Data-Driven Structural Neuroimaging Metrics to Quantify Aging and Cardiovascular Disease

By

Camilo Bermudez Noguera

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Biomedical Engineering

May 31, 2020

Nashville, Tennessee

Approved:

Bennett A. Landman, Ph.D.

Mark D. Does, Ph.D.

Adam W. Anderson, Ph.D.

Daniel O. Claassen, M.D.

Quinn S. Wells, M.D., Pharm.D.

Mark T. Wallace, Ph.D.

COPYRIGHT

DEDICATION

To my wife Elizabeth, who makes every day the best day of my life.

ACKNOWLEDGMENTS

This work was made possible by a large number of collaborations within Vanderbilt University as well as several institutions around the country. I am grateful to all the investigators, faculty, and students who believed in our work and were always excited to meet and discuss ways to improve on what we were doing. I know many of these collaborations will continue for many years to come. I want to thank all of our funding sources, especially the American Heart Association for believing in our work and granting me a predoctoral training grant to study how chronic heart disease affects brain anatomy.

I want to send special thanks to every member of my Thesis Committee. I am very grateful to have many memories with each one of you where I learned and grew as a scientist and as a person. I want to thank Drs. Claassen and Wells in particular for always reminding me of the clinical impact of my work and encouraging me to find ways to change patient care through practice and through research. A special thanks to Drs. Does, Anderson, and Wallace, for doing exactly the opposite, and reminding me that without engineering rigor and scientific excellence in my work, clinical applications may not go far.

Above all, I want to thank Dr. Landman. He always guided me with kindness and professionalism. He listened my ideas and then encouraged me to find ways to improve them. He was the best mentor for me to learn how to merge my interest in quantitative methods and clinical care through imaging. He has supported me as a physician-scientist at every step of the way, even though we were both figuring it out for the first time. I hope I will continue to learn from you for many years on how to become an excellent scientist but, above all, an excellent person.

Of course, none of this could be done without the friends by your side. Inside the MASI lab, I am thankful for the mentors I had early on like Andrew and Yuankai, as well as the friends who were with me at the finish line like Vish, Cailey, Sam, and Karthik. You all kept graduate school caffeinated and fun. Outside the MASI lab, I was surrounded by some of the best people I know: Caleb, Joe, Jack, Ben, Cody and your respective partners. We've been through this whole journey together and we've seen each other through difficult times and through the happiest times. I hope to always have you as friends by my side. I know I am a better person for it.

Most importantly, I want to share the fruits of this labor with my family. Their support, love, encouragement, and patience made a world of difference. Thanks to my Mom and Dad, who always made the right choices for me, despite sacrifices along the way. Thanks to Julian for your kindness and wise words. I am forever grateful for the family I joined along the way. The Sherrills gave me countless restful days in the sun when I needed them the most. Above all, this work belongs to my wife Elizabeth as much as anyone else. We are in an incredible journey together and I know that together, we can do anything.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER

I. Introduction

1. Overview

The relationship between the brain tissue and human behavior has intrigued people for centuries. Neurosurgical exploration can be traced back to the ancient Incan Empire and the ancient Egyptians, who opened the skull to treat head injuries [3-5]. Consequently, these cultures noticed functional deficits as a result of injuries to the brain or the spinal cord. In ancient Rome, Galen performed neurological dissections and identified the effects of peripheral nerves extending from the brain and spinal cord [5, 6]. It wouldn't be until the Renaissance, when Thomas Willis described the vascular supply to the brain in 1664, that the term neurology was first used [5, 7]. In the following centuries, two different perspectives on neurologic physiology emerged: One approach used discoveries in electricity and magnetism to put forth the idea of a specialized cell called a neuron that can transmit electricity to communicate with other cells far away using long extensions called axons [5]. A bundle of neuronal axons form nerves. This allowed for the notion of an intricate system of communicating cells that form a central nervous system, consisting of nerves in the brain and spinal cord, and a peripheral nervous system, consisting of peripheral nerves that communicate with the rest of the body. However, complex neurologic functions such as memory, speech, planning, and visuo-spatial reasoning could not be easily explained with cell theory alone [5].

The second approach involved a series of clinical observations in patients with lesions to parts of the brain that caused specific defects [5]. Landmark cases like Broca's localization of speech, Fritsch and Hitzig localization of motor movement to the cortex, or even the strange case of Phineas Gage, who survived having an iron bar going through his skull harming his frontal lobe, resulting in a dramatic change in personality [5]. Observations on the effects of lesions continued to provide an insight into complex neurological function well into the 20th century. For example, cases like patient H.M., who was unable to

form new memories after a temporal lobectomy to treat his seizures, illustrated the importance of the hippocampus for memory formation [8]. These cases suggested a structural cause for higher cognitive function and therefore a possible localization of disease [5]. As human longevity increased, an association between age and cognitive changes emerged, raising the question of what structural changes inside the brain may be driving this change in behavior [9, 10]. Capturing the structural changes present in the brain *in vivo* would allow for a further understanding the effects of aging and its associated diseases. Medical imaging would provide the opportunity to look inside a patient's body without resorting to an invasive procedure.

Medical imaging began with the discovery of X-Rays in 1896 by Wilhelm Röntgen [11]. The first clinical application of X-Rays occurred only a few days after Röntgen's publication, where a radiograph was used to identify an industrial needle inside a worker's hand and assist in its removal, making this the first image-guided surgery [12, 13]. Since then, many medical imaging modalities have been developed and implemented in clinical practice, such as computed tomography (CT), ultrasound, or magnetic resonance imaging (MRI). MRI has the advantage over CT and X-Rays in that it can generate 2D and 3D volumes of the brain without using ionizing radiation. Instead, it uses magnetic fields and radio waves to produce high-quality images [1]. By specifying different acquisition parameters, an MR signal can be obtained from different tissue types, thus highlighting soft-tissue contrast [1]. This has particular utility in the brain, which is composed of gray matter, consisting mostly of neuronal bodies, and white matter, consisting of neuronal axon projections. White matter and gray matter, both soft-tissue, have distinct material compositions, which can be seen in MRI [14].

With the advent of neuroimaging, a large amount of quantifiable data was made available to clinicians and researchers alike. The modern field of diagnostic radiology consists of human experts who rely on medical imaging for diagnosis and tracking of disease. However, achieving that level of expertise and precision takes years of training and decades of practice [15]. As an increasing number of patients seek medical care and neuroimaging becomes faster and more accessible, there is a growing need to retrieve the information contained in medical imaging in a time- and cost-efficient manner for both researchers and physicians [16]. The field of medical image analysis seeks to extract quantitative information from medical

imaging with comparable accuracy and reliability as human experts. In an effort to replicate human accuracy, much attention has been dedicated to supervised learning, a subfield of machine learning. In supervised learning, an algorithm is trained to find patterns in the data that best represent a known output or "ground truth" [17]. However, obtaining the "ground truth" requires experts to provide high-quality examples. As algorithms become more complex, more examples are required, which does not scale well with expert human raters [17].

Recent advancements in data retrieval, sharing, and storage has provided access to large amount of contextual information, particularly data derived in the electronic health record (EHR), thus introducing the idea of Big Data to medical image analysis [18, 19]. This is a paradigm shift. In the past, medical image analysis studies consisted of small cohorts of subjects with low variability between them and high-quality outcomes curated by experts [17, 20-25]. Unfortunately, such studies are difficult to generalize to other sites, different populations, or other imaging scanners due to user and hardware variability [22, 24, 25]. Big Data provides the opportunity of deriving meaningful imaging phenotypes from a large sampling of the general population [20]. The goal then is to leverage medical image analysis on a large dataset to assist in higher-level clinical reasoning [22, 23]. Despite the technical limitations of crafting accurate and reliable neuroimaging phenotypes, a key component in medical image analysis lies in describing clinical utility in the metrics derived [22]. Therefore, medical image analysis requires an iterative process of technical validation paired with clinical validation [22, 26, 27]. Clinical validation can be defined as the ability of an imaging biomarker to reliably discriminate or track the progression of disease. If the method does not meet the mark for clinical validation, the technique should to be revisited. As algorithms continue to increase in performance, clinical utility will be a key component of integrating machine learning and medical image analysis in the healthcare workflow.

As mentioned earlier, supervised learning consists of reproducing examples, or labels, provided in an annotated training set. Strong labels include annotations on medical imaging provided by experts, such as medical diagnoses, localization coordinates of lesions or pathologies, and even a manual delineation of a volume of interest (i.e., pathology, anatomical structure, or functional domain). However, strong labels

require reliable visualization by humans and enough resources to produce a ground truth for a given application [24]. More training data may be readily available if weaker labels are used. Weak labels include data that do not necessarily come from experts, but can be accessed in a semi-automatic manner, such demographic information (i.e., age, gender, socio-economic status), disease classification obtained from Electronic Health Record (EHR) billing codes, a bounding box localization, or even the presence of a lesions instead of its location. Importantly, both strong and weak labels have sources of variability. For example, the delineation of a volume of interest may be affected by patient anatomy, bias inherent to the expert, and noise in the raw image. Weak labels also have inherent variability as well, which is sought to be minimized by using a larger sample size and robust models. A key challenge in medical image processing is to develop robust and reliable models that can achieve similar performance with weak and strong labels to solve important clinical problems.

As the percentage of the elderly population in the United States and worldwide is projected to increase, clinical problems associated with aging and its associated diseases have become a priority in order to ensure good health during later years. However, aging is a complex process, characterized by the interplay of physiological, structural, functional, and external factors as well as large interindividual differences [28]. Consequently, understanding the physiological processes of aging as well as its interactions with diseases associated with aging, such as cardiovascular disease (CVD) or dementia, is an active area of research. Much of this research has focused on the link between cognitive performance and structural changes observed, particularly in the brain [28-30]. Recent advances in storing and accessing large amounts of clinical data and its annotations make it possible to study aging directly from patients undergoing clinical interactions as opposed to subjects recruited for research. The benefit of studying aging through clinical data is to measure directly how changes in patient health affect the progression of aging and its associated disease in a large, cross-sectional patient population. However, clinical data can be unstructured, unreliable, and highly variable [31].

New technical approaches are needed to best integrate clinical and imaging information in the context of aging to answer the following open questions: (1) Does information from image intensity alone

provide enough information to predict age, or is there benefit in introducing hand-crafted features to guide algorithm training? (2) Does predicting age from imaging data provide a reliable, normative metric for aging that can be used to detect disease? (3) Can we generalize existing brain segmentation algorithms to be robust to large intensity changes in clinical imaging? (4) Can clinical and imaging data be used to identify the neuroimaging signature of heart failure with reduced ejection fraction in the setting of dementia?

