2017

Towards Precision Psychiatry: gray Matter Development And Cognition In Adolescence

Efstathios Dimitrios Gennatas
University of Pennsylvania, gennatas@gmail.com

Towards Precision Psychiatry: gray Matter Development And Cognition In Adolescence

Abstract
Precision Psychiatry promises a new era of optimized psychiatric diagnosis and treatment through comprehensive, data-driven patient stratification. Among the core requirements towards that goal are: 1) neurobiology-guided preprocessing and analysis of brain imaging data for noninvasive characterization of brain structure and function, and 2) integration of imaging, genomic, cognitive, and clinical data in accurate and interpretable predictive models for diagnosis, and treatment choice and monitoring. In this thesis, we shall touch on specific aspects that fit under these two broad points. First, we investigate normal gray matter development around adolescence, a critical period for the development of psychopathology. For years, the common narrative in human developmental neuroimaging has been that gray matter declines in adolescence. We demonstrate that different MRI-derived gray matter measures exhibit distinct age and sex effects and should not be considered equivalent, as has often been done in the past, but complementary. We show for the first time that gray matter density increases from childhood to young adulthood, in contrast with gray matter volume and cortical thickness, and that females, who are known to have lower gray matter volume than males, have higher density throughout the brain. A custom preprocessing pipeline and a novel high-resolution gray matter parcellation were created to analyze brain scans of 1189 youths collected as part of the Philadelphia Neurodevelopmental Cohort. This work emphasizes the need for future studies combining quantitative histology and neuroimaging to fully understand the biological basis of MRI contrasts and their derived measures. Second, we use the same gray matter measures to assess how well they can predict cognitive performance. We train mass-univariate and multivariate models to show that gray matter volume and density are complementary in their ability to predict performance. We suggest that parcellation resolution plays a big role in prediction accuracy and that it should be tuned separately for each modality for a fair comparison among modalities and for an optimal prediction when combining all modalities. Lastly, we introduce rtemis, an R package for machine learning and visualization, aimed at making advanced data analytics more accessible. Adoption of accurate and interpretable machine learning methods in basic research and medical practice will help advance biomedical science and make precision medicine a reality.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Neuroscience

First Advisor
Ruben C. Gur

Second Advisor
Geoffrey K. Aguirre

Keywords
Adolescence, Development, Gray matter, Machine Learning, MRI, Neuroimaging

Subject Categories
Neuroscience and Neurobiology

TOWARDS PRECISION PSYCHIATRY:
GRAY MATTER DEVELOPMENT AND COGNITION IN ADOLESCENCE

Efstathios D. Gennatas, MBBS AICSM

A DISSERTATION

in

Neuroscience

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2017

Supervisor of Dissertation

_____

Ruben C. Gur, PhD
Professor of Psychiatry

Graduate Group Chairperson

_____

Joshua I. Gold, PhD
Professor of Neuroscience

Dissertation Committee
Geoffrey K. Aguirre, MD PhD, Associate Professor of Neurology
Brian B. Avants, PhD, Associate Director, Global Biomarker Discovery and Development, Biogen
Lyle H. Ungar, PhD, Professor of Computer and Information Science
Daniel H. Wolf, MD PhD, Assistant Professor of Psychiatry

TOWARDS PRECISION PSYCHIATRY: GRAY MATTER DEVELOPMENT AND COGNITION IN ADOLESCENCE

To my parents, Eleni and Costas, my brother, Spyros,
and my mentor, Bill Seeley.
Philadelphia, May 2017

ABSTRACT

TOWARDS PRECISION PSYCHIATRY:
GRAY MATTER DEVELOPMENT AND COGNITION IN ADOLESCENCE

Efstathios D. Gennatas

Ruben C. Gur

Precision Psychiatry promises a new era of optimized psychiatric diagnosis and treatment through comprehensive, data-driven patient stratification. Among the core requirements towards that goal are: 1) neurobiology-guided preprocessing and analysis of brain imaging data for noninvasive characterization of brain structure and function, and 2) integration of imaging, genomic, cognitive, and clinical data in accurate and interpretable predictive models for diagnosis, and treatment choice and monitoring. In this thesis, we shall touch on specific aspects that fit under these two broad points. First, we investigate normal gray matter development around adolescence, a critical period for the development of psychopathology. For years, the common narrative in human developmental neuroimaging has been that gray matter declines in adolescence. We demonstrate that different MRI-derived gray matter measures exhibit distinct age and sex effects and should not be considered equivalent, as has often been done in the past, but complementary. We show for the first time that gray matter density increases from childhood to young adulthood, in contrast with gray matter volume and cortical thickness, and that females, who are known to have lower gray

matter volume than males, have higher density throughout the brain. A custom preprocessing pipeline and a novel high-resolution gray matter parcellation were created to analyze brain scans of 1189 youths collected as part of the Philadelphia Neurodevelopmental Cohort. This work emphasizes the need for future studies combining quantitative histology and neuroimaging to fully understand the biological basis of MRI contrasts and their derived measures. Second, we use the same gray matter measures to assess how well they can predict cognitive performance. We train mass-univariate and multivariate models to show that gray matter volume and density are complementary in their ability to predict performance. We suggest that parcellation resolution plays a big role in prediction accuracy and that it should be tuned separately for each modality for a fair comparison among modalities and for an optimal prediction when combining all modalities. Lastly, we introduce *rtemis*, an R package for machine learning and visualization, aimed at making advanced data analytics more accessible. Adoption of accurate and interpretable machine learning methods in basic research and medical practice will help advance biomedical science and make precision medicine a reality.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

1. INTRODUCTION

The assessment and management of psychiatric disorders have always been greatly challenging. Psychiatric research and clinical management have come a long way over the past century, yet diagnosis still suffers from low accuracy rates and current treatment efforts enjoy limited success, both in terms of numbers effectively treated and the extent of their improvement. Research and clinical assessment tools, including neuroimaging, genomic sequencing, and clinical and cognitive testing, are helping accumulate large datasets on healthy subjects and patients with psychiatric symptoms. Advanced data analysis methods are becoming increasingly available and promise to deliver important insights to fill in the gaps in our understanding of psychopathology and suggest better ways to manage it.

Precision medicine refers to clinical decision-making tailored to the individual. The term has emerged in recent years to describe the goal of capturing and addressing individual idiosyncrasy in order to optimize clinical decision making and outcomes by capitalizing on a) the increasing amounts of available clinical data b) increasing computational power, and c) advanced data analytic methods. Consider the stark contrast between the common approach in biomedical research versus the reality of clinical practice. The former has in large part focused on comparing groups of patients vs. healthy controls to test hypotheses, while the latter has always focused on assessment of the individual. At the same time, the available classification

and diagnostic manuals, DSM 5 and ICD 10, suggest clinicians fit patients into groups using a discrete set of labels. It is becoming increasingly clear that these categories correspond very poorly with underlying brain pathology (Hyman, 2007; Insel and Cuthbert, 2015). Perhaps more than clinical practice, this has affected psychiatric research, as researchers end up studying inhomogeneous groups of subjects based on their DSM, or similar, labels and often deriving divergent results. To address this, the National Institute of Mental Health (NIMH) introduced in 2010 the Research Domain Criteria (RDoC) project to shift focus from symptoms to underlying neuropathology, thus providing a framework for enhanced patient stratification that would better support ongoing psychiatric research and would help shape future brain-based clinical classification schemes (Insel et al., 2010). Importantly, the RDoC project, an ongoing experiment (https://www.nimh.nih.gov/about/director/messages/2017/the-future-of-rdoc.shtml), stresses the dimensional aspect of behavior and psychopathology and the need for a robust, data-driven discovery process. Symptoms are not either present or absent and do not come in discrete sets, but can be present at a variable extent in overlapping combinations.

Magnetic resonance neuroimaging affords researchers and clinicians the ability to study the human brain in vivo in a safe and noninvasive way. MRI scanners can be programmed to create different contrasts to focus selectively on different brain tissue or processes, e.g. gray or white matter,

water diffusion, blood oxygenation level, etc. Trade-offs between temporal and spatial resolution, among other parameters, can be exploited to create sequences that create one or more high-resolution images or a whole timeseries. A T1-weighted image can differentiate protons based on their immediate environment and provides a high-resolution structural image of the brain with good tissue contrast. A diffusion-weighted image (DWI) measures water diffusion which can be used to estimate direction of myelinated white matter tracts. Blood-oxygen-level-dependent (BOLD) signal can be acquired in rapid succession to study changes in blood flow over time across the brain. Human subjects can be imaged at any age, from infancy to advanced age, even *in utero*, to build extensive cross-sectional and longitudinal datasets on healthy subjects and patients. As such, MR neuroimaging is one of the core tools for the study of human subjects in neurologic and psychiatric research.

Data analysis, in general, can be divided into two main steps: data preprocessing, and statistical analysis / modeling. Preprocessing, which includes data inspection, cleaning, and transformation forms the bulk of the work and commands most of the attention both because it consists of multiple steps, each involving multiple parameters, and because the success of any subsequent modeling is directly dependent on it. It is also not unique: different analysis methods may benefit from, or require, different preparation of the same data. A weak modeling approach on solid data will generally yield

far more meaningful and useable results than the most powerful algorithm trained on bad data. “There is no substitute for good data” (Luck, 2014).

Each type of MRI requires its own set of preprocessing steps before any statistical analysis and modeling can be applied. A main challenge across MRI modalities remains the scarcity of validation data. Little is known about the direct relationship between MRI contrasts and the underlying neurobiology, which makes tuning of preprocessing parameters particularly tricky. As no gold standards exist in preprocessing, variability in methods remain a source of uncertainty and heterogeneity across studies and their results.

Following preprocessing, approaches for hypothesis testing and modeling are drawn from all of statistics and machine learning. Formal statistical methods come with specific assumptions that must be met if they are to be employed, while other methods can be applied more universally. For example, the Generalized Linear Model has been the *de facto* standard for neuroimaging data analysis, and while it remains a valid choice for many applications, a lot of datasets it is commonly applied on violate some of its core assumptions, commonly the assumptions of normality and linearity. Even after careful consideration of modeling assumptions, it is not possible to accurately predict which combination of methods will yield best results. This can lead to repeated attempts at data preprocessing and modeling until a specific preconceived relationship is found or any significant result is obtained, leading to high bias and reduced validity / reproducibility of

published results. The need is evident for informed preprocessing and modeling of biomedical data.

In this thesis, we shall focus on specific aspects of data preprocessing and analysis of neuroimaging data that we believe form part of core considerations for neuropsychiatric imaging research. Specifically, we shall explore structural brain development around adolescence and its relation to cognition. The first chapter has four broad goals:

- Recommend a pipeline for T1-weighted MRI preprocessing and estimation of gray matter measures: gray matter density, volume, mass, and cortical thickness
- Propose a method for high-resolution gray matter parcellation
- Characterize age-related and sex effects on different gray matter measures from childhood to young adulthood and clear longstanding confusion by showing they are unique and complementary, not equivalent
- Provide an overview of factors known to affect structural MRI signal and emphasize the need for combined histology and MRI studies to fully understand the biological basis of MRI contrasts and derived measures

In the second chapter, we explore how well these gray matter measures predict cognitive performance in different age groups:

- We hypothesize that structural-functional coupling grows stronger with age.

- Based on this, we predict that prediction accuracy will be highest in the oldest subjects.

Finally, in the third chapter we present ***rtemis***, an R package for machine learning and visualization, which was developed to support the above work and is aimed at making advanced data analytics more accessible to biomedical and other researchers.

2. AGE-RELATED EFFECTS AND SEX DIFFERENCES IN GRAY MATTER DENSITY, VOLUME, MASS, AND CORTICAL THICKNESS FROM CHILDHOOD TO YOUNG ADULTHOOD

INTRODUCTION

Structural neuroimaging provides insights into the spectrum of typical and non-typical brain and neurocognitive development. T1-weighted imaging is the most commonly acquired MRI sequence and offers high-resolution, low-noise images of brain structure with good tissue contrast. Several structural measures can be derived from a single T1-weighted image, including gray matter density (GMD), gray matter volume (GMV), and cortical thickness (CT). Since the early days of MRI, a large body of research has utilized these measures to study healthy and clinical populations. Perhaps surprisingly, confusion exists in the field as GMD, GMV, and CT are often wrongly assumed to be equivalent or highly related measures of regional gray matter quantity. GMV and CT are measured in mm^3 and mm, respectively. GMD, on the other hand, is a unitless, scalar measure derived from image segmentation and related to T1 signal intensity. In one form or other, gray matter abnormalities have been described in all major neurologic and psychiatric diseases. Voxel Based Morphometry (VBM) analyses have suggested syndrome-specific regional atrophy patterns in neurodegenerative diseases (Seeley et al., 2009). Gray matter abnormalities are widely reported in psychiatric disorders as well but paint a more complex picture (Brent et al., 2013; Bakhshi and Chance,

2015), likely reflecting both increased neuropathological heterogeneity and diagnostic variability. Much of psychopathology emerges around adolescence, a period characterized by rapid changes in behavior. Detecting and interpreting what may often be subtle and diffuse disease-related differences on top of profound and variable age-related changes is particularly challenging. A clear, multidimensional understanding of normative structural brain development is therefore essential.

The first two years of life are characterized by rapid gray matter growth, which reaches its lifetime maximum at around age 2-3 (Matsuzawa et al., 2001; Knickmeyer et al., 2008). In contrast, myelination of white matter tracts continues well into adulthood, until the late 30s (Grydeland et al., 2013). Several developmental neuroimaging studies have described modest decreases in gray matter during adolescence using measures derived from gray matter volume and cortical thickness (Sowell et al., 2003; Gogtay et al., 2004; Sowell et al., 2004; Shaw et al., 2008; Brain Development Cooperative Group, 2012). It should be noted that some of the early studies used the term "gray matter density" to refer to the proportion of gray matter voxels around a sphere of fixed diameter following hard segmentation of the brain (Sowell et al., 2003; Gogtay et al., 2004; Sowell et al., 2004), and suggested this quantity reflected local cortical thickness. Today, cortical thickness can be measured directly using automated methods and GMD usually refers to a different measure, specifically the output of soft segmentation. Unlike hard

segmentation, where each voxel is labeled as “gray”, “white”, or “CSF” (for the common 3-class case), soft segmentation creates a GMD map by assigning voxels a value between zero and one, which is considered to reflect the amount of gray matter in each voxel. It is related to the T1 signal and thus to the regional proton density as well as the tissue microenvironment. To complicate things further, one of the most common measures employed in the literature, and the default option in many VBM pipelines, is “modulated” gray matter density. This is equal to GMD multiplied by a scaling factor to account for volume change from the individual’s native space image to the registration template. It adds to the confusion because the relative contribution of each measure is unclear and likely variable spatially and temporally (with regards to age). To date, no study has compared age-related effects on these four commonly used gray matter measures.

