022): Tabular data: Deep learning is not all you need. Information Fusion 81:84–90.

Sinha N, Berg CN, Tustison NJ, Shaw A, Hill D, Yassa MA, Gluck MA (2018): APOE $\varepsilon 4$ status in healthy older african americans is associated with deficits in pattern separation and hippocampal hyperactivation. Neurobiol Aging 69:221–229.

Smith SM, Elliott LT, Alfaro-Almagro F, McCarthy P, Nichols TE, Douaud G, Miller KL (2020): Brain aging comprises many modes of structural and functional change with distinct genetic and biophysical associations. Elife 9.

Thompson DK, Wood SJ, Doyle LW, Warfield SK, Egan GF, Inder TE (2009): MR-determined hippocampal asymmetry in full-term and preterm neonates. Hippocampus 19:118–23.

Tustison NJ ANTsXUKBB GitHub. https://github.com/ntustison/ANTsXUKBB.

Tustison NJ, Altes TA, Qing K, He M, Miller GW, Avants BB, Shim YM, Gee JC, Mugler JP 3rd, Mata JF (2021a): Image- versus histogram-based considerations in semantic segmentation of pulmonary hyperpolarized gas images. Magn Reson Med 86:2822–2836.

Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, Gee JC (2010): N4ITK: Improved N3 bias correction. IEEE Trans Med Imaging 29:1310–20.

Tustison NJ, Cook PA, Holbrook AJ, Johnson HJ, Muschelli J, Devenyi GA, Duda JT, Das SR, Cullen NC, Gillen DL, Yassa MA, Stone JR, Gee JC, Avants BB (2021b): The ANTsX ecosystem for quantitative biological and medical imaging. Sci Rep 11:9068.

Tustison NJ, Cook PA, Klein A, Song G, Das SR, Duda JT, Kandel BM, Strien N van, Stone JR, Gee JC, Avants BB (2014): Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. Neuroimage 99:166–79.

Tustison NJ, Holbrook AJ, Avants BB, Roberts JM, Cook PA, Reagh ZM, Duda JT, Stone JR, Gillen DL, Yassa MA, Alzheimer's Disease Neuroimaging Initiative (2019): Longitudinal mapping of cortical thickness measurements: An Alzheimer's Disease Neuroimaging Initiative-based evaluation study. J Alzheimers Dis.

Tustison NJ, Johnson HJ, Rohlfing T, Klein A, Ghosh SS, Ibanez L, Avants BB (2013): Instrumentation bias in the use and evaluation of scientific software: Recommendations for reproducible practices in the computational sciences. Front Neurosci 7:162.

Tustison NJ, Shrinidhi KL, Wintermark M, Durst CR, Kandel BM, Gee JC, Grossman MC, Avants BB (2015): Optimal symmetric multimodal templates and concatenated random forests for supervised brain tumor segmentation (simplified) with ANTsR. Neuroinformatics 13:209–25.

UKBB. https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=110.

Vieira BH, Pamplona GSP, Fachinello K, Silva AK, Foss MP, Salmon CEG (2022): On the prediction of human intelligence from neuroimaging: A systematic review of methods and reporting. Intelligence 93:101654. https://www.sciencedirect.com/science/article/pii/S0160289622000356.

Wang H, Suh JW, Das SR, Pluta JB, Craige C, Yushkevich PA (2013): Multi-atlas segmentation with joint label fusion. IEEE Trans Pattern Anal Mach Intell 35:611–23.

Weiner MW, Veitch DP, Aisen PS, Beckett LA, Cairns NJ, Green RC, Harvey D, Jack CR, Jagust W, Liu E, Morris JC, Petersen RC, Saykin AJ, Schmidt ME, Shaw L, Siuciak JA, Soares H, Toga AW, Trojanowski JQ, Alzheimer's Disease Neuroimaging Initiative (2012): The alzheimer's disease neuroimaging initiative: A review of papers published since its inception. Alzheimers Dement 8:S1–68.

Weintraub S, Dikmen SS, Heaton RK, Tulsky DS, Zelazo PD, Bauer PJ, Carlozzi NE, Slotkin J, Blitz D, Wallner-Allen K, Fox NA, Beaumont JL, Mungas D, Nowinski CJ, Richler J, Deocampo JA, Anderson JE, Manly JJ, Borosh B, Havlik R, Conway K, Edwards E, Freund L, King JW, Moy C, Witt E, Gershon RC (2013): Cognition assessment using the NIH toolbox. Neurology 80:S54–64.

Wolpert DH, Macready WG (1997): No free lunch theorems for optimization. IEEE Transactions on Evolutionary Computation 1:67–82.

Yoo TS, Metaxas DN (2005): Open science–combining open data and open source software: Medical image analysis with the insight toolkit. Med Image Anal 9:503–6.

Yushkevich PA, Amaral RSC, Augustinack JC, Bender AR, Bernstein JD, Boccardi M, Bocchetta M, Burggren AC, Carr VA, Chakravarty MM, Chételat G, Daugherty AM, Davachi L, Ding S-L, Ekstrom A, Geerlings MI, Hassan A, Huang Y, Iglesias JE, La Joie R, Kerchner GA, LaRocque KF, Libby LA, Malykhin N, Mueller SG, Olsen RK, Palombo DJ, Parekh MB, Pluta JB, Preston AR, Pruessner JC, Ranganath C, Raz N, Schlichting ML, Schoemaker D, Singh S, Stark CEL, Suthana N, Tompary A, Turowski MM, Van Leemput K, Wagner AD, Wang L, Winterburn JL, Wisse LEM, Yassa MA, Zeineh MM, Hippocampal Subfields Group (HSG) (2015): Quantitative comparison of 21 protocols for labeling hippocampal subfields and parahippocampal subregions in in vivo MRI: Towards a harmonized segmentation protocol. Neuroimage 111:526–41.

Yushkevich PA, Wang H, Pluta J, Das SR, Craige C, Avants BB, Weiner MW, Mueller S (2010): Nearly automatic segmentation of hippocampal subfields in in vivo focal T2-weighted MRI. Neuroimage 53:1208–24.

Zhang Y, Brady M, Smith S (2001): Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. IEEE Trans Med Imaging 20:45–57.