In this dissertation, I present several new methods to address the challenge of using heterogenous annotations on medical data in order to extract meaningful information from neuroimaging to better understand aging and associated disease. For the remainder of this chapter, I will provide a contextual overview of structural imaging in the brain, neuroimaging in aging, and key methods in medical image analysis that have informed and influenced the work presented here. This chapter concludes by defining open problems in the field and the respective contributions provided.

2. Context for Structural Imaging in the Brain

2.1. Magnetic Resonance Imaging (MRI)

Magnetic resonance (MR) imaging is a versatile noninvasive imaging modality capable of providing contrast between soft tissues in the human body [14]. MR scanners use a strong magnetic field to align the nucleus of hydrogen atoms. It is then possible to excite regions within the body using a radio frequency (RF) pulse, which causes hydrogen atoms to tip away and oscillate around the vector of the applied magnetic field [1, 14]. As the protons return to equilibrium alignment with the magnetic field, a radio-frequency signal is emitted. This signal may be sensed and measured to produce an MR signal [1, 14].

Using MRI, protons with hydrogen atoms in different compartments of the body can be excited. The tissue composition in each compartment determines the rate at which the protons return to equilibrium alignment with the applied magnetic field [1]. Therefore, imaging acquisition parameters can be exploited to highlight contrast between different kinds of tissues. In particular, two time-constants are used to describe the return to equilibrium alignment: T_1 and T_2. T_1 is the time constant of the longitudinal relaxation, or the

parallel vector component of the oscillating proton to the external magnetic field. Conversely, T_2 is the time constant for transverse relaxation, or the vector component perpendicular to the applied magnetic field [1, 14]. A set of specific acquisition parameters on the RF excitation, such as Echo Time (TE), Repetition Time (TR), and tip angle are used to highlight contrast between the three main components of the brain: white matter, gray matter, and cerebrospinal fluid (CSF). Consider Equation I.1, which describes the signal equation for a spoiled Gradient Echo sequence, which is a common sequence for T1- and T2-weighted images:

$$S = k \cdot PD \cdot \left(\frac{\sin\alpha \cdot (1 - e^{-TR/T1})}{1 - (\cos\alpha \cdot e^{-TR/T1})}\right) \cdot \left(e^{-TE/T2^*}\right) \quad \text{(I.1)}$$

where PD represents the proton density, α is the flip angle generated from the RF excitation, and k is a constant. To obtain T1-weighted contrast, differences in longitudinal relaxation between tissues are emphasized. A shorter TR increases T1-weighting. TE is short to deemphasize the transverse relaxation effects from $T2^*$. Table I.1 shows the relative T1, T2, and PD values for CSF, white matter, and gray matter. Since white matter has a lower T1 than gray matter, which is in turn lower than CSF, the signal is brighter in the white matter (Figure I.1) [1]. Conversely, T2-weighted images produce contrast that highlights the differences in transverse relaxation. A long TE will maximize the T2 signal while a long TR with a small flip angle will minimize T1-weighting [1]. As shown in Table I.1, CSF has a faster transverse decay (larger T2) than the tissue around it, resulting in stronger signal and high contrast between CSF and brain tissue. However, it is also possible to see contrast between gray matter and white matter due to respective differences in transverse relaxation [1]. Another common imaging modality is fluid-attenuated inversion recovery (FLAIR), which has acquisition parameters similar to those of T2-weighted MRI, except for an initial RF pulse that suppresses the intensity of CSF to improve contrast [14].

Tissue Type	T_1 (ms)	T_2 (ms)	Relative PD
CSF	2,650	280	1.00
Gray Matter	760	77	0.69
White Matter	510	67	0.61

Table I.1 Typical brain tissue parameters as measured at 1.5 T

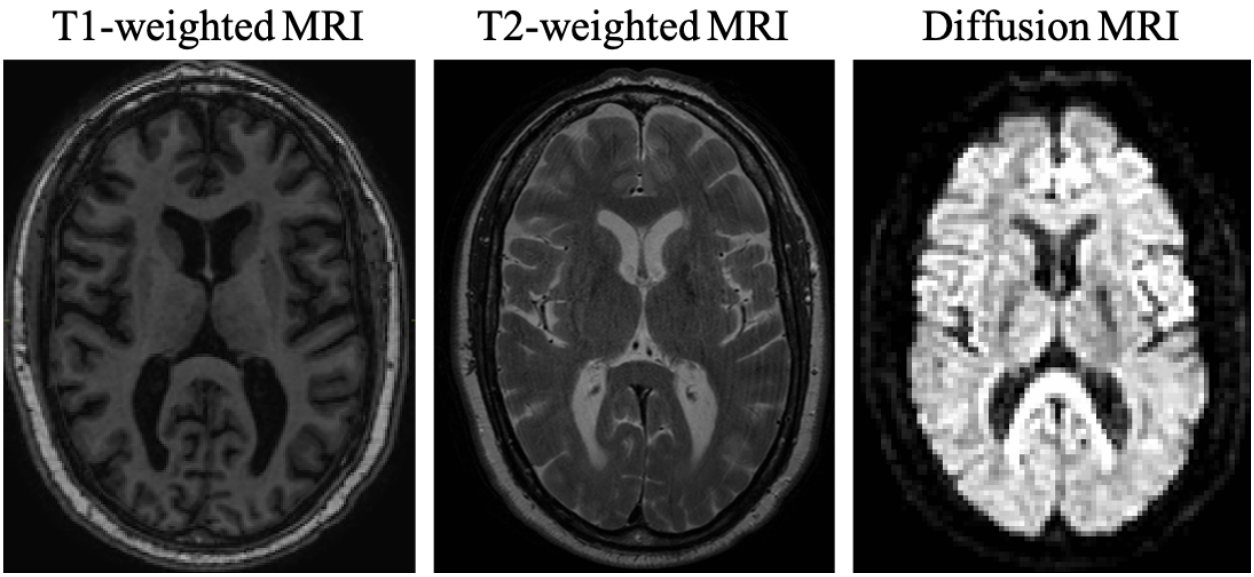

Figure I.1 Three different modalities of brain Magnetic Resonance Imaging (MRI) in a single subject. Each imaging modality highlights contrast between types of tissue in the brain [1].

In addition to the modalities mentioned above, MRI parameters can be manipulated along with magnetization gradients to visualize motion effects within the tissue, such as the diffusion of water molecules. Consider Equation I.2, which describes the diffusion-weighted signal:

$$S = S_0(e^{-bD}) \text{ where } b = (\gamma G\delta)^2(\Delta - \frac{\delta}{3}) \quad \text{(I.2)}$$

S_0 is the MR signal without diffusion weighting and D is the diffusion coefficient intrinsic to the tissue. The exponential terms is analogous to the T2-weighting term in Equation I.1 [32]. The b-value is an acquisition parameter that describes the sensitization of diffusion in one gradient direction, with values for the gradient pulse magnitude (G), duration (δ), and time interval (Δ), while γ is the gyromagnetic ratio. Each compartment in the brain will have a distinct diffusion signature and therefore different contrast between them [14, 33]. It is expected that CSF will show the largest diffusion in all directions, since it consists of largely unrestricted water molecules, thus showing low signal (Figure I.1). White matter is believed to have diffusion restricted to the direction perpendicular to white matter tracts. Therefore, a diffusion signal will be high parallel to a gradient applied in the same direction, whereas it will be low in the perpendicular direction [33]. Gray matter tissue has largely isotropic tissue, so diffusion of water is

uniformly restricted by the extracellular space, thus resulting is a low signal [33]. Diffusion motion is particularly sensitive to pathologies, such as the presence of inflammation or blood, which can be visualized using the magnetization signal, often called the diffusion-weighted image, or as apparent diffusion coefficient (ADC) map, which is a measure of the magnitude of diffusion in each voxel [14].

Diffusion Tensor Imaging (DTI) is a technique that enables the measurement of the restricted diffusion of water in 3D space using a large number of gradient directions. By measuring the diffusion signal in several directions, it is possible to quantify the directionality of diffusion, which reflects the underlying tissue architecture [34]. The displacement of diffusion in 3D space is described by a diffusion tensor, which is a 3D matrix describing the covariance along each direction (Equation I.3).

$$D = \begin{bmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{yx} & D_{yy} & D_{yz} \\ D_{zx} & D_{zy} & D_{zz} \end{bmatrix} \tag{I.3}$$

Diagonalization of the diffusion tensor will yield eigenvalues $(\lambda_1, \lambda_2, \lambda_3)$ and corresponding eigenvectors (e_1, e_2, e_3), which describe the direction and diffusivity along the axes of principal diffusion [35]. If the eigenvalues are equal, then diffusion is considered isotropic, as seen in gray matter, and if eigenvalues are largely different, then the diffusion is considered anisotropic, suggesting an organized microarchitecture, such as the white matter [35]. It is possible to represent these data as diffusion metrics for each voxel, such as mean diffusivity (MD) across all directions (Equation I.4), or the fractional anisotropy (FA) in each voxel (Equation I.5) [35].

$$MD = {(\lambda_1 + \lambda_2 + \lambda_3)}/{3} \tag{I.4}$$

$$FA = \sqrt{\frac{(\lambda_1 - MD)^2 + (\lambda_2 - MD)^2 + (\lambda_3 - MD)^2}{2({\lambda_1}^2 + {\lambda_2}^2 + {\lambda_3}^2)}} \tag{I.5}$$

These metrics will highlight contrast between different kinds of tissue architecture and even provide a general sense for the directionality of white matter tracts. This can be very useful to find structural connections between non-adjacent gray matter regions [34].

2.2. Computed Tomography (CT)

Another common imaging neuroimaging modality is computed tomography (CT). CT uses x-rays emitted in a fan formation and allowed to pass through a patient's tissue and absorbed before being detected and measured [1]. The amount of x-ray energy received by the detector is inversely proportional to the amount of energy absorbed by the tissue. Thus, each fan beam generates a projection of the tissue [1]. The fan beam is rotated around the patient, and a high-resolution rendering of the internal organs can be reconstructed from the set of projections [1]. CT is used to visualize contrast between soft-tissue and hard tissue, such as bones. Contrast is achieved by the relative absorption of X-ray energy by different tissues. For example, bones will absorb more energy than the lungs, so bones will appear bright while lungs, which are filled with air, will appear dark. Due to its high resolution and low acquisition time, CT has many clinical applications in neuroimaging, such as detecting acute bleeding or infarct inside the skull, fractures, and vascular malformations [1].