In this study, we used the extensive cross-sectional neuroimaging dataset collected on the Philadelphia Neurodevelopmental Cohort (PNC) to characterize age effects and sex differences on native space gray matter density, volume, and mass (defined as density times volume; equivalent to modulated gray matter density), as well as cortical thickness.

MATERIALS AND METHODS

Subjects and MRI acquisition

All data was collected as part of the Philadelphia Neurodevelopmental Cohort as previously described (Satterthwaite et al., 2014). Procedures were approved by the Institutional Review Boards of the Children's Hospital of Philadelphia and of the University of Pennsylvania. 1189 subjects (648 females) aged 8 to 23 years were selected from a starting total of 1601 after excluding those with neurological or psychiatric history, use of psychoactive medication or incidental findings and those whose structural data failed quality control. Scanning of all subjects was performed on the same Siemens TIM Trio scanner (Erlangen, Germany) at the Hospital of the University of Pennsylvania. T1-weighted imaging was obtained using a magnetization prepared, rapid-acquisition gradient-echo (MPRAGE) sequence (TR = 1810, TE = 3.5, TI = 1100; FOV = 180 RL / 240 AP).

MRI preprocessing

A custom T1 preprocessing pipeline was created using ANTs (https://github.com/stnava/ANTs; RRID: SCR_004757). Raw T1 volumes were first corrected for bias due to field inhomogeneity using the N4 algorithm (Tustison et al., 2010). The bias-corrected volumes were then registered to a whole-head MNI template (whole-head-to-whole-head registration). The inverse transformation was used to map the MNI brain mask to native space,

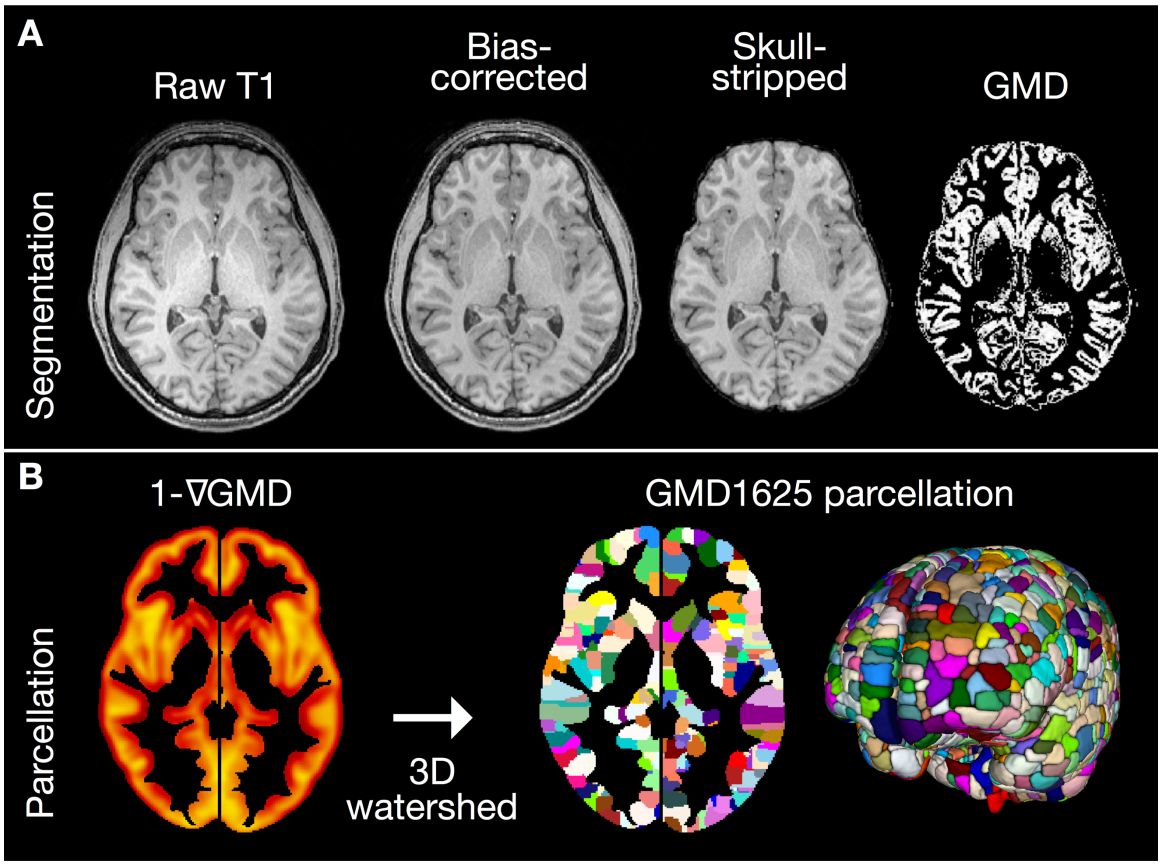

Figure 2.1 T1 preprocessing and high-resolution gray matter parcellation. **A**, Raw T1 MPRAGE volumes were first corrected for field inhomogeneity and then skull stripped by transforming the MNI brain mask to native space. Gray matter segmentation was performed without the use of tissue priors to produce unbiased estimates of GMD. **B**, The GMD maps of an age- and sex-balanced subsample of 240 subjects were averaged and smoothed; 1 minus the gradient of the resulting image was calculated and passed to a 3D watershed algorithm, resulting in 1625 regions covering the whole-brain gray matter.

which was used to isolate the brain in native space (skull-stripping). The skull-stripped volume was then registered to the skull-stripped MNI template (brain-to-brain registration), which results in improved registration accuracy compared to the whole-head-to-whole-head registration (Klein et

al., 2010). Registrations were performed by a sequence of rigid, affine and symmetric diffeomorphic (SyN) transformations (Avants et al., 2008; Klein et al., 2009).

Gray matter density and cortical thickness estimation

MRI brain tissue segmentation is commonly guided by a set of tissue priors. Given the wide age range of our sample, we wanted to avoid using a single set of priors for all subjects or different sets of priors for different age bins. We therefore implemented an iterative process based on Atropos (Avants et al., 2011) that requires no tissue priors. On the first iteration, K-means initialization was used to derive 3 classes. The following two iterations used the segmentation output of the previous step for initialization. This procedure resulted in a 3-class hard segmentation and a GMD map (soft segmentation) for each subject in native space (**Figure 2.1A**). Cortical thickness was obtained using ANTs' diffeomorphic registration-based cortical thickness (DiReCT) estimation procedure (Das et al., 2009) as implemented in the ANTsCT pipeline, following registration of all T1 images to a study-specific template. This method offers reliable CT estimation (Tustison et al., 2014) and, by providing a voxelwise measure in native volumetric space, allows the use of the same brain parcellation as the other modalities.

Quality assurance

To assess the quality of the T1 acquisition and segmentation, we calculated pairwise spatial correlations among all subjects for two sets of images: bias-field corrected, normalized T1s and normalized GMD maps. All images whose spatial correlation was more than two standard deviations lower than the mean in either case were excluded (n = 56). Visual check confirmed variable extent of motion artifact in the excluded images, with those near the threshold being only minimally affected (still excluded). Motion artifact is known to significantly affect tissue segmentation and all our derived measures (Blumenthal et al., 2002; Savalia et al., 2016) and our large sample afforded us this perhaps conservative exclusion threshold.

High-resolution gray matter parcellation

Multiple methods for whole brain parcellation have been previously proposed. Anatomical parcellations, like the AAL atlas (Tzourio-Mazoyer et al., 2002), and the Harvard-Oxford Atlas (distributed with FSL; https://fsl.fmrib.ox.ac.uk/; RRID:SCR_002823) are based on neuroanatomy but consist of a small number of relatively large regions. Using large parcels or regions of interest (ROIs) runs the risk of averaging over inhomogeneous regions, resulting in signal loss. On the other hand, a number of approaches have been proposed for parcellation based on functional connectivity derived from task-free functional MRI data (also known as resting state fMRI), but none based on T1-weighted images. A recent approach used multimodal MRI

data to create a parcellation of 180 regions in each hemisphere. As the authors note, the parcels show high variance in shape and size and consider their number to be "a lower bound, as some parcels are probably complexes of multiple areas" (Glasser et al., 2016).

Our goal was to develop a high-resolution parcellation derived from structural data where parcels are centered around GMD peaks, i.e. cortical gyri and subcortical nuclei. An age- and sex-matched subsample of the 1189 subjects was created by first splitting the initial sample by sex, then splitting each set into deciles based on its age range, and finally randomly selecting 12 subjects from each resulting subset (i.e. 12 subjects per sex per age decile), giving a total of 240 subjects. A mean image was created from the normalized, smoothed GMD maps of these subjects. In order to identify GMD peaks, the gradient of the mean GMD image was calculated, subtracted from 1, and smoothed. A 3D watershed algorithm was applied on the resulting image, producing 1625 parcels covering the whole brain gray matter (**Figure 2.1B**).

Native space parcelwise data extraction

The PNC-GMD1625 parcellation was transformed to each subject's native space by applying the inverse of the brain-to-brain transformation (i.e. MNI-to-native space) and masked by each subject's gray matter hard segmentation. Volume and mean GMD and CT values were estimated for each

parcel for each subject using the c3d utility (part of ITK-SNAP; http://www.itksnap.org/; RRID: SCR_002010). CT values were measured for 1339 of the 1625 regions, after excluding subcortical regions. In order to get a native space equivalent of modulated density, we derived gray matter mass (GMM) as the product of GMD and GMV. Native space analysis allows direct measurement of GMV and extraction of mean GMD and CT values with no interpolation. Averaging GMD and CT values within each parcel instead of applying Gaussian smoothing avoids smoothing-related artifacts which are exaggerated in a segmented image. Cortical thickness (and therefore the gray matter segmentation) varies around 2-5mm while smoothing kernels are commonly at least 8mm full width at half maximum (FWHM). In a gray matter segmentation, this results in voxel intensities being averaged with surrounding empty voxels (i.e. voxels of zero intensity), causing a drop in signal. The extent of signal drop depends on the number of surrounding empty voxels, which varies both by brain region and age. This makes intensity values from different locations incomparable and directly confounds age-related effects. Interpolation results in a similar artifact, equivalent to smoothing at the single voxel level.

Age-related effects and sex differences

Generalized additive models (GAMs) were used to characterize age-related effects and sex differences on GMD, GMV, GMM, and CT using the mgcv package (Wood, 2011; Wood, 2012) in R (R Project for Statistical Computing; https://www.r-project.org/; RRID:SCR_001905). A GAM is similar to a generalized linear model where predictors can be replaced by smooth functions of themselves, offering efficient and flexible estimation of non-linear effects. Three sets of models were fit. Full models included age and age-by-sex interaction terms represented using penalized smoothing splines with smoothing parameters selected by restricted maximum likelihood. For each modality in turn, for each gray matter parcel *p*, a model of form (1) was fit:

$$\{GMD, GMV, GMM, CT\}_p \sim Sex + s(Age) + s(Age * Sex) \quad (1)$$

where s() represents a penalized smoothing spline. The dimension of the basis used to represent the smooth terms was limited to a maximum of 5 in all models. Reduced models were fit in order to obtain accurate p-values for the main effects of sex and age. Specifically, model (2) omits the interaction term and was fit for each parcel in order to obtain p-values for the main effect of sex. Model (3) omits sex entirely and was therefore fit separately for each sex *s* in order to obtain p-values for the main effect of age.

$$\{GMD, GMV, GMM, CT\}_p \sim Sex + s(Age) \quad (2)$$

	GMD			GMV			GMM			CT			N
MNI label	**Mean**	**SD**	**CV**	**Mean**	**SD**	**CV**	**Mean**	**SD**	**CV**	**Mean**	**SD**	**CV**	**total**
Frontal	0.82	0.03	3.96	359.75	99.73	27.72	295.18	83.19	28.18	3.57	0.53	14.82	514
Temporal	0.82	0.03	3.77	352.64	94.07	26.68	290.09	79.32	27.34	4.03	0.60	14.86	319
Parietal	0.82	0.03	3.79	385.09	114.59	29.76	317.25	95.28	30.03	3.34	0.53	15.88	294
Occipital	0.80	0.03	3.54	290.55	95.01	32.70	235.55	78.16	33.18	3.27	0.62	18.86	190
Insula	0.83	0.03	3.76	578.78	106.87	18.47	478.39	88.37	18.47	4.53	0.61	13.49	22
Caudate	0.84	0.03	3.20	374.18	61.72	16.50	313.19	52.72	16.83	-	-	-	25
Putamen	0.76	0.03	4.12	317.26	92.32	29.10	242.50	74.38	30.67	-	-	-	23
Thalamus	0.77	0.03	3.38	316.30	80.05	25.31	246.39	64.68	26.25	-	-	-	25
Cerebellum	0.80	0.03	3.74	340.28	94.78	27.85	273.04	78.83	28.87	-	-	-	213

Table 2.1 Summary statistics of regional gray matter measures averaged by MNI label. **SD**, standard deviation; **CV**, coefficient of variation (= SD/mean * 100); **N total**, total number of parcels within MNI label (of 1625).

$$\{GMD, GMV, GMM, CT\}_{s,p} \sim s(Age) \quad (3)$$

Models of form (3) were also fitted at the whole brain level, using mean GMD and CT (weighted by number of voxels in each parcel), and total GMV and GMM, as separate dependent variables.

$$\{MeanGMD, MeanCT, TotalGMV, TotalGMM\}_{s,p} \sim s(Age) \quad (4)$$

In each case, p-values were corrected for multiple comparisons by controlling the false discovery rate (FDR; Benjamini and Hochberg method; q-value = 0.05).

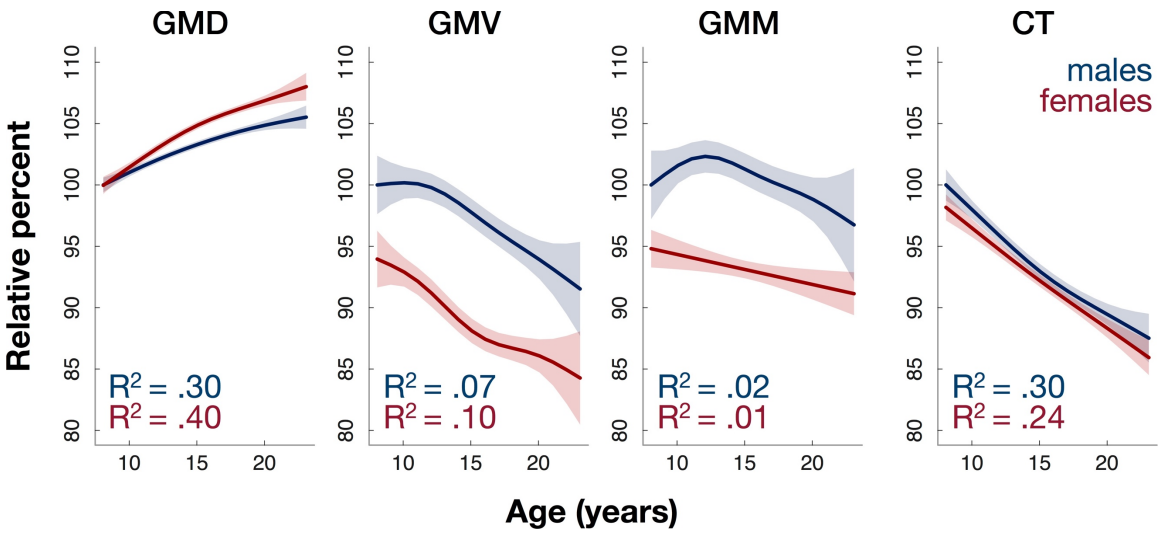

Figure 2.2 Density increases in adolescence while other measures largely decrease. Females have higher density and lower volume. Plots show fitted values of whole-brain gray matter measures against age for the two sexes. GMD and CT were averaged across the brain (weighted by *N* voxels in each parcel), and GMV and GMM were summed. To make results comparable across measures, they are plotted as percentages: 100% is defined as the fitted value for males at 8 years of age. Shaded bands correspond to 2 SE of the fit (~95% confidence interval)

RESULTS

Whole-brain age-related effects: Gray matter density increases while volume and thickness decrease

We sought to characterize age-related effects and sex differences at the whole brain and regional level on three independent gray matter measures, GMD, GMV, and CT, and a derived measure, GMM = GMD * GMV. At the whole brain level, we find that total brain GMV and CT decrease from childhood to young adulthood (8 to 23 years) in accordance with previous studies. In contrast, mean brain GMD increases during the same period. Whole brain GMM shows only a modest decrease. **Figure 2.2** shows plots of fitted values converted to

relative percentages (fitted values for 8 y.o. males are defined as 100%) derived from whole brain models (model of form 4 under Materials and Methods, Age-related effects and sex differences). Importantly, GMD is most sensitive to age: 30% and 40% of variance of mean brain GMD is explained by age at scan time for males and females respectively. CT follows with respective values of 30% and 24%, while only 7% and 10% of variance of total brain GMV is explained by age for males and females respectively.

Whole-brain sex differences: Females have lower volume, higher density than males

Females were found to have lower total GMV than males, as expected by known sex differences in average head and brain size. At the same time, however, we show that females have higher mean GMD than males. Total CT was not significantly different between the two sexes in our analysis (**Figure 2.2**).

Regional variability in age-related and sex effects

To achieve regional specificity, we created a high-resolution parcellation covering the whole brain gray matter and consisting of 1625 regions. To summarize the large number of regional results, each of the 1625 gray matter parcels was assigned one of nine MNI labels (Frontal Lobe, Temporal Lobe, Parietal Lobe, Occipital Lobe, Insula, Caudate, Putamen, Thalamus, and Cerebellum), as defined by the MNI atlas in FSL. **Table 2.1** presents summary statistics for each measure aggregated by MNI label. GMM and GMV had the

MNI label	Main Effect of Age								Main Effect of Sex				Age x Sex Interaction				N total
	GMD		GMV		GMM		CT		GMD	GMV	GMM	CT	GMD	GMV	GMM	CT	
	M	F	M	F	M	F	M	F									
Frontal	99	100	44	67	35	27	92	91	100	88	81	24	79	4	7	1	514
Temporal	100	100	43	65	35	34	75	78	99	95	91	56	94	2	2	1	319
Parietal	98	100	64	74	46	38	96	93	99	84	80	52	91	4	6	4	294
Occipital	99	100	69	47	49	18	92	84	96	74	78	44	92	8	16	15	190
Insula	100	100	14	59	32	32	100	100	100	100	100	14	45	0	0	0	22
Caudate	96	100	24	92	20	32	-	-	100	96	96	-	44	4	4	-	25
Putamen	83	91	100	100	100	96	-	-	70	91	87	-	61	9	13	-	23
Thalamus	100	100	64	76	48	64	-	-	88	96	96	-	32	8	8	-	25
Cerebellum	100	100	51	54	48	46	-	-	79	87	85	-	66	2	4	-	213

Table 2.2 Generalized Additive Models: Main effects and interaction by MNI label: Percentage of parcels with significant effects after FDR correction. **M**, Male; **F**, Female

			Threshold	Median
Main effect of age	GMD	M	0.0494	2.005E-24
		F	0.0500	3.096E-47
	GMV	M	0.0276	2.359E-04
		F	0.0337	2.516E-04
	GMM	M	0.0227	2.612E-04
		F	0.0197	2.516E-04
	CT	M	0.0448	2.195E-09
		F	0.0442	2.989E-10
Main effect of sex	GMD		0.0480	1.437E-12
	GMV		0.0446	1.448E-09
	GMM		0.0427	9.395E-08
	CT		0.0231	5.327E-04
Age x Sex Interaction	GMD		0.0414	6.223E-04
	GMV		0.0029	3.427E-04
	GMM		0.0038	6.872E-04
	CT		0.0032	6.512E-05

Table 2.3 Generalized additive models: FDR threshold and median p values. **M**, Male; **F**, Female; **Threshold**, unadjusted p value corresponding to FDR q value of 0.05; **Median**, median of unadjusted p values surviving FDR correction.

highest coefficient of variation (mean CV = 26.7 and 26.0, respectively), followed by CT (mean CV = 15.6). GMD showed the lowest CV (mean CV = 3.7). Parcel-wise GAMs were fitted to investigate the regional variability of age-related and sex effects in our sample (1625 parcels for GMD, GMV,

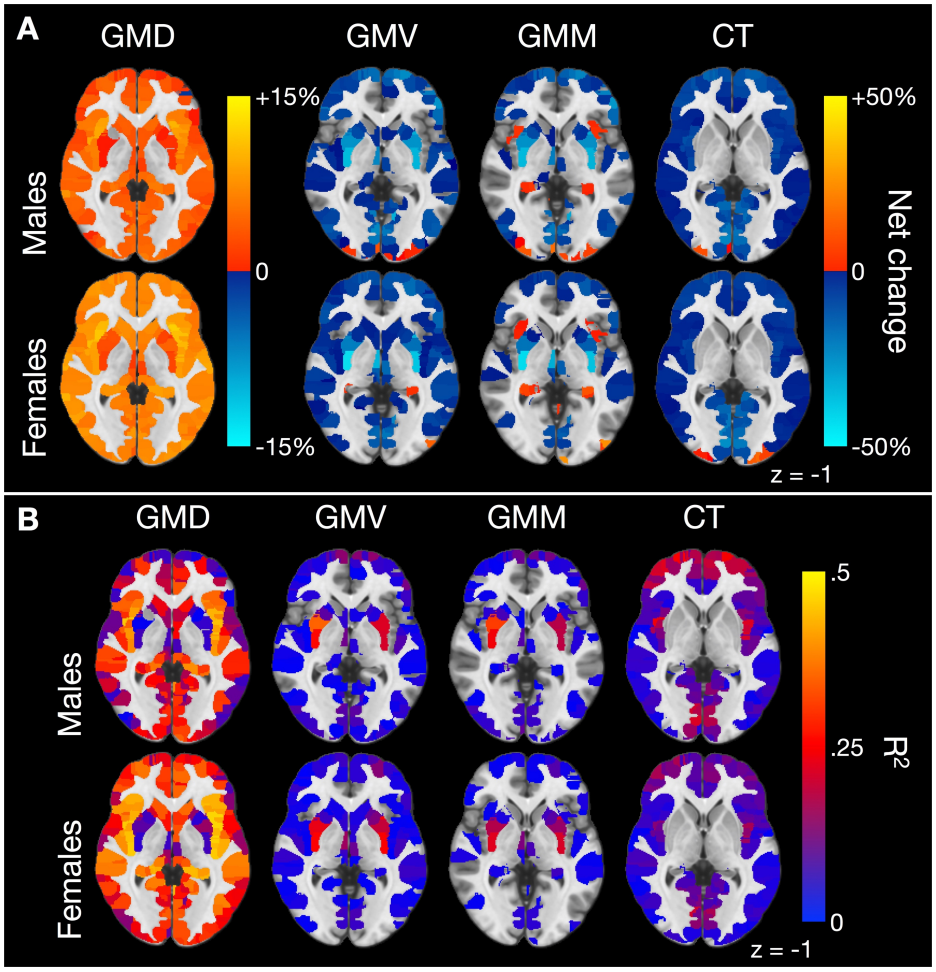

Figure 2.3 Percentage net change and variance explained by sex and modality. **A**, For each parcel, the percentage net change was calculated as follows: (fitted value at 23 fitted value at 8)/(fitted value at 8) 100%. GMD increased virtually throughout the brain, while the other modalities show mostly decreases. Females showed a greater increase in density than males throughout the brain. **B**, Percentage variance of each measure explained by age. GMD showed the highest R^2 values, followed by CT. High bilateral symmetry on all maps suggests biological plausibility. Interactive movies including all axial slices in this figure are available on-line at https://egenn.github.io/gmdvdev

and GMM; 1339 for CT). **Table 2.2** shows the percentage of parcels with significant main effect of age, main effect of sex, and age-by-sex interaction after FDR correction (q = 0.05), aggregated by MNI label. GMD showed

significant age effects throughout the brain (99% of parcels in males, 99.9% in females), followed by CT (89% males, 88% females). GMV, on the other hand, showed significant age effects only in 52% and 65% of parcels in males and females respectively, while the numbers for GMM were 41% and 33%. Sex effects were strong for GMD, GMV, and GMM. Indeed, main effect of sex was more widespread than main effect of age in GMV and GMM. Sex effects in CT were present in a minority of regions across the whole brain, but in just over half of all temporal and parietal parcels. Age-by-sex interactions were virtually limited to GMD. **Table 2.3** shows the unadjusted p-value corresponding to FDR q-value of 0.05 and the median of unadjusted p-values surviving FDR correction.

To study the direction of age-related effects in each parcel, the net change from youngest to oldest was estimated by subtracting the fitted value at 8 years from the fitted value at 23 years for each modality, sex, and region and converted to a percentage (by dividing with the fitted values at 8 years). Net change for parcels not surviving FDR correction was set to zero. GMD increased, on average, within all MNI labels, while GMV and CT decreased. Mean GMM decreased in all MNI regions other than the temporal lobe, insula, and cerebellum. The bilateral insula stands out showing the highest increase in GMD and GMM of all MNI regions. To characterize each parcel's sensitivity to age, we examined each model's adjusted R^2, denoting percent variance of each modality's regional values explained by age. **Table 2.4** lists R^2 and

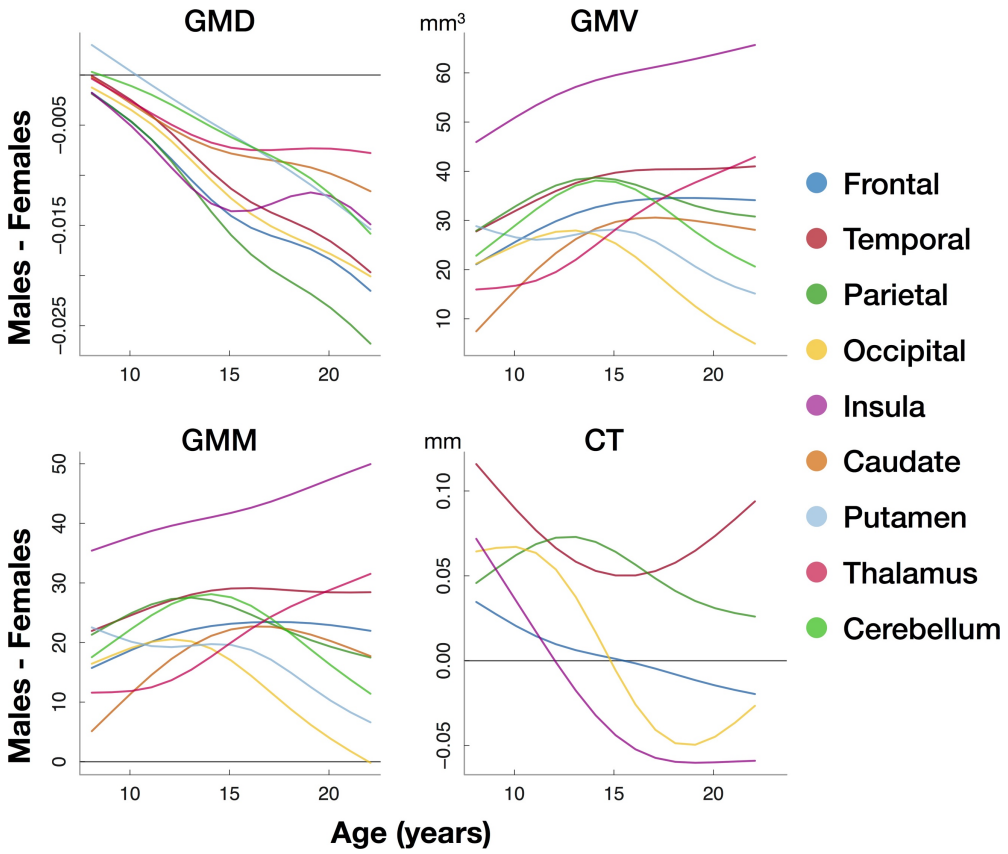

Figure 2.4 Sex differences by modality by MNI label against age. The difference of male and female fitted values for each modality for each MNI label was calculated at each year from 8 to 23 years of age. This plot highlights qualitatively how sex differences vary with age, in most cases in a nonlinear fashion (a constant sex difference in any measure would appear as a horizontal line). Note that only in CT the direction of the difference changes in frontal and occipital lobes as well as the bilateral insula from a male to a female advantage.

percent change for each modality. Averaging by MNI label masks the variability within each label, for example, a lobe may on average decrease in volume, but some parcels within it may increase. For this reason, **Table 2.4** includes numbers of individual parcels with a net positive and net negative change from 8 to 23 years within each MNI label. Brain slices mapping net