2.3. Challenges

There are still many limitations with neuroimaging in clinical and research applications. Like all imaging modalities, both MRI and CT are affected by issues such as noise, artifact, and resolution limitations [1]. Research scans are usually high-resolution, with a low signal-to-noise ratio since these are acquired with a study in mind and relaxed constraints like acquisition time. Therefore, an image with high resolution and reduced artifact could be obtained in the research setting with enough resources and time. It was already described above that such studies also require large amount of resources and may not generalize well to images under different circumstances. Conversely, clinical data is affected by motion artifact, low resolution due to limited acquisition times, and a prior probability of disease that may affect the utility of the scan for research purposes. However, large amounts of data can be obtained after parsing the entire medical record. It is therefore an important challenge to develop medical image analysis that are robust and generalizable across both domains of clinical and research neuroimaging.

3. Context for Medical Image Analysis in Neuroimaging

Due to the increasing amount of imaging data produced by medical centers and research laboratories, there is an increasing demand for extracting quantitative and actionable information that can enhance and inform a clinical pipeline, whether it be in diagnosis, lesion detection, or normative biomarkers [22, 24, 36]. At Vanderbilt University Medical Center alone, the electronic medical record shows there are over 100,000 MRI of the brain and almost 200,000 head CT. The field of medical image analysis has adapted to this need, influenced by two key trends: (1) the development of robust models able to handle big data representative of a large population, and (2) the development of deep learning, which does not require hand-crafted, expert features in order to perform machine learning on imaging datasets [17, 20, 37, 38]. Automated medical image analysis promises the opportunity of timely, efficient, and robust imaging biomarkers to assist a clinical workflow [18, 19]. Previous work in this field has focused on smaller datasets, so new challenges will arise as the field strives to provide high-quality biomarkers on larger datasets, to solve more difficult clinical problems, and find meaningful signal in weak annotations [18-20].

3.1. Machine Learning on Medical Images

Large-scale structural image analysis seeks to automate the process of extracting valuable, quantifiable information from an image and present it in the context of a larger population or disease cohort. Historically, this is done by a human expert who is able to extract meaningful features from the image, either implicitly or directly by annotating the image, and associating these findings with a broader context relevant to the patient or the research study, such as age, pathologies, or normal variants [17]. However, with increasing demand of fast and reliable analysis of an even larger supply of medical imaging, many have strived to use machine learning in order to provide these services to scale [25]. Consequently, medical image analysis consists of two key steps: image processing, where important features are extracted from

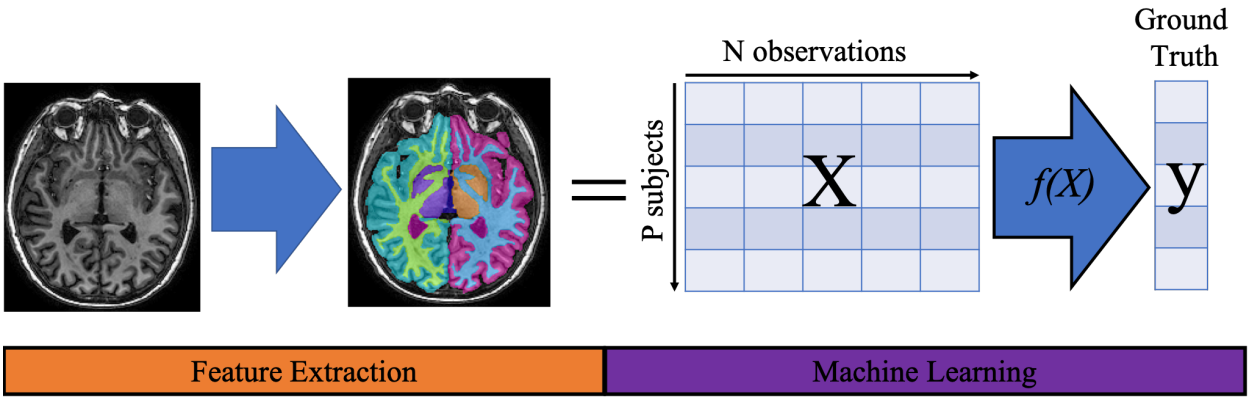

Figure I.2 Machine learning framework for medical image analysis. Medical data is acquired and processed to extract features. Features can be organized as the matrix X, which includes N observations from P subjects. The goal of machine learning is to find the best function *f* that best maps X to y.

the raw image, and image analysis, where these features are placed in the context of a population or statistical model.

Machine learning is a subfield of computer science and artificial intelligence, where algorithms are used to fit statistical and mathematical models to a set of sample data called training data in order to understand a desired outcome [25, 39]. In other words, given a set of *N* observations from *P* subjects, herein represented as *X,* we wish to find the function *f* that best maps the input data into desired output *y,* also referred to as "ground truth" or "labels". The goal is then to apply the function learned from the training data to a separate dataset called the testing set in order to assess accuracy and generalizability (Figure I.2). Broadly, machine learning applications to medical imaging can be summarized into two main tasks: regression and classification [39]. In regression, the outcome is a continuous variable, while the input data again is a set of *N* observations from *P* subjects. Regardless of what model is chosen to represent the function *f,* most algorithms seek to minimize the error in prediction, which is the difference between *y* and *f(X)*. In a classification problem, the output *y* is not continuous, but discrete, and the algorithm seeks to find the function *g* that best places the data *X* into its respective categories *g(X)* while minimizing the error of misclassification.

Machine learning is well-suited for medical image analysis for several reasons. Not only does it perform its analyses in a timely manner, but it can represent highly nonlinear relationships between input

data and outcomes that are difficult for humans to compose intuitively [39]. A main drawback of machine learning, up to recent years, was its reliability on the quality of the input data and its annotations [17, 22, 24, 25]. In other words, machine requires an initial preprocessing step to convert the raw into the organized format of X with N observations from P subjects. In order to organize the input data, a set of features would have to be extracted. These features also require an expert annotation. For instance, consider the task of modelling the volume of different brain regions across an individual's lifetime, as done by Huo et al [40, 41]. In this case, the raw data consists of a series of brain MRI and the corresponding age. However, we do not know the volume of the brain regions or how to draw the brain regions on the brain MRI. Therefore, an expert would have to manually trace the regions of interest on all subjects or there would need to be a trustworthy automatic segmentation algorithm that can provide the delineations. From each region of interest, the volume can be computed and thus a regression model can be crafted to explain how volume changes with age. The first step of extracting the volume delineations, or "features," is the image processing step and, until the recent development of deep learning, was a key element of medical image analysis.

3.2. The Benefits and Pitfalls of Deep Learning

Deep learning, or deep neural networks, is a subfield of machine learning developed in the 1980's, where the goal is to learn the relationship between the input data and a desired output. These algorithms were designed to emulate neuronal connections in the brain by linking a series of linear activation functions, or "neurons," in order to learn complex relationships in the data [42, 43]. A deep neural network consists of a series of "layers," which, in turn, are a weighted linear combination of the elements of a previous layer. By stacking several layers deep and adding an activation criteria to each element, it is possible to approximate any nonlinear function, given that the model and training data are large enough [42]. This made it possible to model increasingly complex relationships in the data [17, 42]. In the 1990's neural networks were largely forgotten due to hardware restrains: the models were too large and too complex to solve and test in a time-efficient manner. Deep networks re-emerged in the late 2000's when Graphics

Processing Units (GPUs) were adapted to training deep neural networks, resulting in algorithms training 10 to 20 times faster with current hardware [43, 44].

The meaningful contribution of deep learning to medical image analysis came with the advent of convolutional neural networks to solve image-based classification problems with networks such as LeNet [45] and AlexNet [46]. The key contribution here was that instead of learning the weight of individual nodes in each layer, the network will learn the kernel of a convolution operation. Through a series of convolution operations, it is possible to automatically extract information from the image that best predicts the outcome [17, 38, 43, 47]. By doing so, deep learning effectively merged several steps of image processing and image analysis in a "end-to-end" framework. This was a paradigm shift in medical image analysis, since there is no longer the need to craft and extract features from raw images in order to train an algorithm (cf. Figure I.2). Useful features can be automatically learned from raw images without having an expert rater provide strong labels. The powerful feature-extraction of deep learning, along with weak labels obtained from biomedical informatics, opens the possibility to understand complex relationships between healthcare data and patient outcomes. Due to the many arrangements of elements in a deep convolutional neural networks, there are many possible network configurations to solve a problem in medical image analysis. In this proposal, we intentionally limit the scope of network architecture by using three common networks as general purpose functions to solve challenges in medical imaging: A fully-connected convolutional network based on LeNet [45], the U-Net [48], and the ResNet [49]. More detail on each of these networks is provided below.

Like any other machine learning algorithm, deep learning has many limitations. First, deep learning models have a tendency to overfit if the size of the training data is too small [17, 38, 42]. Overfitting is a phenomenon in machine learning where the model is learning the training data too well [39]. In other words, the real signal in the training data is too weak due to a small sample size, so the model will capture individual variability, or noise, as a representation of the signal. Consequently, the model will not generalize well when presented with the testing set, and therefore the applicability of this model to future test-cases is reduced. Second, deep learning models have been shown to be unstable. Small perturbations in the input

data can cause severe deviations in prediction [50]. This is an important drawback, as machine learning and deep learning algorithms are implemented in life-or-death situations such as self-driving cars or healthcare decision making. Lastly, deep learning models are difficult to interpret compared to linear models [51-54]. Due to the many connections between elements in each layer, it is often very difficult to provide inference on what features are extracted from a raw image and how each contributes to the final prediction. Previous attempts have used attention maps [55], saliency maps[56], sensitivity maps [57], or direct visualization of the convolutional kernels [46], but this area remains an open problem in the field [51-53].

4. Neuroimaging Medical Image Analysis on Aging and Associated Diseases

It is estimated that by 2050, the United Stated population over the age of 65 will reach 83.7 million (20.9%), almost double that in 2012 at 43.1 million (13.7%) [58]. Additionally, the proportion of Americans aged over 65 who reported having one or more chronic disease increased from 86.9% to 92.2% between 1998 and 2008 [59]. This trend is believed to continue given the increased longevity and increasing obesity [59]. Diseases associated with the aging population include cardiovascular disease, diabetes, cerebrovascular disease, hypertension, dementia, all of which are systemic diseases with a wide array of risk factors, complex physiological changes, and a need for well-orchestrated care [28, 59]. Studying clinical data in the elderly is imperative in order to improve our understanding of the interplay between these diseases as well as provide evidence-based and cost-effective care to all.