		GMD				GMV				GMM				CT				N
	MNI label	R^2	Pct	N ↑	N ↓	R^2	Pct	N ↑	N ↓	R^2	Pct	N ↑	N ↓	R^2	Pct	N ↑	N ↓	total
Males	Frontal	0.181	5.48	507	1	0.015	-5.34	35	193	0.012	-1.37	66	116	0.088	-12.75	23	450	514
	Temporal	0.207	5.79	319	0	0.013	-2.69	45	92	0.015	1.70	69	44	0.027	-5.43	30	209	319
	Parietal	0.168	4.99	288	0	0.021	-10.17	12	176	0.015	-5.22	24	112	0.097	-15.87	4	278	294
	Occipital	0.170	5.09	189	0	0.026	-11.02	13	118	0.018	-6.33	15	78	0.063	-14.11	5	169	190
	Insula	0.393	8.51	22	0	0.002	-1.04	0	3	0.008	3.63	7	0	0.140	-17.49	0	22	22
	Caudate	0.202	4.78	24	0	0.009	-3.15	0	6	0.006	-0.27	3	2		-	-	-	25
	Putamen	0.058	2.60	18	1	0.122	-28.70	0	23	0.097	-26.76	0	23		-	-	-	23
	Thalamus	0.222	5.59	25	0	0.034	-10.60	2	14	0.028	-5.92	3	9		-	-	-	25
	Cerebellum	0.216	5.78	213	0	0.028	-3.02	32	76	0.025	0.91	50	53		-	-	-	213

		GMD				GMV				GMM				CT				N
	MNI label	R^2	Pct	N ↑	N ↓	R^2	Pct	N ↑	N ↓	R^2	Pct	N ↑	N ↓	R^2	Pct	N ↑	N ↓	total
Females	Frontal	0.263	8.21	514	0	0.017	-8.68	22	321	0.007	-0.84	40	100	0.075	-11.15	25	444	514
	Temporal	0.296	8.51	319	0	0.016	-6.48	40	167	0.013	1.89	60	47	0.026	-5.27	36	214	319
	Parietal	0.293	8.41	294	0	0.022	-12.16	6	213	0.009	-3.36	20	93	0.092	-14.86	8	266	294
	Occipital	0.262	7.66	190	0	0.012	-6.65	5	85	0.004	-1.48	7	27	0.045	-10.81	15	146	190
	Insula	0.430	10.39	22	0	0.010	-5.42	0	13	0.006	2.92	7	0	0.095	-15.16	0	22	22
	Caudate	0.243	6.36	25	0	0.032	-10.65	0	23	0.006	-2.97	0	8		-	-	-	25
	Putamen	0.107	5.24	21	0	0.125	-26.15	0	23	0.091	-21.99	0	22		-	-	-	23
	Thalamus	0.256	6.61	25	0	0.067	-19.71	0	19	0.046	-14.91	2	14		-	-	-	25
	Cerebellum	0.259	8.16	213	0	0.021	-3.70	26	88	0.016	2.18	49	48		-	-	-	213

Table 2.4 Net percent change from 8 to 23 years and variance explained by MNI label. **Pct**, Percent change of fitted values from 8 to 23 years. **N** ↑, Number of parcels with positive net change (increase); **N** ↓, Number of parcels with negative net change (decrease). Only parcels that survived FDR correction have been considered.

change and R2 for each modality are shown in **Figure 2.3** and are available in interactive format online (https://egenn.github.io/gmdvdev.html).

Figure 2.4 helps describe how development modulates sex effects by plotting the average difference of male and female fitted values per modality per MNI region by age from 8 to 23 years. Males and females have no differences in GMD at age 8, but females start to lead soon thereafter throughout the brain. Males have higher GMV and GMM on average in each MNI region throughout this age range. Only CT shows a change in the direction of sex differences with age. Males have higher CT in bilateral insula until about age 12 and in frontal and occipital lobes until age 15, at which

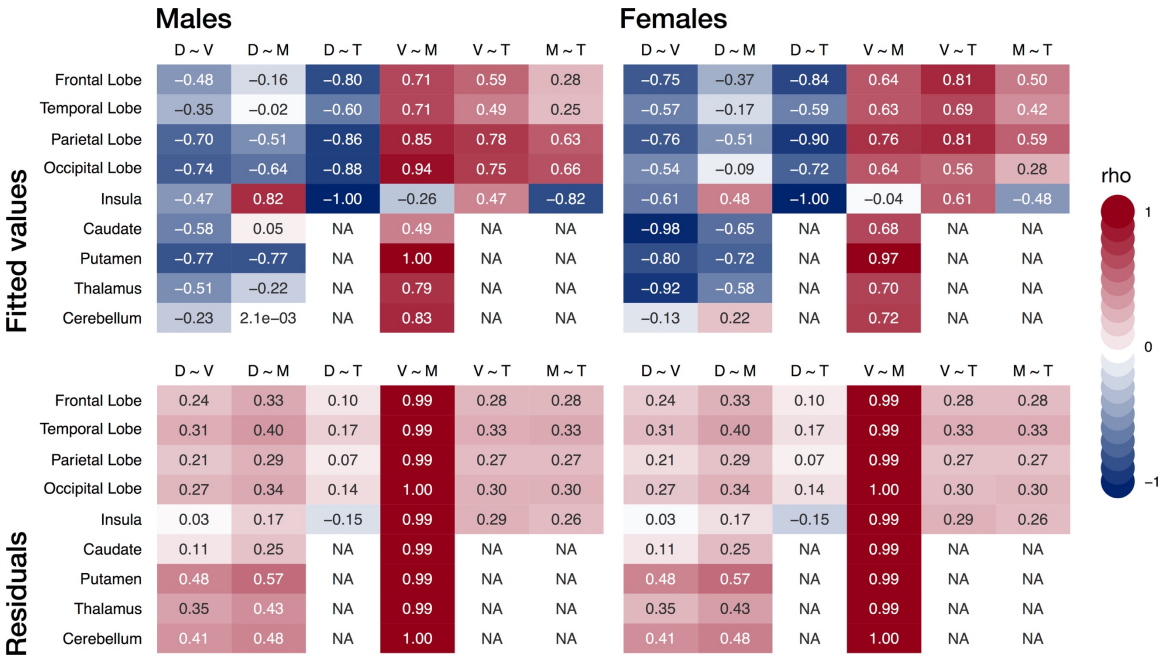

Figure 2.5 Intermodal correlations averaged by MNI label. Pairwise spearman correlations (rho) were estimated between the fitted values of model 3 (top row) of all gray matter measures to summarize the similarity of age-related effects among modalities and between their residuals (bottom row). Brain slices with these results are available on-line at https://egenn.github.io/gmdvdev/imcor.html. **D**, Gray matter density; **V**, gray matter volume; **M**, gray matter mass; **T**, cortical thickness.

points the effect reverses leading to a female advantage. In most cases, the sex differences have a nonlinear relationship with age.

GMM largely resembles GMV, not GMD

We defined GMM as the product of GMD and GMV in order to study a native-space equivalent of modulated density, a very popular measure in structural neuroimaging studies. GMM showed age-related effects that, for the most part, closely paralleled those of GMV (**Figure 2.5**). This is probably because GMV has much higher variance than GMD (**Table 2.1**), and consequently

contributes much more to the variance in GMM explained by age than GMD does. Despite this, three MNI regions showed, on average, opposite direction of net change in GMM than in GMV, i.e. an increase instead of a decrease. This was observed in both males and females, in descending order of magnitude, in the insula, the temporal lobe, and the cerebellum (**Table 2.4**).

Development modulates intermodal relationships among structural measures

To summarize the differences in age-related effects among the four measures with a single quantity, we calculated pairwise correlation coefficients of fitted values for our sample's age range (8 - 23 y.o.; **Figure 2.5**, top row). Spearman correlation was used as most fitted values are non-linear. As expected from the results above, GMD was negatively correlated with GMV and CT throughout the brain and cortex respectively. GMM was positively correlated with GMV and CT in all MNI regions, on average, with the exception of the insula. Looking at the same pairwise correlations among the residuals, i.e. after removing the effect of age, we see that most correlations are positive, with notable exception of the density and thickness pair in the insula (**Figure 2.5**, bottom row). Intermodal correlations of residuals help suggest what relationships may look like in the absence of an age effect but are no substitute for directly examining separate age bins, which should ideally extend across the lifespan.

DISCUSSION

Despite extensive use of different MRI-derived gray matter measures in the literature, very few attempts have been made to directly compare them or their developmental patterns, and they are often wrongly assumed to be equivalent. This study shows distinct age-related effects and sex differences on whole brain and regional measures of gray matter density (GMD), volume (GMV), mass (GMM = GMD * GMV), and cortical thickness (CT) in a cross-sectional dataset of 1189 youths aged 8 to 23 years drawn from the Philadelphia Neurodevelopmental Cohort. A custom T1 preprocessing pipeline and a novel high-resolution gray matter parcellation were created in order to produce unbiased gray matter segmentations without use of priors, and to extract native space measures without any interpolation or smoothing. Our findings partly challenge the widely-held, though vague, view that "gray matter declines" from childhood to young adulthood, and provide a more complete description of developmental gray matter differences.

Not all gray matter declines in adolescence

Mean brain GMD increases from childhood to young adulthood, while total brain GMV and mean CT decrease. Total GMM only shows a slight decrease from 8 to 23 years, suggesting that an increase in density may partly counter a decrease in volume. Regionally, GMD increases virtually throughout the brain. GMV, on the other hand, decreases on average in all lobes and

subcortical regions, but there are parcels within those broad regions whose volume increases, particularly in frontotemporal cortex. Future work will determine whether areas that expand during adolescence despite an overall decline in volume support the enhancement of specific neurocognitive functions.

Age-related effects and sex differences in density may help understand cognitive abilities

We know that higher GMV correlates with higher neurocognitive performance in adults (Gur et al., 1999; McDaniel, 2005), which gives rise to two apparent paradoxes: 1. Adolescence is characterized by a sharp rise in neurocognitive performance (Gur and Gur, 2016), despite a decline in GMV. 2. There are no significant sex differences in general intelligence (Halpern et al., 2007), despite a male advantage in GMV. Our results suggest that age-associated volume decrease might be compensated for by increasing gray matter density during adolescence and lower volume in females might be compensated for by higher density throughout the brain.

Biological basis of structural MR measures: the need for large scale, quantitative histological – MRI studies

The above findings beg the question: What do GMD and the other gray matter measures mean in terms of biology? Multiple studies have shown that the T1 signal is sensitive to myelin and iron content, whose distributions overlap

significantly within cortical gray matter (Stüber et al., 2014). Surprisingly, only few attempts have been made to quantify the relationships among histological features and MRI-derived structural measures. A large number of studies using T1-weighted imaging to quantify gray matter have focused on neurodegeneration. Such diseases result in neuronal loss which causes direct decreases in all gray matter measures. This may partly explain the confusion that has led to these measures often being considered highly correlated or equivalent, and even grouped together in meta-analyses (for example, see (Shao et al., 2014). However, in the context of normal brain structure, or in brain disease without extensive neuronal loss, including most psychiatric disorders, regional and global variation in different gray matter measures may be less correlated, even anti-correlated, as seen here between GMD and GMV. We expect that MRI-derived gray matter measures are differentially determined by a set of histological factors, including neuronal and glial number and size, dendritic arborization, number of axonal projections and extent of myelination. Their effects will vary by age, brain region, and cortical layer.

In adolescence, MRI-estimated decline in gray matter volume is generally attributed to a combination of synaptic pruning of exuberant connections, a regressive event, and increasing myelination, a progressive event, both essential aspects of normal development (Stiles and Jernigan, 2010). While pruning results in a direct reduction in neuropil, myelination

may have multiple direct and indirect effects on T1-based gray matter quantification. White matter myelination and expansion may result in a physical outwards shift of the gray-white matter boundary causing gray matter to compact and leading to decreases in GMV and CT and increases in GMD. Alternatively, myelination near the gray-white border may increase signal intensity in voxels nearest the border enough to switch their classification from gray to white, which would lead to reduction of volume and thickness measurements but have no effect on density, since these voxels would be now excluded from any gray matter parcels. At the same time, cortical gray matter also contains substantial amounts of myelinated fibers with significant regional variability (Nieuwenhuys, 2013) and intracortical myelin also increases during adolescence (Grydeland et al., 2013). Increasing cortical myelination would lead to a decrease in estimated GMD, which means that GMD increases reported in this study are possibly underestimates. Rabinowicz et al performed stereologic morphometry in six males and five females aged 12 to 24 years and reported significantly higher neuronal densities and neuronal number estimates in males than females, but no sex differences in cortical thickness, suggesting higher neuropil mass / increased neuronal processes in the female cortex (Rabinowicz et al., 2009), which might explain our findings of higher GMD in females.

We chose to compare four different measures of regional gray matter in volumetric space. Other morphometric and morphological measures like

cortical surface area and gyrification index can also be derived from T1 images in surface space analyses. The limited positive correlation we found between GMV and CT age-related effects is probably explained by independent changes in surface area and gyrification (Raznahan et al., 2011). While the majority of brain regions show significant sex effects on GMV as expected, a minority of regions showed a significant sex effect on CT. Considering that gray matter volume roughly equals surface area times cortical thickness, we expect surface area to exhibit more extensive sex differences than thickness. We limited our analysis to volumetric space measures in order to use the same parcellation for each measure and avoid the extra resampling and registration errors introduced in the conversion between the two spaces (Klein et al., 2010). For the same reason, care must be taken when comparing volumetric and surface space analyses.

Given the important gaps in our understanding of the links between biology and imaging, it is crucial to design large-scale, combined MRI and histological quantification studies to fully characterize the neurobiological basis of raw MRI signals and derived measures. Biophysical modeling of MR-derived measures will enable accurate noninvasive in vivo prediction of histological features (Stiles and Jernigan, 2010). This will be crucial in elevating the potential of neuroimaging in the investigation of nervous system physiology and pathology, disease diagnosis, and treatment monitoring.

Limitations and implication for future work: Phenotypes of structural brain development and links to cognition

The cross-sectional design of this study was its main limitation. Ongoing longitudinal studies will provide true measures of developmental change and allow the analysis of inter-individual differences in development. Future studies would also benefit from inclusion of more MRI modalities. New diffusion-weighted MRI techniques like neurite orientation dispersion and density imaging (NODDI) may provide rich information on gray matter structure and complement T1 and T2 signals (Zhang et al., 2012). Histological morphometry has shown cortical layer- and type-specific changes in neuronal cell bodies (Rabinowicz et al., 2009), which cannot be resolved with today's common MRI sequences but this may be possible in the future. We must note that while different segmentation software employ similar methods for GMD estimation, results are dependent on parameter selection. Correlation with histology will also help guide these choices and optimize pipelines to produce measures with maximal biological interpretability.