One active area of research has been to study the effect of aging on the brain through the lens of neuroimaging. Several longitudinal research studies such as the Baltimore Longitudinal Study of Aging (BLSA), the Framingham Heart Study group, or the Mayo Clinic Study of Aging have been following subjects for decades to study how aging and associated diseases affect neuroanatomy, with the goal of describing the relationship between longitudinal disease cognitive function [60-65]. Several studies have identified changes in cortical thickness, dilation of sulcal and ventricular spaces, and white matter hyperintensities to be associated with aging independent of disease [30]. These findings help discriminate between normative brain atrophy and degeneration associated with aging compared to a pathological

process [66, 67]. However, the effect of disease interactions on neuroanatomy have not yet been studies in large clinical imaging cohorts. Medical image analysis is well suited to study aging because it provides quantitative, actionable data to better understand the diagnosis and progression of disease. Medical imaging along with healthcare data found in the electronic health record (EHR) provide a non-invasive alternative to measure and track the changes associated with aging and associated diseases directly from the patient population [67].

5. Open Problems

The application of deep learning to medical image analysis has allowed the field to investigate many unanswered questions about the nature of medical imaging and its applicability to clinical practice. In this dissertation we explore the following open problems in medical image analysis when applied to strong and weak labels alike when obtained from a clinical setting.

5.1. Brain Age Estimation

Aging is a complex, multifactorial process by which the human body changes over time. Aging is not only a universal biological phenomenon but an independent risk factor for many human diseases [68]. Understanding how aging affects individuals as well as its effect on specific diseases has incredible therapeutic potential. This has motivated research efforts towards deriving an "aging biomarker" that is predictive of disease risk and residual lifespan [68]. One such strategy has been to predict age from healthcare data, such as brain imaging and imaging-derived metrics [69-71]. The idea of age prediction from healthcare data aims at finding a normative biomarker for aging and disease. Complex data such as raw imaging may contain information encoded in it that is difficult to extract with imaging metrics alone. Moreover, imaging might have information hinting at the onset of disease that is not yet visible through signs and symptoms. Brain age provides the opportunity to place the appearance of an individual's brain in the context of thousands of others in order to assess whether it is within the norm. An advanced brain age has been previously associated with Alzheimer's Disease [72-76], Schizophrenia [77-79], traumatic brain

injury[80], amongst others [68]. Deep learning provides the opportunity to accurately learn a weak label, such as age, from thousands of subjects with minimal preprocessing, thus making a large-scale biomarker feasible. Thus, predicted age can act as a non-invasive biomarker of aging as well as a biomarker for disease. Such tools can be used to study the effects of neurodegenerative and mental health illnesses on brain aging.

Previous studies that predict age have used classical image processing methods to extract features followed by a regression model in order to predict age including kernel methods [72], Markov models [81], and Bayesian models [82]. More recently, deep learning models have used the raw image to predict age with great success [83]. Despite the success of deep learning in achieving similar or better performance at this task, it is still unclear whether deep learning captures the same information available from expert features. In other words, would deep learning benefit from combining expert features, usually rich in anatomical context, along with the features learned automatically? Does adding anatomical context in the form of expert features make age estimation more stable against perturbations to the input image? In this dissertation, we describe a combined age estimator using deep learning and expert features that not only improves on previous accuracy but is more robust against adversarial perturbations.

The following open problems were identified for this dissertation:

- Brain age estimation using expert features as well as raw MRI has been previously described by several groups. However, it is not clear whether these two sources of data contain different information such that integrating dual context would result in improved accuracy.
- Age prediction has been extensively studied in brain MRI, but it has not yet been that age can be predicted from other imaging modalities, such as orbital CT, or whether differences in predicted age from orbital CT can discriminate between diseases of the eye.
- It is not yet known to what degree deep learning tools on medical imaging tools are susceptible to adversarial attacks or the levels of adversarial noise necessary to achieve it.

- While it is well known that deep learning networks are unstable, it is unclear whether including expert features will result in a more robust network against adversarial attacks due to enforced anatomical context.
- Different image-based algorithms have been proposed to compute age. However, many of these methods lack clinical validation via testing differences in brain age gap between disease groups and healthy controls. For instance, it is not known if brain age gap changes in subjects with depression, particularly in the elderly population.
- Similarly, it is now known whether age prediction metrics correlate with changes in cognitive performance in healthy controls or in patients with depression.

5.2. Segmentation of Brain Structures Across Imaging Modalities

Segmentation has constantly been a common application in image processing. In many ways, segmentation can be thought of as a pixel-wise classification problem. The goal of segmentation in a supervised learning framework is to use previously annotated medical images to automatically generate annotations on new samples. This annotation consists of a manual delineation of a region of interest. In the brain, such regions include functional regions, anatomical divisions of brain tissue, or detection of lesions and pathologies. Automatic segmentation can provide additional clinical information in volume estimation, lesion detection, and surgical planning; as well as research support, with region seeding for tractography and functional studies and feature extraction. Due to small variability in the anatomy between subjects, generalizability across subjects is a key challenge in segmentation.

In the past, segmentation was achieved with considerable success using deformable registration of labelled atlases. A labelled atlas consists of an image-delineation pair, where the delineation is often generated by an expert rater and considered the ground truth. Initial success was achieved using a single-atlas in some applications [84-86], but a boost in performance was achieved when multiple atlases were registered to the test subject and the labels were fused according to the likelihood of correspondence [87, 88]. Segmentation was revolutionized with the advent of deep convolutional neural networks, particularly

the U-Net [48] and ResNet [49], which showed a significant increase in accuracy across many different modalities. A particular advantage of deep learning models over multi-atlas segmentation is that once an adequate model is trained, generating a segmentation takes a matter of seconds, while multi-atlas segmentation is on the scale of hours to days [89]. However, deep learning models trained on small datasets are difficult to generalize across different contrasts, making it difficult to share trained models across applications that use different imaging modalities or different anatomical locations [17].

Transferring information from one imaging modality to another has great value in medical imaging, since different contrasts may highlight important structures. For example, the dentate nucleus (DN) is a small gray matter structure deep in the cerebellum, which is involved in motor coordination, sensory input integration, executive planning, language, and visuospatial function [90]. The DN and its associated white matter tracts have been implicated in the development of movement disorders, such as PD or essential tremor (ET) [91-95]. Previous studies have used a delineation of the DN for imaging-based studies of connectivity and volumetry to detect structural differences [96-98]. An accurate and reliable segmentation of the DN would allow to perform these studies on a large-scale.

A major challenge in DN segmentation is its generalization across modalities. Many clinical and research MRI uses T1-weighted sequences, in which the DN is difficult to visualize [90]. However it is clearly visible using T2-weighted MRI or DWI. Learning a segmentation encoded in DWI and accurately applying it on T1-weighted MRI would enable the application of such a segmentation across multiple studies [2, 99, 100]. In this dissertation, we propose a segmentation method of the DN learned from strong labels such as manual delineations on DWI and applied to T1-weighted imaging. We assess the reproducibility of the automatic segmentation by showing that volumetric differences between groups are preserved between the manual and automatic segmentations.

The challenges in harmonizing multi-site and multi-sequence imaging datasets for segmentation have been described previously, recognizing that the variability in tissue contrast can come from scanner factors (number and sensitivity of head coils, imaging gradients, magnetic field homogeneity, or scanner noise) or software factors (image reconstruction methods or software updates) [101]. Many of the efforts

directed at image harmonization have sought to find a mapping of tissue contrast intensities to account for small changes in acquisition parameters [101-104], while recognizing that these methods may not be robust to large changes [101, 103], such as those found in clinical data like the presence of intravenous contrast or large variations in flip angle. These large variations result in a different representation of the underlying anatomy and harmonizing across them requires a robust technique that does not only map tissue intensities. Improving algorithm generalizability on clinical data will allow for larger learning datasets and translation to the bedside.

The following open problems were identified for this dissertation:

- An automatic segmentation of the DN has previously only been done using single-atlas methods. Deep learning methods have shown improved technical accuracy and generalizability over single atlas segmentations in other brain structures. There is yet no consistent segmentation of the DN.
- Deep learning can be susceptible to irregularities in the data. However, with an increasing need to translate modern algorithms to the bedside, these should be made robust to heterogeneities in the data. It is still an open problem on how to perform domain adaptation from a research domain to heterogenous clinical data, such as different acquisition parameters or the presence of intravenous contrast. Furthermore, it is still unknown how improvement in performance for local factors may result in a degradation of performance in the whole brain segmentation task on the original dataset.
- In order to achieve domain adaptation towards heterogenous data, new training examples are needed to refine the algorithm via transfer learning. However, generation of manual tracings by experts can be time- and resource-prohibitive. There are large amounts of unlabeled research and clinical imaging data. It is still an open problem as to how to leverage these data to improve local performance using weak labels in a supervised or semi-supervised learning framework.

- Although performance of whole brain segmentation is known to drop between imaging domains, it is not known how bias and variance of volume measurements from segmentations change or if this can be minimized with transfer learning.

5.3. Neuroimaging Effects of Heart Failure in the Setting of Dementia

Alzheimer's Disease and other dementias affect more than 5.3 million Americans and result in over $259 billion in medical expenditures annually [105]. Extensive epidemiologic evidence suggests that HF not only shares many risk factors with dementia, but is itself an important risk factor for dementia [106-109]. Other studies have shown that there are clinical variables in HF associated with a higher risk of dementia, such as hypotension, diabetes, hyperglycemia, anemia, and electrolyte levels [108, 110-112]. However, the clinical comorbidities influencing development of HF-associated dementia have not been studied in a phenome-wide, unbiased fashion; therefore, it is unclear whether HF confers risk independently, in combination with, or through interactions with other comorbidities. Similarly, patients with HF have specific changes in neuropsychological testing and brain morphology, such as atrophy in the parahippocampal gyrus, compared to controls [109, 113, 114]. Epidemiologically, the proportion of vascular dementia to Alzheimer's disease increases in patients with HF [112, 115]. This evidence suggests that HF has a distinct effect on brain morphology, which may result in a different dementia phenotype. An improved understanding of the clinical and imaging predictors of dementia in the setting of HF may help characterize a pathophysiologic mechanism and lead to improved diagnostic and therapeutic strategies.

Recent advances in clinical informatics have allowed investigators to introduce tools such as phenome-wide association (PheWAS) as systematic and efficient methods to leverage EHR phenotype data for research [116, 117]. The PheWAS method uses a validated, curated medical phenome that hierarchically groups the ~18,000 International Classification of Disease (ICD) billing codes into about 1,800 phenotypes, each with defined control groups. The PheWAS phenome compares favorably to other methods such as the Agency for Healthcare Research and Quality's Clinical Classifications Software (CCS), with PheWAS codes representing over twice as many common problems as the CCS. PheWAS phenotypes have been used

successfully to replicate known genetic and clinical associations [118]. These methods are flexible and have been applied to discover patterns in phenotype clustering and to show that patterns of comorbidity associations can provide important insights regarding etiologies of complex diseases [116, 119, 120]. PheWAS can be a powerful tool to identify new, modifiable clinical phenomena in the setting of HF.