Our results demonstrate that GMD, GMV, and CT must be considered distinct and complementary. They also further emphasize the need for nonlinear modeling and accounting for sex differences. We found that GMD and CT are most sensitive to age, which makes them prime candidate biomarkers of brain development. In contrast, modulated density or GMM may not be very informative in a developmental context, and it is best to

consider GMD and GMV independently. We also show that intermodal relationships change with age, which further emphasizes that neuroimaging findings should not be generalized from one age period to another. We have previously shown that structural covariance networks develop during childhood to mirror adult functional intrinsic connectivity networks (Zielinski et al., 2010a). Ongoing work aims to identify how different structural measures can be best applied to study cognition and disease.

As we advance from group-level to individual-level studies, from unimodal to multimodal analyses, and from descriptive to predictive models with the aim of integrating neuroimaging into clinical practice, it is essential to make best use of all available data. The first step is to understand available measures and the relationships among them. Development is a critical dimension on which these relationships may vary and adds to the challenge and the importance of this task.

The work in this chapter was published in the Journal of Neuroscience (Gennatas et al., 2017) and was featured on the cover of the May 17, 2017 issue (http://www.jneurosci.org/content/37/20/i)

3. GRAY MATTER INCREASINGLY PREDICTS COGNITIVE PERFORMANCE DURING ADOLESCENCE

INTRODUCTION

The prospect of predicting people's cognitive ability has always fascinated man. A large part of Neuroscience, and particularly Cognitive Neuroscience, is broadly concerned with understanding how the brain gives rise to the mind. A complete characterization of the vast networks of interactions from genes and molecules to cells, circuits, systems, to the whole brain and, finally, behavior may be very far off. However, it is possible to use brain data to predict clinical and cognitive outcomes, despite an incomplete understanding of the underlying biology. Such work can feed back into both basic neuroscientific research and clinical applications.

Early work in the field looked into correlations of intelligence with measures of head size and, later, MRI-derived estimates of whole brain volume. A meta-analysis of 37 datasets estimated population correlation between brain size and intelligence at 0.33 (McDaniel, 2005). Later studies focused on regional correlations, attempting to localize brain regions most contributing to intelligence, but were limited to mass-univariate analyses (Narr et al., 2007), which ignore relationships among brain regions and interactions. Multivariate predictive models trained on structural brain data have mostly focused on age prediction (Franke et al., 2012), in some cases relating brain development to cognition (Erus et al., 2014).

In this study, we train models to compare the ability of four structural brain measures derived from T1-weighted MRI to predict performance in a verbal reasoning task collected on the Philadelphia Neurodevelopmental Cohort using the Computerized Neurocognitive Battery (CNB) (Gur et al., 2010). The CNB has been widely administered in multiple settings and populations and translated to over fifteen languages. Instead of deriving a study-specific general factor, we chose a verbal reasoning (completing analogies) as the outcome of interest, which is known to correlate strongly with overall performance (Moore et al., 2015) and is one of the most commonly tested domains in standardized and other aptitude tests. We hypothesized that prediction accuracy of cognitive performance from gray matter measures increases with age, but made no prediction as to which measure would be the best predictor. We report that gray matter alone can predict up to 20% of variance in verbal reasoning performance of young adults estimated on out-of-sample data using 10-fold cross-validation.

MATERIALS AND METHODS

Subjects and neuroimaging

Subject selection and quality assessment of T1-weighted imaging was performed as described in **Chapter 2**, *Materials and Methods*. Of the initial 1189 subjects, 899 (478 females) with a valid CNB collected within twelve months of the structural MRI scan were selected for this study. We used the same T1-

derived measures of gray matter described in **Chapter 2**: gray matter density (GMD), gray matter volume (GMV), gray matter mass (GMM = GMD x GMV), and cortical thickness (CT), extracted from the same high-resolution parcellation (PNC-GMD1625, **Figure 2.1**). In order to study the effect of age on the prediction of cognitive performance and check for an interaction between brain data and age on prediction accuracy, the sample was stratified on age by splitting into terciles: Children (N = 299, 153 females; 8 – 12.7 years), Adolescents (N = 373, 192 females, 12.7 – 17.3 years), and Young Adults (N = 227, 133 females, 17.3 – 22 years).

Generalized Additive Models: Whole brain data & age

Models were trained to predict verbal reasoning scores from gray matter measures at two different scales: whole brain data (single value per gray matter measure) and regional brain data derived from the PNC-GMD1625 parcellation (GMD, GMV, GMM: N = 1625; CT: N = 1339 gray matter parcels). Whole brain mean GMD, total GMV, total GMM, and mean CT were used to predict verbal reasoning in each age tercile using Generalized Additive Models (GAMs) within the ***rtemis*** package (see **Chapter 4**). The ***learnCV*** function of rtemis was used to perform 10-fold cross-validation for model testing and average test set mean squared error (MSE) was calculated for each gray matter measure for each age group.

A common concern in developmental neuroimaging studies is controlling or correcting predictors and/or outcomes for age. Since most, if not all, measures of brain and performance and indeed many unrelated measures and artifacts correlate with age, it is easy to derive spurious correlations driven by age. However correcting for age can lead to signal loss and/or introduction of artifact. We chose not to age-regress either the neuroimaging data or the cognitive scores and instead used age stratification as described above. On top of that, cross-validated prediction of verbal reasoning scores from age alone was performed for each modality for each age group using Generalized Additive Models to measure directly the predictive power of age on performance.

Gradient Boosting: High dimensional regional brain data

Predictive models from high dimensional data were trained for each modality for each age group using gradient boosting of linear models as implemented in the *XGBoost* package (Chen and Guestrin, 2016). All training was performed again using the ***learnCV*** function within ***rtemis*** (**Figure** 3.1) to perform nested resampling for model tuning and testing. 10-fold cross-validation was used for testing (outer resampling). For each fold, 10 stratified bootstraps of the training set (inner resampling) were used to tune the L2 regularization weight (range: 0 - .3), and the number of boosting iterations using an early stopping rule (no improvement in validation set MSE for fifty iterations).

Models were trained with this procedure to predict verbal reasoning scores from GMD, GMV, GMM, and CT regional values, in turn, for each age tercile. Boosting of linear models was used for these high-dimensional datasets as execution times are orders of magnitude shorter than boosting trees. Parcelwise (mass-univariate) correlations were estimated between each gray matter measure and verbal reasoning scores to qualitatively compare univariate and multivariate effects at each age group.

RESULTS

Regional gray matter correlates weakly with verbal reasoning

Mean parcelwise correlations of GMV with performance are stable from childhood to adolescence at 0.10 and increase slightly into young adulthood to 0.15. On the other hand, an average correlation of 0.12 between GMD and performance in children diminishes to -0.04 in adolescence and -0.07 in

Example *learnCV* call for model tuning and testing:

```
LAN.GMD.MF3 <- learnCV(x = GMD.MF3, y = LAN.MF3, 'xgblin',
                       params = list(lambda = seq(0, .3, .1),
                                     resampler = 'strat.boot'),
                       outdir = '/Projects/CNBpredict/LAN.GMD.MF3/')
```

Figure 3.1 This ***learnCV*** command will train 10 tuning models + 1 final model for each of 10 folds, save an *rtemis* object containing its full output in an .Rds file along with PDF files of plots for True vs Fitted and True vs. Predicted values and a density plot of MSE in the specified output directory.

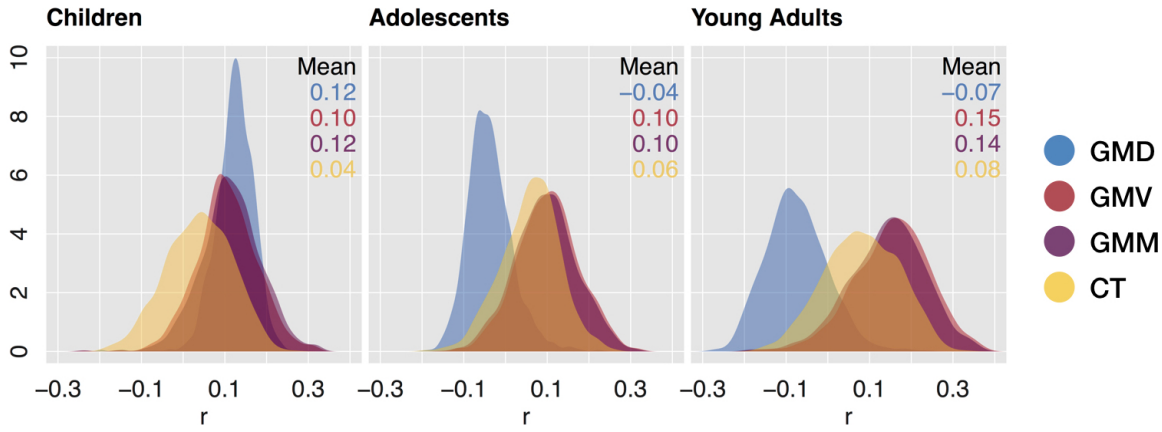

Figure 3.2 Density plots of parcelwise Spearman correlations between regional gray matter measures and verbal reasoning (GMD, GMV, GMM: N = 1625; CT: N = 1339 parcels). **GMD**, gray matter density; **GMV**, gray matter volume; **GMM**, gray matter mass; **CT**, cortical thickness.

young adulthood. Density plots of correlation values between regional gray matter and verbal reasoning are shown in **Figure 3.2**.

Whole brain volume is a good predictor of verbal reasoning

At the whole brain level, Generalized Additive Models reveal that GMV is the best predictor of performance, particularly in the oldest group, where it explains 20% of the variance, estimated after 10-fold cross-validation. Mean whole brain GMD and CT fail to predict performance. GMM (= GMD x GMV) was included for comparison and is shown to track GMV for the most part and will not be discussed further.

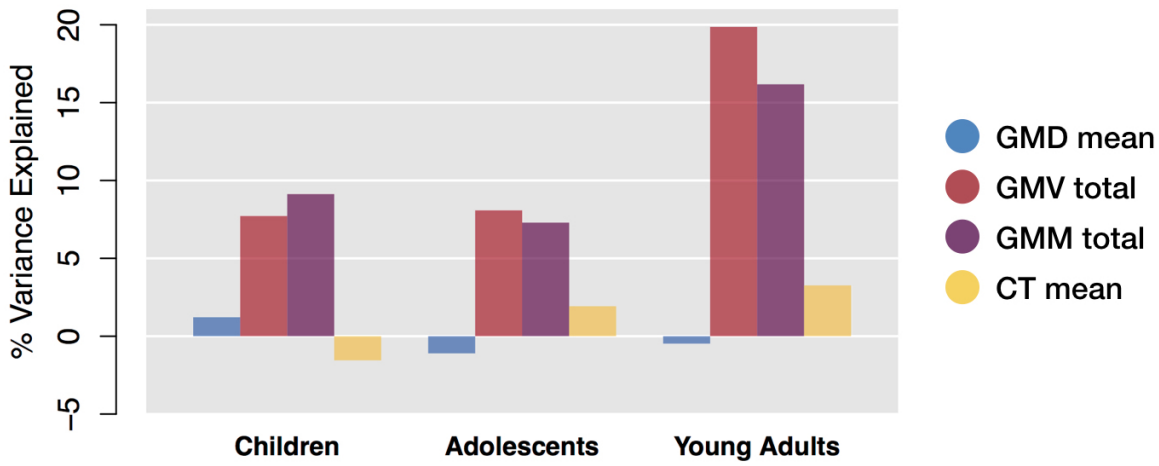

Figure 3.3 Prediction of performance using Generalized Additive Models of whole brain data

Multivariate models of regional GMD increasingly predict performance

Using the full high-dimensional dataset of each modality, we trained multivariate models using an efficient procedure of linear model boosting to predict verbal reasoning. Patterns of GMD are the best predictors of performance, showing an increase in prediction accuracy from childhood to young adulthood, when it reaches 20% of explained variance, on average, after 10-fold cross-validation. GMV and CT trail behind at around 10% of variance explained in the young adult group.

Age predicts performance only in children

To ensure that the above results are not driven by shared correlations of predictors and outcome with age, we trained Generalized Additive Models to predict performance from age alone for each age group. Interestingly, age explained 13% of variance in verbal reasoning scores in the children's group

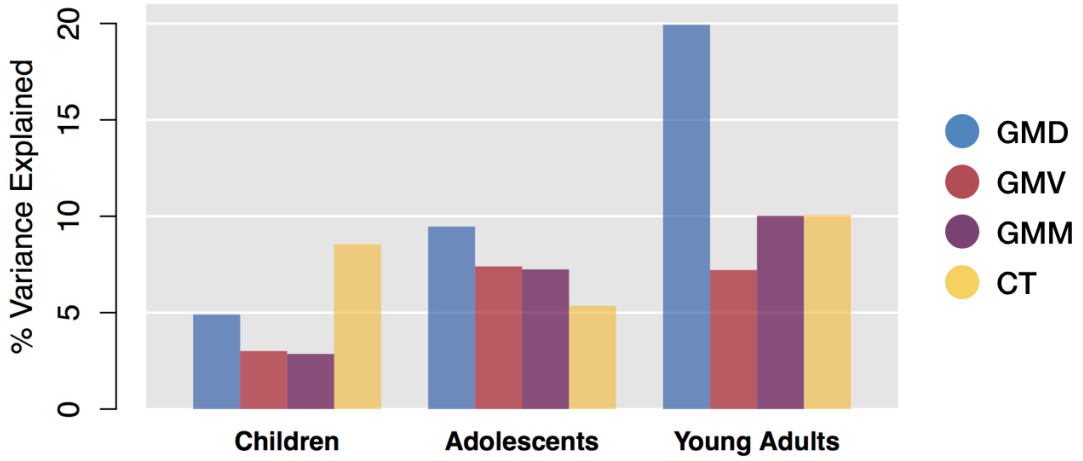

Figure 3.4 Prediction of performance using gradient boosting of regional gray matter measures

and not at all in the other two (R-squared = -0.92% and -1.71% for adolescents and young adults respectively). This suggests that mean regional GMD correlation within the children group may be driven by age (mean r = 0.12, **Figure 3.2)**, and demonstrates that the predictive power of whole brain GMV or regional GMD in the young adult group are not driven by age at all (**Figures 3.3, 3.4**).