The field of neuroimaging has been especially active in the search for imaging disease biomarkers and the brain regions contributing to the signal. These efforts demonstrate associations between cardiovascular disease and imaging features associated with cognitive and brain function. Beason-Held et. al. showed that a high Framingham Cardiovascular Disease Risk Profile (FCRP) score, an aggregate score of cardiovascular risk factors, is associated with longitudinal decline of cerebral blood flow in several regions grey matter of the brain [121]. Similarly, Romero et. al. showed that well-known cardiovascular risk factors such as age, sex, and hypertension influence the amount of cerebral microbleeds seen in T2 MRI [122, 123]. Moore et al. showed that left ventricular mass index, a precursor of left ventricular hypertrophy and HF, is associated with changes in white matter microstructure [124]. These studies show that changes in cardiovascular health have structural sequelae in the brain, making neuroimaging a promising tool for understanding disease as well as early diagnosis and prevention. However, these studies are highly dependent on engineered structural features and pre-selected regions of interest (ROI) – a limitation that can be overcome by integrating high-quality crafted features with the data-driven features obtained from deep learning.

- Recruiting elderly patients with several comorbidities for imaging research studies can be costly and time-consuming. It is an open problem whether retrospective clinical data can be leveraged to extract high-quality imaging and clinical variables relevant for a small clinical cohort, such as patients with comorbid heart failure and dementia.
- Extracting clinical data allows for discovery of clinical comorbidities associated with heart failure. It is not yet known if these associations may lead to novel links between clinical comorbidities to advise and inform future targeted studies.

- Although it is known that heart failure shares neuroimaging features with aging and dementia, a heart failure neuroimaging signature has not been previously described. Furthermore, it is unknown if this signature can be observed from a heterogenous clinical imaging dataset.
- It is unknown if the discovery of an imaging signature can inform future imaging studies targeted at specific brain regions under a research setting or how these morphological changes affect cognition.

6. Contributions

Medical image analysis is undergoing a transformation due to the unprecedented amount of data available, as well as new, complex models suited to interpret it. With the interpretation of Big Data in medical imaging, new challenges will arise as many of these algorithms are translated into clinical practice. It is often straightforward to show technical validation in medical image analysis by producing a small error on a withheld dataset. However, clinical validation will be an essential step in making these algorithms available for clinical practice.

6.1. Brain Age Estimation

- We trained a deep learning network on a large cohort to predict age using both raw MRI and volumetric features to supply additional anatomical context, which shows improvement over each respective input alone.
- We trained a deep neural network on a large cohort of clinically acquired orbital CT to predict age using the raw volume and volumetric features, again showing improvement against separate feature sources.
- We show how some diseases of the orbit resulted in changes in the orbital CT age gap, suggesting this as a possible non-invasive biomarker of eye disease.

- We present an adversarial model to measure the effect of adversarial noise levels on MRI brain age predictions and show that while deep learning networks are largely susceptible to adversarial noise, the inclusion of volumetric features largely diminishes this effect.
- We show the clinical applicability of our MRI brain age estimator in finding differences in the Brain Age Gap between patients with depression and controls in an adult and a geriatric population. Furthermore, we show associations between Brain Age Gap and cognitive performance in the geriatric cohort.

6.2. Segmentation of Brain Structures on Across Imaging Modalities

- We generated an automatic segmentation of the DN on multiple imaging modalities learned from manual tracings done on diffusion imaging. We showed that a diffusion-based segmentation of the DN can reliably reproduce manual labels, but differs greatly from segmentations achieved by single-atlas registration.
- We generated augmented labels for unlabeled brain MRI to perform transfer learning. We explored the use of semi-automated labels via manual correction of labels obtained from multi-atlas segmentation in a pediatric dataset. We also used paired, unlabeled clinical brain MRI to generate automated labels on a post-contrast image using the co-registered pair without contrast.
- We performed transfer learning from the pretrained whole brain segmentation tool SLANT towards heterogenous clinical datasets, such as multiple acquisition parameters and a pediatric sample. These efforts will improve on the generalizability of image processing tools in order to translate to the bedside.
- We showed the importance of domain adaptation for volumetric estimation of brain regions of interest. Specifically, we estimated the error in volume estimation of the hippocampus between imaging pairs of the same subject and showed a drop in error compared to non-harmonized methods such as baseline SLANT or the FreeSurfer segmentation pipeline.

6.3. Neuroimaging Effects of Heart Failure in the Setting of Dementia

- We completed cohort definition and data extraction from the VUMC EHR using ICD-9 and ICD-10 codes for dementia and the presence of heart failure. We further filtered to include only subjects who have at least one head MRI documented in the clinical imaging database ImageVU and evidence of some form of cognitive testing on their medical chart supporting the diagnosis of dementia. We have identified 5,913 subjects who meet all of these inclusion criteria for dementia and brain imaging.
- We completed visual quality control on all brain MRI and identified high-quality imaging T1 and T2 FLAIR brain MRI. We found over 400 subjects with high-resolution, high-quality imaging appropriate for medical image processing. We further analyzed how the clinical profile according to EHR codes changes between patients with and without high resolution imaging.
- We completed automatic segmentation on all T1-weighted MRI using the preciously developed tools on heterogenous clinical data. We described the neuroimaging signature of heart failure with preserved ejection fraction.in this population.

7. Previous Grants & Publications

The contributions in this dissertation have been previously published or are currently under review. The brain age estimator as well as the adversarial were published at *Magnetic Resonance Imaging [125]* and *Neurocomputing* [126], respectively. Clinical validation of brain age gap on a clinical cohort is under review at *Translational Psychiatry.* Our tool for multimodal segmentation of the dentate nucleus was published in the *Journal of Medical Imaging [127].* The initial work on domain adaptation of whole brain segmentation was originally published the conference proceedings for *SPIE: Medical Imaging 2020 [128]* and a full journal version is currently under review at *Journal of Medical Imaging.* The work focused on translating these tools to describe the clinical and imaging features of heart failure was funded by the *American Heart Association Predoctoral Fellowship*. Consequent work on the neuroimaging signature of heart failure in dementia is currently under review at the *Journal of the American heart Association.*

8. Conclusions

The rest of this document is as follows. In Chapter II, we will introduce the literature on brain age estimation and show that deep learning age estimation can result in improved accuracy when expert features are introduced in training in two different imaging modalities. Chapter III is a continuation of this work, which further shows that a model with additional anatomical context is less susceptible to adversarial noise. In Chapter IV, we present the application of predicted brain age to discriminate between patients with depressions in adults and a geriatric population, as well as the relationship between predicted age and cognitive performance. We then focus on deep learning methods of segmentation. Chapter V presents an accurate deep learning segmentation of the DN learned across different modalities of MRI. We show that volumetric differences between groups are preserved when using the automatic segmentation. In Chapter VI we propose a new approach to generalize whole brain segmentation to heterogenous datasets with augmented labels, and in Chapter VII we take this a step further to validate the performance of transfer learning on clinical data for variable acquisition parameters. Chapter VIII then brings these tools together in the description of a neuroimaging signature of heart failure using clinical data. Chapter IX outlines future research directions focused on using clinical data from imaging and medical charts to better understand aging and associated disease. The novelty in this dissertation lies in the integration of high-quality research tools across technical domains and implementing them to heterogenous datasets in order to better understand the effects of aging and associated diseases through neuroimaging.

II. Anatomical Context Improves Deep Learning on the Brain Age Estimation Task

1. Introduction

In a recent special issue, Greenspan et al. reviewed the role of deep learning in medical image analysis, concluding that "In the majority of works presented, use of a deep network is shown to improve over the state-of-the-art. As these improvements seem to be consistent across a large variety of domains, and as is usually the case, development of a deep learning solution is found to be relatively straight-forward, we can view this as a major step forward in the medical computing field." [17] The authors state that networks excel at tasks of lesion detection, segmentation, registration, and predictive models. A key challenge of deep learning applied to medical imaging in a supervised learning framework, is the need for large training sets with high-quality, expertly labelled features. Generating high-quality labels for medical images, in tasks such as detection, diagnosis, or segmentation, requires expert knowledge which does not scale well to the large number of training examples needed for a robust deep learning algorithm. The authors identify a second key issue with the shifting paradigm towards deep learning: "Can we rely on learned features alone or may we combine them with handcrafted features for the task?"[17].

Advances in deep learning have provided an approach for learning a highly non-linear function of a dataset when an appropriate kernel or feature manifold is not known. Historically, feature extraction in the field of computer vision has relied on automatically detecting intensity patterns or textures in an image. However, recent work on deep convolutional neural networks has shown that an adaptive learning of image features through convolutional filters can results in more accurate results. Unlike many datasets used in computer vision, medical imaging is unique in that there is already extensive a priori expert human knowledge associated with the image, such parcellation of tissues into anatomical or functional units. This

knowledge can be formulated as imaging features grounded in medicine and physiology, which can direct computer vision tasks to find better relationships in the data. Decades of work on medical image processing have focused on engineering and refining meaningful features to capture the dimensionality of a small imaging dataset. While deep learning has shown remarkable improvements, it has not been shown whether engineered features are redundant. For example, age prediction from medical imaging is a task that has relied heavily on pre-processed regions of interest (ROIs) from T1w structural brain MRI using the volumes of ROIs such as the white matter, ventricles, and the cortex [69-71]. Recently, these measures have become efficiently extractable from a standard T1-weighted (T1w) brain MRI using a multi-atlas segmentation approach [69].

Recently, Cole and Franke evaluated the clinical utility of predicting brain age from hand-crafted imaging features and showed that predicted brain age can be used to better understand differences between individuals during the aging process, understand disease processes, and design treatment strategies [70, 129]. The authors pose that the absolute difference between predicted age and chronological age, herein called Brain Age Gap (BAG) biomarker, is a valuable imaging metric, since it has been shown to correlate with aging as well as neurodegenerative diseases [129]. A recent study showed a correlation between BAG and mortality in subjects over 73 years old, attributing an increase of 6% to mortality risk for every year predicted older [130]. BAG also correlated with common metrics of age such as decreased grip strength, decreased expiratory volume, and slower walking time [130]. Moreover, an increased BAG has been shown to correlate with several neurodegenerative diseases like Alzheimer's disease, bipolar disorder, diabetes, Down syndrome, epilepsy, major depression, mild cognitive impairment, traumatic brain injury, and schizophrenia [70, 129]. These clinical correlations suggest common secondary effects on the brain, such as inflammation or oxidative stress [129]. Some of the challenges with this technology moving forward include incorporating multimodal data, such as orbital computed tomography (CT) to produce new

biomarkers like orbital brain age gap (OrbitBAG), and the use of deep learning, which the authors identify as beneficial due to the "removal [of] the reliance on data pre-processing to extract meaningful features." [129] Predicting age from neuroimaging may provide a new, noninvasive biomarker of aging as well as a discovery tool for positive and negative effectors on aging.