DISCUSSION

Following the findings in **Chapter 2** where we showed that different gray matter measures exhibit distinct age-related and sex effects during development from childhood to young adulthood, we suggested that they should be treated as independent and complementary. We then set out to examine how they compare in their ability to predict a measure of cognitive

performance. We chose to study verbal reasoning, a measure that correlates very highly with overall intelligence, and which was collected on the Philadelphia Neurodevelopmental Cohort as part of the Computerized Neurocognitive Battery. We have previously shown that structural covariance of modulated gray matter density (volume) increasingly mirrors resting state functional connectivity from childhood to young adulthood (Zielinski et al., 2010b). We hypothesized that structural-functional relationships grow stronger with age as developmental structural changes slow down and functional activation patterns stabilize.

We showed that prediction accuracy of verbal reasoning scores increases with age as predicted. Interestingly, whole brain GMV alone was a good predictor of verbal reasoning, explaining 20% of variance, but multivariate patterns of GMV, using measures from 1625 gray matter regions as predictors, failed to reach the same level of accuracy, explaining only 10% of variance. In contrast, mean brain GMD did not predict performance at all, but multivariate analysis of GMD explained 20% of variance in performance. This example emphasizes the power of multivariate models in neuroimaging even in the absence of strong mass-univariate results.

The main limitation of this work was the sample size after dividing into three age bins, which limited performance of the multivariate models. Because of the small sample size, the two sexes were considered together. Future work would certainly benefit from studying more age bins of narrower

age range each, separately for males and females. Interestingly, our results suggest that a sweet spot exists in the resolution of brain parcellation, which would likely be different for each measure and should ideally be tuned in the future. Ongoing work is looking to address this in two ways: through direct comparison of multiple parcellations of variable resolution, and by sparse decompositions of high resolution data with variable number of dimensions and sparsity.

4. ADVANCED BIOMEDICAL DATA ANALYSIS WITH *rtemis*

INTRODUCTION

Advances in biomedical science are helping generate an increasing volume and variety of data, at increasing velocity, though often of uncertain veracity (http://www.ibmbigdatahub.com/infographic/four-vs-big-data). Along with increased requirements for data warehousing and privacy control, this raises the need for sophisticated analytic methods to extract insights and guide decision making. Group-level hypothesis testing is slowly being replaced by subject-level predictive modeling (Bzdok et al., 2016). Mass-univariate analyses of unimodal data are increasingly supplanted by multivariate analyses of high-dimensional, multimodal data. As data science is embraced across fields and industries, the benefits of research and development at the theoretical and applied level are shared by all. However, in biomedical research, access to the best available algorithms is often limited by researchers' technical expertise. A growing, inhomogeneous ecosystem of software packages running on multiple programming languages and often lacking good documentation adds an extra layer of complexity on top of the variety of algorithms and approaches. We present ***rtemis***, an open source package written in R designed to make advanced data analysis and visualization more efficient and accessible. ***rtemis*** provides a unified framework for data analysis by taking advantage of the R language and some of the best algorithms and packages available.

IMPLEMENTATION

rtemis is implemented in the R language (The R Project for Statistical Computing; https://www.r-project.org), a free and open source language for statistical computing and graphics, the *de facto* programming language of statisticians. It capitalizes on multiple existing, high quality R packages available either through the Comprehensive R Archive Network (CRAN; https://cran.r-project.org), the Bioconductor repository of open source software for bioinformatics (https://www.bioconductor.org), or directly through public GitHub repositories (https://github.com). It runs on all operating systems that support R, which include macOS, Linux, and Windows. Two main advantages of the R language are:

- It is built specifically for quantitative analysis /statistical computing and makes most common and a lot of advanced quantitative and statistical functionality directly accessible.
- The collection of statistics-related contributed packages in R far surpasses that of any other language.

Design Principles

A core principle behind the design of ***rtemis*** was to make it as easy and fast as possible for the user to get from data to results in a reproducible fashion even without much prior experience in data analysis. The following are some of the main design goals of the package:

```
> LAN.Age.MF3.gamCV <- learnCV(Age.MF1, LAN.MF3, 'gam')
[2017-06-07 15:09:25 learnCV] Hello, egenn
Error in dataPrepare(x, y, NULL, NULL) :
 Training set predictors and outcome do not contain same number of cases
```

Figure 4.1 Error reporting in ***rtemis*** attempts to pinpoint the source of the problem and relay in simple language. In this example, ***dataPrepare***, a helper function which prepares data ahead of all model training, checks whether the correct number of cases is present in predictors and outcome.

- Minimize the amount of code that needs to be input manually by the user, thereby minimizing user time and the probability of user error.
- Minimize computation time (running time) by allowing parallel (and distributed) execution where possible.
- Provide a user friendly and intuitive interface; minimize need to consult the manual.
- Make data analysis pipelines more transparent using informative messaging, error reporting, and logging (for example, see **Figure 4.1**).

R6 class system

During early development, ***rtemis*** was implemented using classic *S3* methods (Chambers, 1991). As the project grew, the need arose for formal object and method definitions. Objects and associated methods were built initially using all available class systems for comparison: S4, Reference Class (RC, sometimes referred to as R5), and R6. The last two were preferred for their ability to include methods within the object itself (similar to Python objects; **Figure 4.2**). RC and R6 objects also use pass-by-reference (again similar to

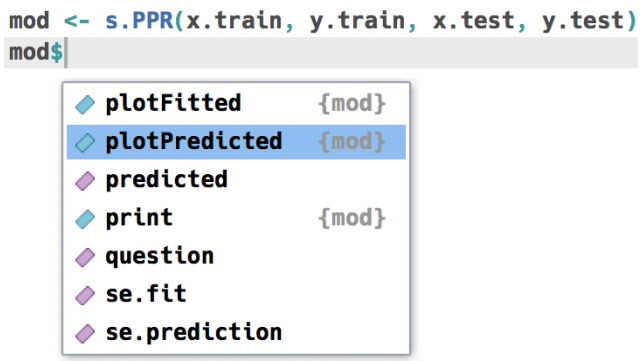

Figure 4.2 Example of an R6 object of class 'rtMod' used for all supervised learning. Attributes (e.g. *predicted* - outcome values predicted by the model from the testing dataset) and functions (e.g. *plotPredicted* - plots Predicted vs. True outcome values) are both accessible directly from within the object.

Python), which can be advantageous when manipulating large datasets as they help reduce memory load. R6 was finally chosen for its lightweight and fast implementation (https://cran.r-project.org/web/packages/R6/vignettes/Performance.html), and its backing by core R projects and developers.

Classes have been implemented for supervised learning (***rtMod***: all models; ***rtModBag***: - bagged models; ***rtModCV***: - cross-validated models), clustering (***rtClust)***, decomposition (***rtDecom***), and cross-decomposition (***rtXDecom***).

VISUALIZATION

It is difficult to overstate the importance of data visualization. It is an essential part of data analysis that can play an invaluable role before and after each preprocessing or modeling step. ***rtemis*** supports both static and dynamic/interactive graphics. The ***mplot3*** family of functions is responsible

Function Name	Input Data	Description
mplot3	vector x / vectors x & y	Alias for mplot3.x and mplot3.xy depending on input
mplot3.x	vector x	Index, Timeseries, Density, Histogram, QQ-line plots
mplot3.xy	vectors x & y	Scatter plot; incl. fit lines estimated with any rtemis learner
mplot3.xym	vectors x & y	Combination of mplot3.xy scatter & mplot3.x marginal plots (density and/or histogram)
mplot3.fit	vectors x & y	Alias for mplot3.xy with equal axes, diagonal, and fit lines
mplot3.bar	vector or matrix x	Barplots
mplot3.box	matrix x	Boxplots
mplot3.heat	matrix x	Heatmap with optional hierarchical clustering
mplot3.conf	confusion matrix	Confusion matrix for classification results
mplot3.roc	rtemis classification	ROC curve for classification models
mplot3.surv	survival::Surv object	Kaplan-Meier survival function
mplot3.img	matrix x	False color 2D image
mplot3.marginal	rtemis regression	Build a scatter plot by varying one independent variable
mplot3.cart	rpart model	Draw a decision tree trained by recursive partitioning
mplot3.adsr	A, D, S, R, I, O	Draw an envelope generator based on Attack time, Decay time, Sustain level, Release

Table 4.1 The ***mplot3*** family for static graphics

for producing static graphics in ***rtemis***. It uses layers of customized base graphics to produce publication-quality plots. **Table 4.1** lists the available ***mplot3*** functions and their description. All plots in this thesis were created using ***mplot3.***

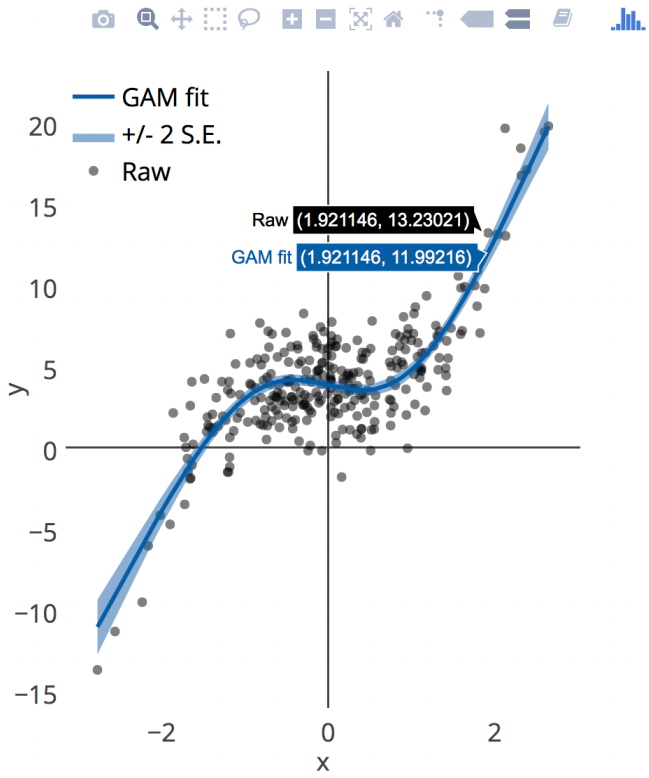

Figure 4.3 Screenshot of a dynamic plot drawn with ***dplot3***. Hovering the mouse over scatter points in this case displays raw and fitted values. Visibility of elements can be toggled by clicking on their name in the top-left legend.

Dynamic graphics are created with the *dplot3* and *dplot3.heat* functions built on the open source *plotly* platform (https://plot.ly/). They are viewable either within the RStudio Integrated Development Environment (IDE) or in a web browser (**Figure 4.3**).

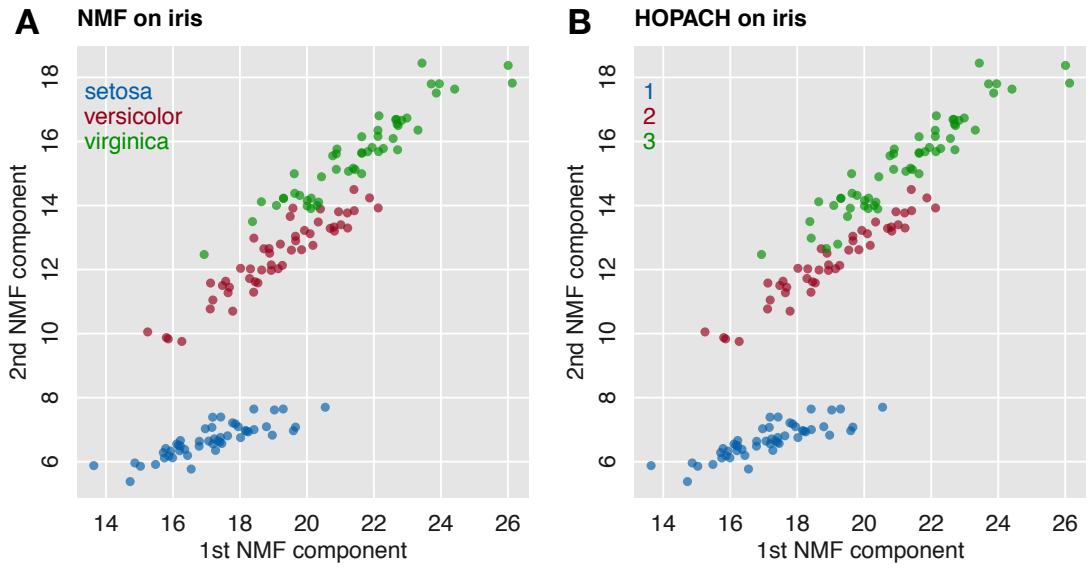

Figure 4.4 Unsupervised learning on the *iris* dataset (Anderson E, 1935). **A** Non-negative matrix factorization projects the dataset to two dimensions. Color indicates true flower species. **B** Hierarchical Ordered Partitioning and Collapsing Hybrid (HOPACH) algorithm (van der Laan and Pollard, 2003) separates cases into three categories with little error without any knowledge of real labels.

UNSUPERVISED LEARNING: Clustering & Decomposition

Unsupervised learning attempts to find structure in unlabeled data, i.e. without being guided by an outcome / dependent variable (cf. *Supervised Learning*). Consider an $n \times p$ dataset (n cases by p variables). Clustering, or Cluster Analysis, divides the n cases into k groups, resulting in a $k \times p$ dataset ($k < n$), based on a similarity / distance measure derived from the p variables. Matrix decomposition or factorization, on the other hand, projects a high dimensional dataset to a lower dimensional space, i.e. from $n \times p$ to $n \times p'$ ($p' < p$). If the original dataset is known or speculated to consist of a large number of measurements (variables) originating from a small number

of generators, or latent variables, decomposition can help recover them and in this can help gain insights into the true structure of the data. This is a common procedure in feature engineering. For example, variance in voxelwise neuroimaging brain data can be considered to result from the the action of a small number of networks which can be identified using decomposition algorithms, commonly Independent Component Analysis (ICA) for functional data. **Tables A.1** and **A.2** in the **Appendix** list algorithms available in ***rtemis*** for clustering and decomposition, respectively. Clustering functions begin with ***u.**** and output an object of class ***rtClust***, while decomposition functions with ***d.**** and output an object of class ***rtDecom***. For an example, se **Figure 4.4**.