Some of the best results that predict age report a BAG between 4-5 years [72], but these models have been difficult to generalize due to small sample size [72, 82, 131], limited age-range [81, 132], or extensive multimodal data requirements [70, 132, 133]. Cole et. al proposed a convolutional deep neural network technique on raw T1w MRI images which showed an mean absolute error of 4.65 years in 2,001 healthy adults ages 18 to 90 years [83]. This method showed competitive results in age prediction without any a priori feature extraction or image preprocessing. The work by Cole [83] exemplifies the improved performance of deep networks over handcrafted features. However, it does not explore how using both types of features affect prediction accuracy. Understanding the functional difference between engineered features and machine learning tasks can provide insight into structural and functional changes seen in medical imaging.

Deep learning has shown a remarkable improvement over hand-crafted feature-based learning, but it is still unclear whether deep neural networks capture all the information available from expert features. Deep neural networks learn convolutional filters that minimize an objective loss function, such as mean squared error, but do not enforce specific anatomic of physiological principles present in the image. Conversely, using engineered features based on anatomy or function requires a priori expertise and will necessarily limit the information available in the image to the chosen features. The rationale for merging expert features with deep learning is to leverage the existing knowledge present in expert features to direct learning of the convolutional network towards intensity patterns predictive of age that are not captured in the engineered features. The principal contribution of this method is to show that deep learning can be used

to enhance prediction along with engineered features. A strength of deep convolutional networks is the ability to find patterns in the data that are not immediately obvious. Therefore, it is best to leverage this powerful tool to find new patterns instead of enforcing the learning of features that we can already obtain through classical methods.

In this chapter, we hypothesize that deep learning is complementary to traditional feature estimation. We propose to build upon previously validated network designs to include traditional structural imaging features alongside deep convolutional ones and illustrate this approach on the task of imaging-based age prediction on two separate datasets: T1-weighted brain MRI and CT of the head. We show that deep learning can enhance tasks in medical image processing when learned features are combined with handcrafted features. As the field of medical image processing continues to adopt techniques in deep learning, it will be important to preserve hand-crafted features that enhance the task at hand, instead of replacing the features altogether. The work presented in this chapter was published at *Magnetic Resonance Imaging* with Camilo Bermudez as first author.

2. Methods

2.1. Imaging Datasets and Preprocessing

The complete MRI cohort aggregates 9 datasets with a total 5,048 T1w 3D images from normal healthy subjects, as curated by [41]. This cohort includes subjects marked as controls from nine studies (Table II.1). The data include subjects with ages ranging between 4 and 94 years old, with a mean age and standard deviation of 29.1 ± 22.6 years. Of 5,048 subjects, 52.4% were male and 47.6% were female. Data were also acquired from different sites so there is a difference in field strength, of which 77% of scans were acquired at 3 Tesla and 23% were acquired at 1.5 Tesla. ROI volumes, sex, and field strength were all used as input features for age prediction.

Site	Mean Age	Young Age (0 - 30 y/o)	Middle Aged (30 -50 y/o)	Older Adult (50 - 96 y/o)	Site total
ABIDE	17.2 ± 7.8	523 (95)	39 (3)	1 (0)	563 (98)
ADHD-200	11.6 ± 3.3	950 (367)	-	-	950 (367)
BLSA	68.1 ± 12.7	1 (0)	61 (31)	552 (311)	614 (342)
Cutting	12.5 ± 5.0	583 (293)	3 (1)	-	586 (294)
FCON-1000	28.3 ± 13.8	823 (469)	130 (56)	116 (68)	1,069 (593)
IXI	48.8 ± 16.4	98 (54)	166 (80)	259 (162)	523 (296)
NDAR	11.0 ± 3.8	328 (168)	-	-	328 (168)
NKI-Rockland	33.9 ± 21.5	58 (26)	21 (8)	24 (14)	103 (48)
OASIS	45.2 ± 23.8	139 (78)	43 (24)	130 (93)	312 (195)
Total	29.1 ± 22.7	3503 (1550)	463 (203)	1082 (648)	5,048 (2,401)

Table II.1Demographics for brain MRI cohort per site. Our study uses brain MRI of subjects marked as healthy controls from nine different sites. Parenthesis indicate number of female subjects.

For feature extraction, 45 atlases are non-rigidly registered [134] to a target image and non-local spatial staple (NLSS) label fusion [135] is used to fuse the labels from each atlas to the target image using the BrainCOLOR protocol [136]. A total of 132 regional volumes were calculated by multiplying the volume of a single voxel by the number of labeled voxels in original image space. Total intracranial volume (TICV) was calculated using SIENAX [137] and used for volume normalization for a total of 132 raw volumes and 132 normalized volumes.

The cohort of head CT images consists of 1,313 clinically acquired scans as part of a larger study on eye disease. All images were acquired at Vanderbilt University Medical Center (VUMC) with variable imaging protocols and scanners (Table II.2). Images were retrieved and deidentified for retrospective study under Institutional Review Board (IRB) approval. Since these subjects were undergoing CT imaging for regular clinical care, it includes patients with eye diseases such as glaucoma, intrinsic optic nerve disease (IOND), optic nerve edema, orbital inflammation, or thyroid eye disease (TED), as well as healthy controls who received imaging but were not clinically diagnosed with these diseases. CT images were normalized to a range between -100 and 200 Hounsfield Units (HU) for optimal visualization of orbital structures. The CT images were processed for structural features of the orbit using the protocol described in [138]. The data includes subjects of ages ranging between 1 and 97 years old, with a mean age and standard deviation

of 52.1 +/- 20.7 years. This dataset includes healthy control subjects (70.0%) as well as subjects with glaucoma (7.7%), IOND (15.2%), optic nerve edema (14.7%), orbital inflammation (2.5%), or TED (5.0%). Seventy-five structural metrics were extracted using multi-atlas segmentation and used as input variables along with disease classes for each subject, using the method proposed by Harrigan et. al. [138].

Cohort	Mean Age	Young Age (0 – 30 y/o)	Middle Aged (30 -50 y/o)	Older Adult (50 – 96 y/o)	Total
Healthy Control	56.5 ± 19.5	108	107	651	866
Glaucoma	61.0 ± 18.7	8	15	78	101
IOND	44.6 ± 20.0	54	64	82	200
Optic Nerve Edema	32.0 ± 14.7	97	75	21	193
Orbital Inflammation	46.1 ± 21.6	9	10	14	33
TED	53.3 ± 14.0	2	25	38	65
Total	52.1 ± 20.7	242	252	819	1,313

Table II.2 Demographics for head CT cohort per disease status. Our study uses clinically acquired head CT from healthy subjects as well as five different eye disease status. Note: Some of these subjects have multiple diagnoses. IOND: Intrinsic optic nerve disease; TED: Thyroid eye disease.

2.2. Model Architecture

In this work, we adopt the network developed and validated by Cole et. al and extend it to include anatomical features derived from multi-atlas segmentation [83]. In Cole's work, a series of convolutional operations result in a high-dimensional convolutional representation of imaging data, which is then used to directly predict age. In this work, we concatenate the anatomic representation, which consists of volumetric estimates of key ROI's identified with multi-atlas segmentation. Figure II.1 shows the convolutional representation obtained using a 3D convolutional neural network. This consists of 5 layers each with two, 3D convolution operations, ReLU activation, and max pooling; resulting in a representation of the brain imaging with 15,360 features or 3,584 features in the case of orbital CT.

After concatenation of the convolutional and volumetric features these features undergo ReLU activation and densely connected layer to a single node with linear activation to directly predict age (Figure II.1). In addition, we trained two more baseline models for a comparison: 1) the baseline Cole et. al model using our data, and 2) a volumetric features only model, which consisted of two densely-connected layers of 128 nodes. As with previous work, the learning rate was initiated at 0.01 with a 3% decay each epoch. All models were trained using stochastic gradient descent optimization with momentum of 0.9. The loss function was mean absolute error. Training was allowed to continue until the loss function on the validation set did not change by more than 0.1 in 20 epochs. All models were developed using Keras version 2.2.4 with Tensorflow 1.5 and trained on an NVIDIA Titan Graphics Processing Unit (GPU).

2.3. Statistical Analysis

We performed a five-fold cross-validation scheme by withholding 20% of the data for testing while using the remaining 80% for training and validation (70% and 10% respectively). This process was repeated five

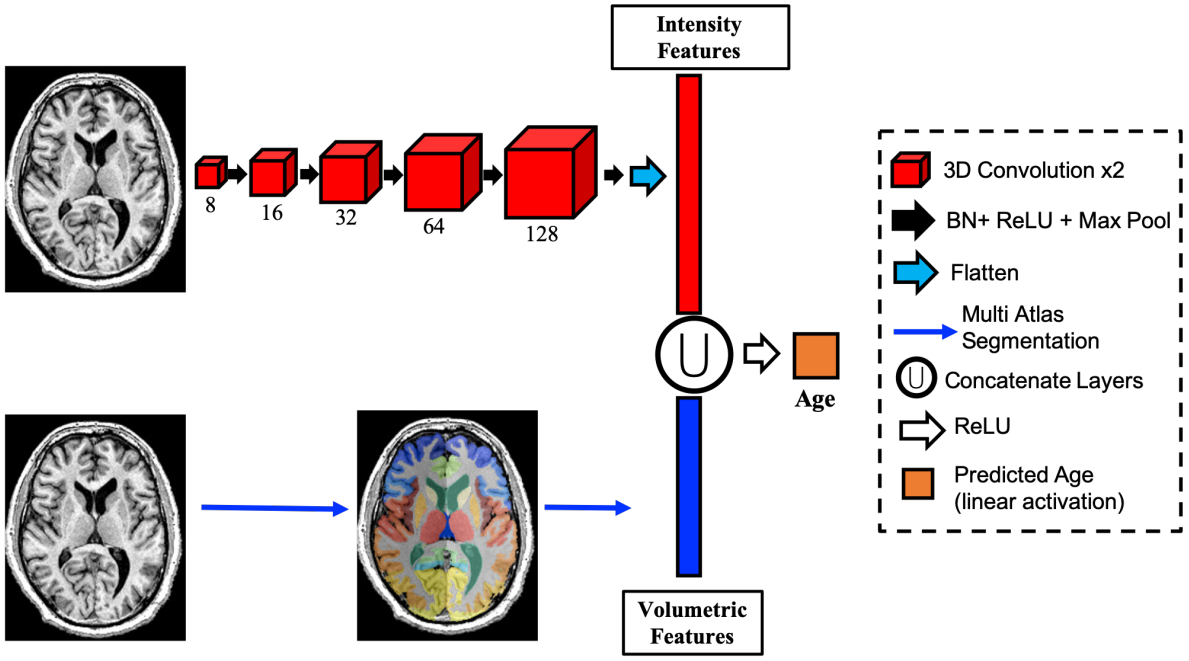

Figure II.1 Pipeline for age prediction. Two sets of features are used: intensity-derived features (red) derived from a convolutional neural network of increasing filter size (red boxes), and structural features (blue) using multi-atlas segmentation (bottom). These features are concatenated and used as inputs to directly predict age. BN: Batch Normalization; ReLU: Rectified Linear Unit Activation; Max Pool: Max Pooling Layer.

times until the entire dataset was used for testing only once. Therefore, a total of five networks were trained for each method (ie. volumetric features only, raw image only, or combined) and evaluated on the testing set for each fold. This way, every subject in the data is used as a testing subject only once. All results shown below represent the evaluation of the testing set in each fold grouped together. The BAG biomarker was calculated for all models and compared using a Wilcox signed rank test. We corrected for multiple comparisons using Bonferroni corrections.

Additionally, we test the hypothesis that the BAG biomarker is increased in subjects with existing eye disease compared to healthy controls in the testing set. We tested for significance of the BAG biomarker between each disease group and healthy controls using a linear regression model with disease state as a dummy variable and true age as a covariate. Significance between groups was set to $p < 0.05$ in the dummy variable.

2.4. Network Visualization with Grad-CAM

We use the Gradient Class Activation Maps (Grad CAM) method described by [55] and implemented by the publicly available library keras-vis [139] to visualize the areas of the raw MRI with higher attention in the combined model. We created nine categories based on three age cutoffs for true age and predicted age: young (< 30 years old), middle aged (30 – 50 years old), and older adult (> 50 years old). The nine categories consist of young predicted young, young predicted middle aged, young predicted older adult, middle aged predicted young, middle aged predicted middle aged, middle aged predicted older adult, older adult predicted young, older adult predicted middle aged, and older adult predicted older adult. We randomly selected 10 subjects from each category, calculated the Grad CAM for each and showed the resulting average for each category overlaid on a subject from that sample.

3. Results

In this work, we show that intensity-derived features from a deep convolutional network can enhance learning from structural features by improving the accuracy of age prediction in both brain MRI and head CT datasets.

3.1. Context-Aware Deep Neural Network Best Predicts Age in T1-Weighted MRI

This study investigates the ability to predict age from brain MRI in healthy individuals. We use two sets of inputs to train and validate a fully-connected network: 1) intensity features derived from a convolutional neural network representation, and 2) context features including volumetric estimates for known regions of interest in the brain obtained via atlas-based segmentation. We also introduce sex and scanner field strength with structural features as additional contextual features. Figure II.2A shows that

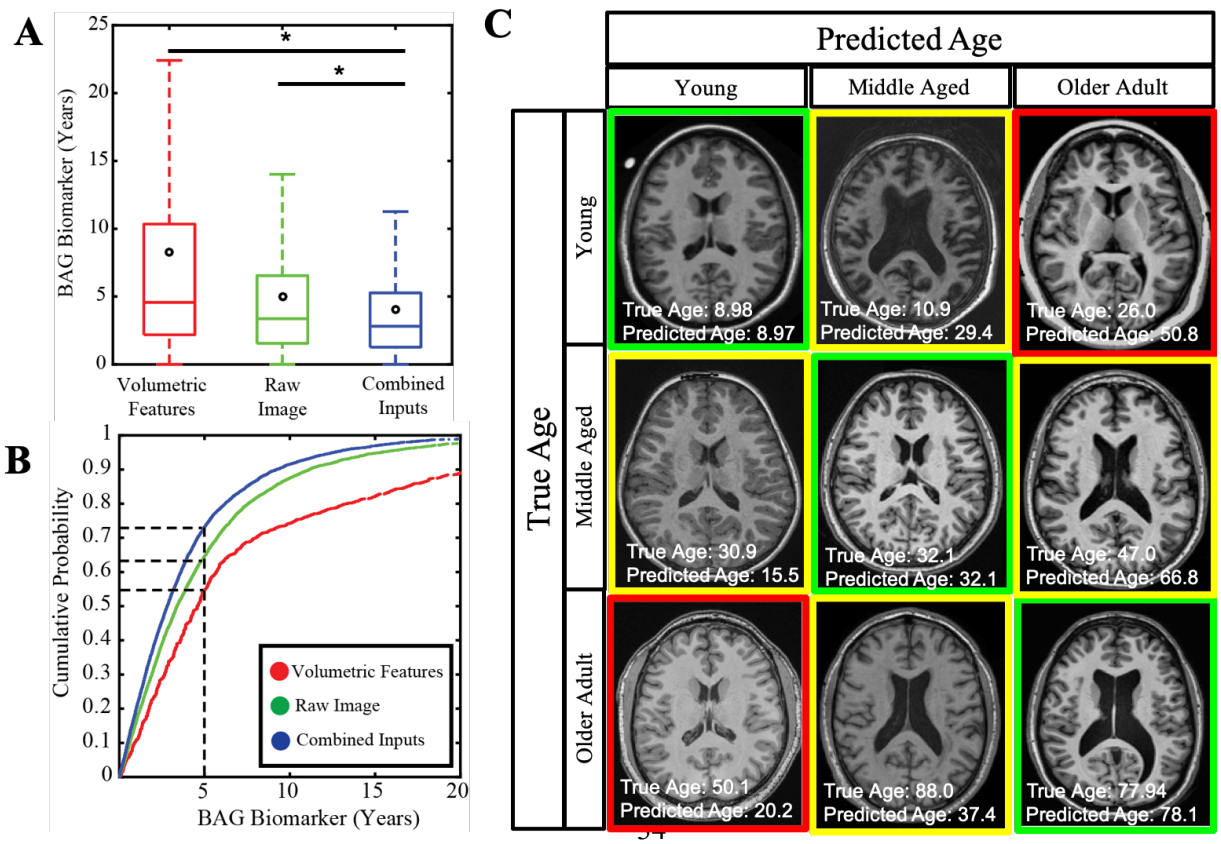

Figure II.2 Deep Learning Improves Age Prediction in Brain MRI. Age can be predicted more accurately when using convolutional and structural features on a fully connected network model (A). Most subjects can be predicted within 2.81 years using the combined model (B). Subjects who are predicted old have features of young patients and vice versa (C).

intensity-derived features outperform structural features with a mean absolute error of 5.00 vs 8.26 years ($p<0.001$). However, age prediction improves when both feature sets are used as inputs, resulting in a median absolute of 4.08 years ($p<0.001$).

Figure II.2B shows the cumulative probability of accurate prediction within an acceptable error range. This shows that 73.2% of subjects fall within an absolute error of 5 years in the combined model. This is an improvement from the intensity features model, which can only predict 64.6% of the subjects within 5 years while the features only can predict 54.9% within 5 years. Figure II.2C presents a representative 3-by-3 matrix of subjects where the vertical axis represents true age of a young individual, a middle-aged adult, and an older adult. Some of the canonical features of old age, such as enlarged ventricles are seen in young patients predicted as old. Conversely, anatomical features of young age such as small cortical sulci are apparent in the older subjects.

Imaging Modality	Input Data	MAE (yrs)	RMSE	R	R^2
Brain MRI	Volumetric Features	8.23	12.91	0.84	0.70
	Raw Image [83]	5.00	7.25	0.95	0.90
	Combined Features	4.08	5.93	0.97	0.93
Orbital CT	Volumetric Features	13.28	18.02	0.45	0.20
	Raw Image [83]	11.02	14.11	0.75	0.56
	Combined Features	9.99	13.19	0.76	0.58

Table II.3 Accuracy of all three models on brain MRI data. MAE: mean absolute error. RMSE: Root mean squared error. R: Pearson correlation coefficient.

3.2. Context-Aware Age Prediction Generalizes to Orbital CT

We applied the same network architecture to predict age on a dataset of head CT used in [140], which includes healthy and eye disease populations. Again, we use two inputs: 1) intensity features derived from a convolutional neural network representation, and 2) volumetric estimates of important orbital structures proposed by [138] as well as structural features. Figure II.3A shows that intensity-derived features alone outperform hand-crafted features, resulting in a mean absolute error of 11.02 years and 13.28 years

respectively ($p<0.001$). However, using both feature datasets as inputs allows for a significant improvement in prediction, resulting in a median absolute error of 9.99 ($p<0.001$).

In the CT dataset, 60.3% of subjects predicted within 10 years (Figure II.3B) using the combined model, whereas the 10-year limit only covers 53.2% and 52.5% of the intensity-only or feature-only models, respectively. Representative images are shown in the 3-by-3 matrix in Figure II.3C, where the vertical axis represents true age of a teenager, a middle-age adult, and an older adult.