SUPERVISED LEARNING: Classification, Regression, Survival

Supervised learning involves the prediction of an outcome of interest, or dependent variable, from a set of predictor variables, or independent variables, or features. The outcome may be a categorical or continuous variable. The process is called classification and regression, accordingly. Survival regression is a related approach that aims to predict time to an event (in medicine, usually death). All supervised learning function names in ***rtemis*** begin with ***s.**** followed by the algorithm alias found in **Table A.3**. Some features of supervised learning in ***rtemis***:

- Input data is checked for consistency and type of model is inferred from type of outcome: vector of factors -> Classification,

numeric vector -> Regression, matrix of time and status -> Survival Regression

- Automatic hyperparameter tuning: If more than a single value is provided for any parameter, grid search is automatically run by resampling the training set to create internal training and validation sets; the error is averaged across resamples for each combination of parameters and the combination minimizing error, on average, in the left-out sample is chosen. The final model is trained on the full training set using the identified parameters. Grid search can be exhaustive or randomized.
- Sensible defaults: Algorithm hyperparameters are set to values likely to perform well under common conditions. If no such values exist, functions are set to automatically tune hyperparameters.
- All learners output an object of class ***rtMod***, which supports all standard R methods for trained models: *coef, fitted, plot, predict, print, residuals, summary* (**Figure 4.5**).
- If an output directory is specified, the ***rtMod*** object is saved as an .Rds file (serialized R data file) along with plots of True vs. Fitted (training set) and True vs. Predicted (testing set) values in PDF format and a log text file with the full console output of the function.

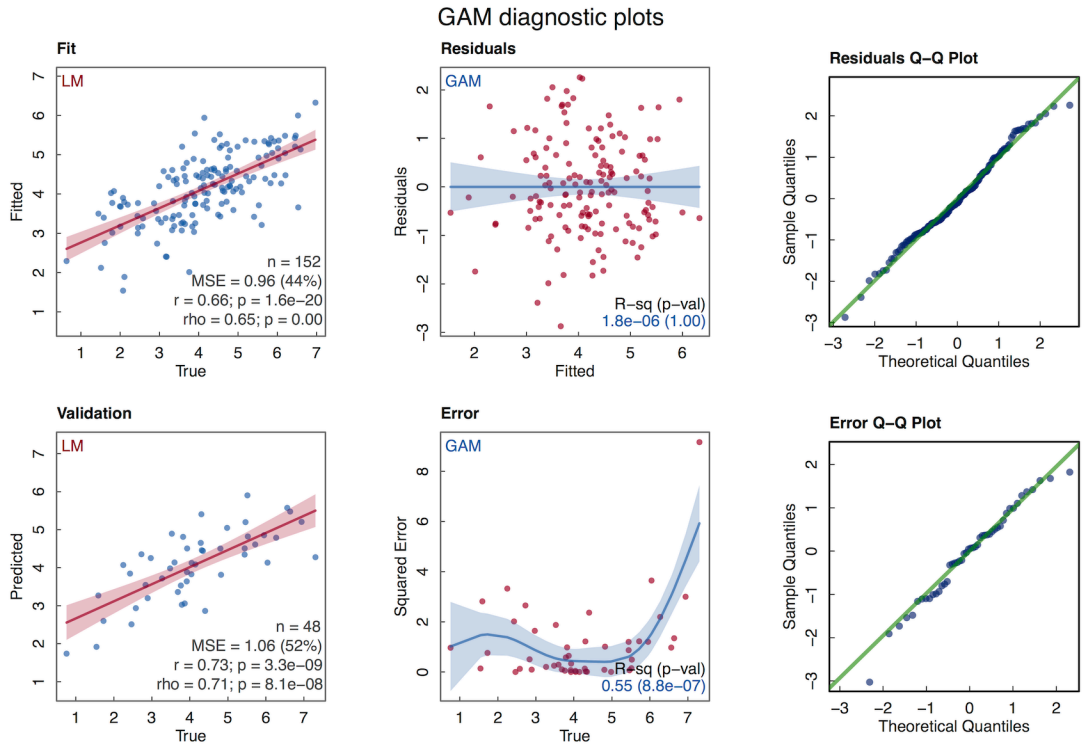

Figure 4.5 *summary* method on an ***rtMod*** object draws a panel of informative plots using ***mplot3***

learnCV: One-step model tuning and testing

learnCV is the main function for predictive modeling in ***rtemis***. It accepts a matrix of predictors *x* and an outcome vector *y*, creates resamples using ***resample*** (**Table A.5**), and trains any ***rtemis*** learner (**Table A.3**) on each resample. The output is saved to an ***rtModCV*** object, which also inherits from the ***rtMod*** object. The function aggregates fitted, predicted, and true values across resamples and estimates error across resamples.

bagLearn: Bootstrap aggregating

Bootstrap aggregating (bagging) can be run automatically for most learners (those that do not include it by design). Specifying the *bag.resampler.rtSet* argument triggers ***bagLearn***, an internal function that calls the originating learner function to train multiple resamples of the training set, produce predicted values from each using the testing set *x.test* and output each model's prediction and their average in an ***rtModBag*** object, which inherits from the ***rtMod*** object.

decomLearn: Decompose and learn

decomLearn takes advantage of the modular design of ***rtemis*** to tune a decomposer and train a learner using the low dimensional projections as predictors. Specifically, the function:

- Accepts training and testing sets of predictors and outcome, x.train, y.train, x.test, y.test
- Uses a *resampler* (**Table A.5**) to create resamples of x.train and y.train
- For each resample:
    - Uses a *decomposer* (**Table A.2**) to decompose internal training sets using exhaustive or randomized grid search on parameter combinations - e.g. sparseness = seq(.1, 1, .1), nvecs = c(3, 5, 12)
    - Uses a specified *tuner* (any learner function; **Table A.3**) to identify combination of parameters that minimizes prediction error

- Trains decomposition on full training set using identified parameters
- Trains final *learner* (**Table A.3**)
- Outputs an object of class ***rtDecomLearn***

CROSS-DECOMPOSITION

Cross-decomposition refers to methods like canonical correlation analysis (CCA), which decompose two or more datasets in parallel. They aim to derive sets of projections, one projection from each input dataset in each set, such that projections in eash set are maximally correlated. ***rtemis*** supports sparse CCA using the PMA package (available on CRAN) modified to run in parallel, and more advanced sparse decompositions provided in the ANTsR package (https://github.com/stnava/ANTsR). Cross-decomposition functions available in ***rtemis*** begin with ***x.**** and are listed in **Table A.4**.

META-MODELING

Meta-models are models whose input is the output of other models – i.e. models whose predictors are the estimates of other models. The process is commonly referred to as *stacking* (or *blending*, or *stacked generalization)* and has proven highly successful in many real world scenarios. The idea is that by pooling predictions from multiple base learners you can take advantage of different models' strengths and produce a final prediction better than the best individual prediction - ideally. Top performing entries in data science

competitions have almost invariably used some form of stacking. ***rtemis*** currently includes functions to automate training of three types of meta-models: the common approach of stacking outputs of different algorithms, here referred to as *model stacking*, and two custom meta-models we refer to as *modality stacking*, and *feature-weighted stacking*.

Model stacking

Model stacking is probably the most common and straightforward type of stacking. Assume you want to predict an outcome *y* given an input matrix *x*. You have multiple learning algorithms available but do not know ahead of time which one will perform best (Wolpert, 1996). In model stacking, suppose you create training and testing sets *x.train*, *y.train*, *x.test*, and *y.test*, you would:

- Split *x.train* and *y.train* into further training and testing sets based on *r* resamples: $x.train'_{1...r}$, $y.train'_{1...r}$, $x.test'_{1...r}$, $y.test'_{1...r}$
- For each resample *r*, train a set of *i* base learners to map $x.train'_{1...r}$ to $y.train'_{1...r}$ *and* get predictions $y.hat.test'_{(1...r,\ 1...i)}$ from data $x.test'_{1...r}$
- Concatenate across *r* and train a meta learner to map predictions of base learners' concatenated $y.hat.test'.cc_{1...i}$ to outcomes *y.test'.cc*
- Train base learners on full training set *x.train* and get predictions $y.hat.test_{1...i}$ from data *x.test*
- Pass $y.hat.test_{1...i}$ to the trained meta model to get final predictions *y.hat.test.meta*; estimate error by comparing to *y.test*

Modality stacking

Modality stacking is similar to model stacking but in this case base models differ by being trained on a separate dataset (modality) each (and may or may not use the same learning algorithm). For example, gray matter density, gray matter volume, fractional anisotropy, and regional cerebral blood flow can each be used to predict an outcome of interest. This procedure will often produce superior results to concatenating the datasets of different modalities into one extra wide dataset, as this exaggerates the p >> n problem (having many more predictors than cases), among other issues. It is implemented in the ***metaFeat*** function.

Group-weighted stacking

In group-weighted stacking (GWS), base models differ by being trained on differently weighted versions of the full sample. This is useful if you suspect that a different pattern of features will predict the outcome in each subset. Each base model is trained on the full set of cases, but cases not part of the group are down-weighted. A parameter *alpha* ($0 \leq \alpha < 1$) determines the weight of non-group cases. For example, if we expect sex differences in the pattern of brain regions that predict cognitive performance, we can use GWS to obtain better prediction accuracy than if we trained a single, non-stacked, model on males and females together. The α parameter should be tuned for performance. GWS is implemented in the ***metaGroup*** function.

Figure 4.6 PNC Explorer allows rich interactive data exploration and visualization using ***mplot3*** with point-and-click simplicity

rtemis-POWERED WEB APPLICATIONS

Many real-world scenarios require immediate access to visualization and data analytics, where the need for coding would be a major hindrance or simply prohibitive. Online dashboards, powered by open source or proprietary platforms, are becoming increasingly popular across fields and businesses and provide advanced functionality with point-and-click simplicity. We have taken advantage of the *shiny* web application framework (https://shiny.rstudio.com/) to create online, interactive web applications powered by ***rtemis***. These applications load on any web browser and allow the user to access ***rtemis*** functionality without the need to use any R code. The

Figure 4.7 PNC IMcor allows dynamic heatmap visualization of intra- and intermodal correlations using ***dplot3.heat***

web server is running ***rtemis*** in the background, obviating the need to install R, ***rtemis***, and its dependencies.

A pair of web applications were created to visualize the complete data release of the Philadelphia Neurodevelopmental Cohort (PNC). **PNC Explorer** provides access to some of the main ***mplot3*** plotting functionality in an interactive manner. It supports univariate and bivariate plotting: index, histogram, density, and scatter plots (**Figure 4.6**). It is paired with the **PNC IMcor**, which allows dynamic heatmap visualization of intra- and inter-modal correlations of multiple imaging datasets: gray matter density, gray

matter volume, mean diffusivity, regional cerebral blood flow, regional homogeneity, and amplitude of low frequency fluctuations (**Figure 4.7**).

The goal is to provide free online access to a series of apps, where users can upload and visualize their own data. Such functionality can be very useful in biomedical research, and essential in data-driven clinical applications.

DISCUSSION

rtemis aims to make advanced data analysis accessible to all. Some of its primary target groups are biomedical researchers and, eventually, clinical practitioners. A cheatsheet which highlights the core components of the package is available online at: https://egenn.github.io/docs/rtemisCheatsheet.pdf. The complete R-style manual can be found at: https://egenn.github.io/docs/rtemisCheatsheet.pdf. A vignette with examples of code and corresponding output (also viewable within RStudio's help viewer), is available at: https://egenn.github.io/rtemis/rtemis-vignette.html.

The design of ***rtemis*** and its core of shared internal functions allows for easy expansion and addition of new algorithms for supervised or unsupervised learning in the future. The modular architecture makes it simple to build custom meta-models and other combinations of supervised and unsupervised learning.

Current work on ***rtemis*** is focused on implementing interpretable machine learning algorithms. Current state-of-the-art algorithms provide

high accuracy at the expense of interpretability. Algorithms that are both highly accurate and interpretable will be profoundly beneficial to basic research by providing insights into effects and interactions of features within massive multivariate datasets and will also make possible the use of machine learning in clinical decision making.

5. DISCUSSION & FUTURE WORK

NEUROIMAGING & THE BRAIN

MR neuroimaging allows us to study human brain structure and function in a safe and noninvasive way and is an invaluable tool in advancing our understanding of normal brain physiology and pathology. It has already helped gain great insights into brain function, especially perception. At the same time, little progress has been made towards applying neuroimaging in clinical practice. Countless papers present weak or questionable findings on MR-derived measures without understanding of the underlying biology and do not hesitate to make extravagant promises that unlocking of the mysteries of brain disease and discovery of treatments are but a small step away. While such discoveries have not yet materialized and may be overdue, they are certainly possible. If neuroimaging is to deliver on its translational potential, studies require a solid link to biology and a path to application.

In **Chapter 2**, we attempted to clear some of the confusion surrounding different gray matter measures. We showed that gray matter volume (GMV) and gray matter density (GMD) show opposite age and sex effects in adolescence and should be considered complementary. Our findings may help explain how cognitive performance improves sharply during adolescence while GMV is reduced and how males and females show no differences in overall performance. An investment in the careful characterization of the relationship between brain histology and MR-derived measures is essential to

bridge neuroimaging with neurobiology. They may be treated as separate fields, but of course remain two highly complementary methods of studying the same organ system. We focused on gray matter measures but the same applies to all MR modalities.

ACCURATE & INTERPRETABLE: MACHINE LEARNING FOR BASIC RESEARCH AND PRECISION MEDICINE

In **Chapter 3**, we showed that structural measures can predict cognitive performance even in relatively small sample sizes and discussed the importance of multivariate predictive modeling over traditional mass-univariate hypothesis testing. However, larger sample sizes are necessary to build accurate and reliable models. In **Chapter 4**, we introduced an R package to make using and comparing different supervised and unsupervised learning algorithms faster and easier.

Other than limits to researchers' technical expertise, the second and fundamental reason why advanced analytic methods are not yet widely employed in biomedical research or clinical applications is reduced interpretability. Current state-of-the-art machine learning and deep learning approaches are highly successful in an array of specialized applications and advancing at a relatively fast pace. One of their main weaknesses remains their lack of transparency. "Black box" methods may offer good predictions, at best, but do not help us understand how and why the algorithm is making its decisions. This limits the insights we can gain into the question at hand,

and, more importantly, prevents human supervision of the process which is ultimately prone to catastrophic failure (Caruana R. et al., 2015). On the one hand, this limits the utility of machine learning methods in basic scientific discovery. On the other, legal, moral, and practical constraints prohibit their use in clinical practice. Unlike other applications of machine learning, there is minimal room for failure, or trial-and-error when human lives are involved. The development of interpretable machine learning methods will give researchers deeper insights into their data. More importantly, they will allow physicians to check and correct, as necessary, the learning algorithm's rules. Such technologies will be transformational for biomedical research, and usher in the era of precision medicine.

SHARING & CARING: THE NEED FOR TEAM SCIENCE

Brain research requires vast resources in terms of funding, personnel, infrastructure, and time. There are clearly limits to what can be achieved by a single investigator or lab. However, real progress can be achieved by collaborations among labs, institutions, and industries. The importance of team science in biomedical research is well understood (Hall et al., 2008). Its adoption may be hindered by the established tradition of competition for funding and recognition, but it is hopefully only a matter of time before it is embraced widely. Partnerships among universities, health systems, private and industrial Research and Development units are growing stronger and will

eventually be the norm. Some of the factors that will drive the success of such collaborations include:

- Homogenization of data collection protocols
- Public sharing of basic research data
- Standardization of evidence-based, free, and open source software
- Publication of data and code along with each research article
- Open review process
- Systematic replication of research findings

Neuroscience, Neurology and Psychiatry are set to benefit greatly from large-scale collaborative work. Many challenges remain to be addressed before effective treatments can be created, but team science is our best bet to get there.

APPENDIX – rtemis Algorithms

Table A.1 Clustering algorithms

Alias	Description
CMEANS	Fuzzy C-means Clustering
HARDCL	Hard Competitive Learning
HOPACH	Hierarchical Ordered Partitioning And Collapsing Hybrid
H2OKMEANS	H2O K-Means Clustering
KMEANS	K-Means Clustering
NGAS	Neural Gas Clustering
PAM	Partitioning Around Medoids
PAMK	Partitioning Around Medoids with k estimation
SPEC	Spectral Clustering

Table A.2 Decomposition algorithms

Alias	Description
CUR	CUR Matrix Approximation
H2OAE	H2O Autoencoder
H2OGLRM	H2O Generalized Low-Rank Model
ICA	Independent Component Analysis
ISOMAP	ISOMAP
KPCA	Kernel Principal Component Analysis
LLE	Locally Linear Embedding
NLCR	Non-Linear Cluster Reduce
NMF	Non-negative Matrix Factorization
PCA	Principal Component Analysis
SPCA	Sparse Principal Component Analysis
SVD	Singular Value Decomposition
TSNE	t-distributed Stochastic Neighbor Embedding

Table A.3 Supervised learning algorithms

Alias	Description	Class	Reg	Surv
ADABOOST	Adaptive Boosting	T	F	F
BART	Bayesian Additive Regression Trees	T	T	F
BRUTO	BRUTO Additive Model	F	T	F
CART	Classification and Regression Trees	T	T	T
CFOREST	Conditional Random Forest	T	T	T
CTREE	Conditional Inference Trees	T	T	T
C50	C5.0 Decision Tree	T	F	F
ET	Extra Trees	T	T	F
EVTREE	Evolutionary Learning of Globally Optimal Trees	T	T	F
GAM	Generalized Additive Model	T	T	F
GBM	Gradient Boosting Machine	T	T	T
GLM	Generalized Linear Model	T	T	F
GLMNET	Elastic Net	T	T	T
GLS	Generalized Least Squares	F	T	F
H2ODL	H2O Deep Learning	T	T	F
H2OGBM	H2O Gradient Boosting Machine	T	T	F
H2ORF	H2O Random Forest	T	T	F
KNN	k-Nearest Neighbor	T	T	F
LDA	Linear Discriminant Analysis	T	F	F
LIGHTGBM	Light Gradient Boosting Machine	T	T	F
LM	Ordinary Least Squares Regression	F	T	F
LOESS	Local Polynomial Regression	F	T	F
LOGISTIC	Logistic Regression	T	F	F
MARS	Multivariate Adaptive Regression Splines	T	T	F
MLGBM	Spark MLlib Gradient Boosting	T	T	F
MLMLP	Spark MLlib Multilayer Perceptron	T	F	F
MLRF	Spark MLlib Random Forest	T	T	F
MULTINOM	Multinomial Logistic Regression	T	F	F
MXFFN	MXNET Feed Forward Neural Network	T	T	F
NBAYES	Naive Bayes	T	F	F
NW	Nadaraya-Watson Kernel Regression	F	T	F
POLY	Polynomial Regression	F	T	F
POLYMARS	Multivariate Adaptive Polynomial Spline	T	T	F
PPR	Projection Pursuit Regression	F	T	F
PPTREE	Projection Pursuit Trees	T	F	F
QRNN	Quantile Neural Network Regression	F	T	F
RF	Random Forest	T	T	F
RFSRC	Random Forest (Survival, Regression,	T	T	T
RLM	Robust Linear Model	F	T	F
SPLS	Sparse Partial Least Squares	F	T	F
SVM	Support Vector Machine	T	T	F
TLS	Total Least Squares	F	T	F
XGB	Extreme Gradient Boosting	T	T	F
XGBLIN	Extreme Gradient Boosting of Linear Models	F	T	F

Class: Classification **Reg:** Regression **Surv**: Survival regression

Table A.4 Cross-decomposition algorithms

Alias	Description
CCA	Sparse Canonical Correlation Analysis
SD2RES	ANTsR sparse decomposition
SD2RESDEF	ANTsR sparse decomposition by deflation

Table A.5 Resampling methods

Alias	Description
kfold	Stratified k-fold cross-validation
strat.sub	Stratified subsampling
bootstrap	Bootstrap (sampling with replacement)
strat.boot	Stratified bootstrap

REFERENCES

Anderson E (1935) The irises of the Gaspe Peninsula. Bulletin of American Iris Society, 59, 2-5.

Avants BB, Epstein CL, Grossman M, Gee JC (2008) Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med Image Anal 12:26–41.

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

Bakhshi K, Chance SA (2015) The neuropathology of schizophrenia: A selective review of past studies and emerging themes in brain structure and cytoarchitecture. Neuroscience 303:82–102.

Blumenthal JD, Zijdenbos A, Molloy E, Giedd JN (2002) Motion artifact in magnetic resonance imaging: implications for automated analysis. Neuroimage 16:89–92.

Brain Development Cooperative Group (2012) Total and regional brain volumes in a population-based normative sample from 4 to 18 years: the NIH MRI Study of Normal Brain Development. Cerebral Cortex 22:1–12.

Brent BK, Thermenos HW, Keshavan MS, Seidman LJ (2013) Gray matter alterations in schizophrenia high-risk youth and early-onset schizophrenia: a review of structural MRI findings. Child Adolesc Psychiatr Clin N Am 22:689–714.

Bzdok D, Varoquaux G, Thirion B (2016) Neuroimaging Research: From Null-Hypothesis Falsification to Out-of-Sample Generalization. Educational and Psychological Measurement.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1721-1730). ACM.

Chambers JMHTJ (1991) Statistical Models in S. CRC Press, Inc.

Chen T, Guestrin C (2016) XGBoost: A Scalable Tree Boosting System. arXiv cs.LG.

Das SR, Avants BB, Grossman M, Gee JC (2009) Registration based cortical thickness measurement. 45:867–879.

Erus G, Battapady H, Satterthwaite TD, Hakonarson H, Gur RE, Davatzikos C, Gur RC (2014) Imaging Patterns of Brain Development and their Relationship to Cognition. Cerebral Cortex:bht425.

Franke K, Luders E, May A, Wilke M, Gaser C (2012) Brain maturation: predicting individual BrainAGE in children and adolescents using structural MRI. Neuroimage 63:1305–1312.

Gennatas ED, Avants BB, Wolf DH, Satterthwaite TD, Ruparel K, Ciric R, Hakonarson H, Gur RE, Gur RC (2017) Age-related effects and sex differences in gray matter density, volume, mass, and cortical thickness from childhood to young adulthood. J Neurosci. 37(20):5065-5073.

Glasser MF, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E, Ugurbil K, Andersson J, Beckmann CF, Jenkinson M, Smith SM, Van Essen DC (2016) A multi-modal parcellation of human cerebral cortex. Nature 536:171–178.

Gogtay N, Giedd JN, Lusk L, Hayashi KM, Greenstein D, Vaituzis AC, Nugent TF, Herman DH, Clasen LS, Toga AW, Rapoport JL, Thompson PM (2004) Dynamic mapping of human cortical development during childhood through early adulthood. Proc Natl Acad Sci USA 101:8174–8179.

Grydeland H, Walhovd KB, Tamnes CK, Westlye LT, Fjell AM (2013) Intracortical myelin links with performance variability across the human lifespan: results from T1- and T2-weighted MRI myelin mapping and diffusion tensor imaging. J Neurosci 33:18618–18630.

Gur RC, Richard J, Hughett P, Calkins ME, Macy L, Bilker WB, Brensinger C, Gur RE (2010) A cognitive neuroscience-based computerized battery for efficient measurement of individual differences: standardization and initial construct validation. J Neurosci Methods 187:254–262.

Gur RC, Turetsky BI, Matsui M, Yan M, Bilker W, Hughett P, Gur RE (1999) Sex differences in brain gray and white matter in healthy young adults: correlations with cognitive performance. 19:4065–4072.

Gur RE, Gur RC (2016) Sex differences in brain and behavior in adolescence: Findings from the Philadelphia Neurodevelopmental Cohort. Neurosci Biobehav Rev.

Hall KL, Feng AX, Moser RP, Stokols D, Taylor BK (2008) Moving the Science of Team Science Forward. American Journal of Preventive Medicine. 35(2):S243-S249.

Halpern DF, Benbow CP, Geary DC, Gur RC, Hyde JS, Gernsbacher MA (2007) The Science of Sex Differences in Science and Mathematics. Psychol Sci Public Interest 8:1–51.

Hyman SE (2007) Can neuroscience be integrated into the DSM-V? Nat Rev Neurosci 8:725–732.

Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, Sanislow C, Wang P (2010) Research Domain Criteria (RDoC): Toward a New Classification Framework

for Research on Mental Disorders. American Journal of Psychiatry.

Insel TR, Cuthbert BN (2015) Medicine. Brain disorders? Precisely. Science 348:499–500.

Klein A, Andersson J, Ardekani BA, Ashburner J, Avants B, Chiang M-C, Christensen GE, Collins DL, Gee J, Hellier P, Song JH, Jenkinson M, Lepage C, Rueckert D, Thompson P, Vercauteren T, Woods RP, Mann JJ, Parsey RV (2009) Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. 46:786–802.

Klein A, Ghosh SS, Avants B, Yeo BTT, Fischl B, Ardekani B, Gee JC, Mann JJ, Parsey RV (2010) Evaluation of volume-based and surface-based brain image registration methods. Neuroimage 51:214–220.

Knickmeyer RC, Gouttard S, Kang C, Evans D, Wilber K, Smith JK, Hamer RM, Lin W, Gerig G, Gilmore JH (2008) A structural MRI study of human brain development from birth to 2 years. 28:12176–12182.

Luck SJ (2014) An Introduction to the Event-Related Potential Technique. MIT Press.

Matsuzawa J, Matsui M, Konishi T, Noguchi K, Gur RC, Bilker W, Miyawaki T (2001) Age-related volumetric changes of brain gray and white matter in healthy infants and children. Cereb Cortex 11:335–342.

McDaniel MA (2005) Big-brained people are smarter: A meta-analysis of the relationship between in vivo brain volume and intelligence. Intelligence 33:337–346.

Moore TM, Reise SP, Gur RE, Hakonarson H, Gur RC (2015) Psychometric properties of the Penn Computerized Neurocognitive Battery. Neuropsychology 29:235.

Narr KL, Woods RP, Thompson PM, Szeszko P, Robinson D, Dimtcheva T, Gurbani M, Toga AW, Bilder RM (2007) Relationships between IQ and Regional Cortical Gray Matter Thickness in Healthy Adults. Cereb Cortex 17:2163–2171.

Nieuwenhuys R (2013) The myeloarchitectonic studies on the human cerebral cortex of the Vogt-Vogt school, and their significance for the interpretation of functional neuroimaging data. Brain structure & function 218:303–352.

Rabinowicz T, Petetot JM-C, Khoury JC, de Courten-Myers GM (2009) Neocortical maturation during adolescence: change in neuronal soma dimension. Brain Cogn 69:328–336.

Raznahan A, Shaw P, Lalonde F, Stockman M, Wallace GL, Greenstein D, Clasen L, Gogtay N, Giedd JN (2011) How does your cortex grow? 31:7174–7177.

Savalia NK, Agres PF, Chan MY, Feczko EJ, Kennedy KM, Wig GS (2016) Motion-

related artifacts in structural brain images revealed with independent estimates of in-scanner head motion. Human brain mapping.

Seeley WW, Crawford RK, Zhou J, Miller BL, Greicius MD (2009) Neurodegenerative diseases target large-scale human brain networks. Neuron 62:42–52.

Shao N, Yang J, Li J, Shang H-F (2014) Voxelwise meta-analysis of gray matter anomalies in progressive supranuclear palsy and Parkinson's disease using anatomic likelihood estimation. Front Hum Neurosci 8:63.

Shaw P, Kabani NJ, Lerch JP, Eckstrand K, Lenroot R, Gogtay N, Greenstein D, Clasen L, Evans A, Rapoport JL, Giedd JN, Wise SP (2008) Neurodevelopmental trajectories of the human cerebral cortex. 28:3586–3594.

Sowell ER, Peterson BS, Thompson PM, Welcome SE, Henkenius AL, Toga AW (2003) Mapping cortical change across the human life span. Nat Neurosci 6:309–315.

Sowell ER, Thompson PM, Toga AW (2004) Mapping changes in the human cortex throughout the span of life. The Neuroscientist 10:372–392.

Stiles J, Jernigan TL (2010) The basics of brain development. Neuropsychol Rev 20:327–348.

Stüber C, Morawski M, Schäfer A, Labadie C, Wähnert M, Leuze C, Streicher M, Barapatre N, Reimann K, Geyer S, Spemann D, Turner R (2014) Myelin and iron concentration in the human brain: A quantitative study of MRI contrast. Neuroimage 93:95–106.

Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, Gee JC (2010) N4ITK: improved N3 bias correction. IEEE Trans Med Imaging 29:1310–1320.

Tustison NJ, Cook PA, Klein A, Song G, Das SR, Duda JT, Kandel BM, van Strien N, Stone JR, Gee JC, Avants BB (2014) Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. 99:166–179.

Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M (2002) Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. 15:273–289.

Van der Laan MJ, Pollard KS (2003) A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap. Journal of Statistical Planning and Inference, 117(2), 275-303.

Wolpert DH (1996) The lack of a priori distinctions between learning algorithms. Neural computation, 8(7), 1341-1390.

Wood SN (2011) Fast stable restricted maximum likelihood and marginal likelihood

estimation of semiparametric generalized linear models. J R Stat Soc Series B Stat Methodol 73:3–36.

Wood SN (2012) On p-values for smooth components of an extended generalized additive model. Biometrika:ass048.

Zhang H, Schneider T, Wheeler-Kingshott CA, Alexander DC (2012) NODDI: practical in vivo neurite orientation dispersion and density imaging of the human brain. Neuroimage 61:1000–1016.

Zielinski BA, Gennatas ED, Zhou J, Seeley WW (2010) Network-level structural covariance in the developing brain. Proc Natl Acad Sci USA 107:18191–18196.