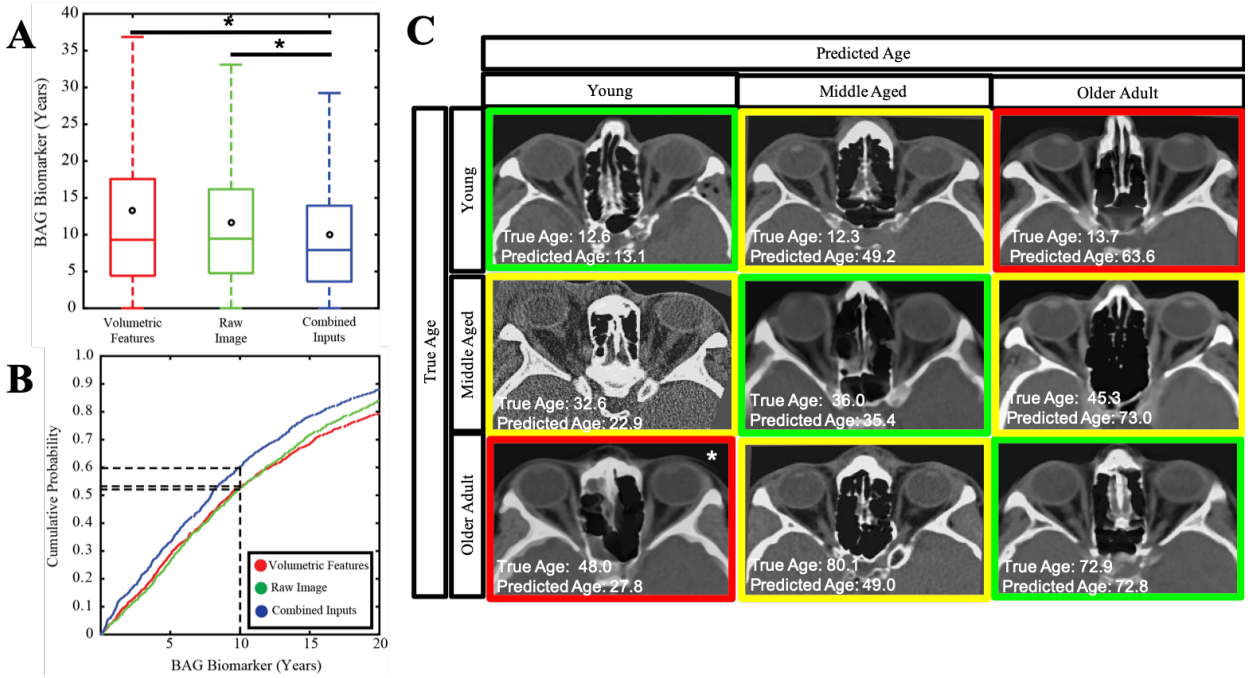

Figure II.3 Deep Learning Improves Age Prediction in head CT. Age can be predicted more accurately when using convolutional and structural features on a fully connected network model (A). Most subjects can be predicted within 7.90 years using the combined model (B). Accuracy of prediction does not show characteristic features of aging with our model (C). Note that while the older adult predicted young is 48 years old and technically in the middle aged bin, this was the oldest subject predicted within the young category, so we include it as a proxy.

3.3. OrbitBAG as a Marker of Disease

We further extend the development of a new imaging biomarker and validate it against 5 different diagnoses of eye disease. We estimate OrbitBAG for healthy controls in the testing set as well as unseen subjects with glaucoma, intrinsic optic nerve disease, orbit nerve edema, orbital inflammation, or thyroid eye disease. Figure II.4 shows that the OrbitBAG biomarker is significantly elevated in patients with intrinsic optic nerve disease ($p < 0.001$) and orbital edema ($p < 0.001$) when controlling for true age. A trend of increased OrbitBAG is observed with glaucoma, orbital inflammation, and thyroid eye disease, albeit not significant ($p = 0.48$, $p = 0.51$, $p=10$, respectively). In the case of orbital edema and thyroid eye disease, OrbitBAG biomarker was not markedly elevated. OrbitBAG values for each condition are shown in Table II.4.

Cohort	MAE (yrs)	RMSE	R	R^2
Controls	9.62	12.25	0.82	0.68
Glaucoma	9.97	12.97	0.73	0.53
Intrinsic Optic Nerve Disease	13.92	17.12	0.66	0.44
Optic Nerve Edema	15.66	18.32	0.62	0.38
Orbital Inflammation	11.36	14.11	0.77	0.60
Thyroid Eye Disease	11.69	15.03	0.43	0.19

Table II.4 OrbitBAG accuracy results for each disease cohort. MAE: mean absolute error. RMSE: Root mean squared error. R: Pearson correlation coefficient.

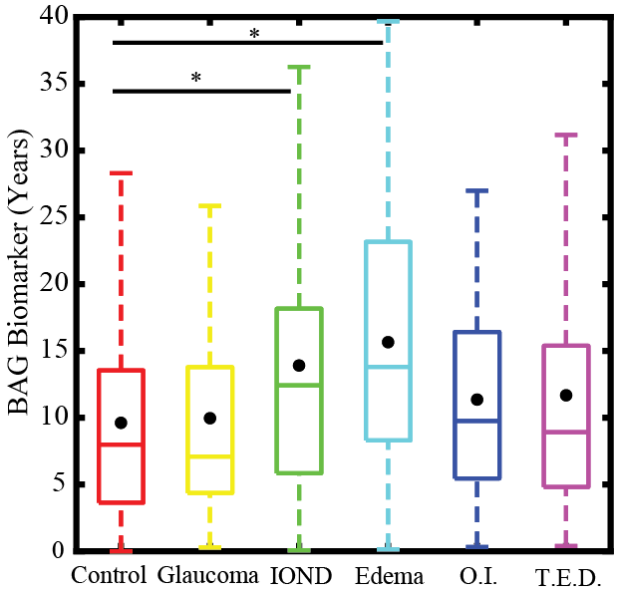

Figure II.4 OrbitBAG biomarker for respective cohorts within the Orbital CT dataset. Intrinsic optic nerve disease (IOND) shows a significant increase in OrbitBAG. A similar trend is observed in glaucoma and orbital inflammation, but not statistically significant. Edema and thyroid eye disease (TED) do not show an increase in OrbitBAG.

4. Discussion

In the task of age prediction, deep learning performance can be enhanced by providing context features, such as volumetric estimates of important anatomical structures. We provide a more accurate method of age prediction compared to the literature. Additionally, we show that the proposed method of integrating deep learning with anatomical, hand-crafted features is not unique to brain MRI, but generalizes to orbital CT. We provide a new biomarker OrbitBAG based on orbital CT to estimate age and show it is elevated in patients with intrinsic optic nerve disease. Together, these results provide an accurate biomarker of aging in two different imaging modalities and demonstrate that there is still valuable information in classical image processing that is not being captured by deep learning.

4.1. Image Processing Features Enhance Deep Learning

Several studies have shown that deep neural networks are highly susceptible to changes in output to small adversarial perturbations or irregularities in the data [141-146]. For example, scene recognition, a task mastered by deep learning, has been shown to be disrupted by single pixel changes [141, 143]. Similarly, Nguyen et. al were able to produce images that are unrecognizable to humans but deemed recognizable by a convolutional neural network with 99.99% confidence [144] .In this regard, anatomical features provide a robust representation of the image, compared to the learned convolutional filters. By providing contextual information, the model proposed here can help stabilize the network and capture information not available in the known features. Moreover, the large spatial extent of deep convolutional networks can learn to ignore unnecessary information such as background noise or artifacts. Together, features derived from a convolutional representation as well as features derived from classical image processing can enhance prediction tasks. Here, we see a significant improvement in age prediction accuracy in both brain MRI and

head CT, suggesting that the improvement achieved by incorporating anatomical features is not specific to one modality.

Besides a general improvement in accuracy in computer vision tasks like object recognition and segmentation, deep learning has also shown improvements in computational time required for evaluating new images. Although training a deep neural network may take hours to days, evaluating a new image for segmentation or age prediction takes a matter of seconds. This shows a dramatic improvement over common image processing techniques such as multi-atlas segmentation, which can take hours to days to evaluate a single subject. In the work presented here, we propose the incorporation of deep learning with expert features. One limitation of our method is that producing an estimate of age will be dependent on the time needed to craft expert features for a test subject. If these are generated using multi-atlas segmentation, it may take hours. However, new techniques in deep learning are emerging that can replicate common deep learning tasks in a fraction of the time. For instance, Huo et. al. were able to generate a whole-brain volumetric segmentation using deep learning with similar accuracy to multi-atlas segmentation [147]. We believe that incorporating anatomical context via hand-crafted features in deep learning pipelines can boost performance and generalizability while still being time-efficient.

Decades of work in image processing before deep learning have developed contextual features such as functional and anatomic regions of interest, surface parcellation, regional connectivity, and deformation models. Unlike the feature maps used in deep learning, these hand-crafted features are often built on underlying principles of anatomy and physiology. A large sample size of these features can possibly capture the entire manifold of possible human values. However, such restrictions do not exist in deep learning and extensive datasets are needed to arrive at plausible solutions. Contextual features could be used to limit the search-space of deep neural networks, diminish the number of parameters needed and avoid overfitting. This work also raises the question of robustness against adversarial attacks when a neural network is

grounded by contextual features. Future work may benefit of exploring the role of such features in preventing adversarial attacks in medical imaging.

An alternative method to enhance image processing techniques is to iterate between deep learning and traditional feature engineering. New attention networks like the one used by Huo et. al. [148] could help identify key regions of valuable anatomical information for the specified task. These maps can be used along with anatomic and physiological context to craft better manual features. These hand-crafted features can then inform a future network to improve algorithm accuracy and refine feature maps. It is possible that with more data and more complex deep learning models, similar accuracy can be achieved with imaging alone. In this work, we propose a method to guide training with imaging features that are already available. Instead of focusing on improving prediction accuracy, a stable model that integrates clinical and imaging context can be a fruitful ground for inference on key anatomical features preferentially affected by ageing.

4.2. Brain Age Gap Used as an Imaging Biomarker of Aging

Previous studies have shown that BAG correlates with aging in a healthy population as well as neurodegenerative diseases in brain MRI [129]. In this study, we show that an age prediction biomarker can be developed on a new imaging modality, Orbital CT, and validated against diseased populations. We show that the OrbitBAG biomarker is generally increased in populations that show structural changes such as glaucoma, edema, and orbital inflammation. We also observe a wide distribution of error in age prediction (Figure II.2 and II.3) in a healthy population. It is possible that these changes are a result of normal anatomical variability or susceptibility of disease. A longitudinal study of these subjects, along with clinical data, would show whether having an outlier BAG is predictive of future disease.

Here, we have demonstrated that while much attention has been dedicated to predicting chronological age from brain MRI, it is also possible to predict age from different imaging modalities obtained from different parts of the human body. With the increase in usage of medical imaging and processing capabilities, it may be possible to predict age from multiple body parts and multiple imaging modalities at once. A whole-body age prediction algorithm may better reflect the state of the entire body and find interesting associations in the aging process of different human organs.

4.3. Network Visualization on Raw MRI Input using Grad CAM

We proposed the integration of contextual anatomical figures along with raw imaging to inform the task of age prediction. A key open question in the field of medical image processing with deep learning is whether deep neural networks capture all available information in an image, or if these algorithms can be better trained by enforcing high-quality a priori information, such as volumetric estimates in regions of interest. It is often difficult to interpret the meaning of each convolutional layer in a deep convolutional network. However, a common tool for data visualization in deep learning are Gradient Class Activation Maps (Grad CAMs), which highlight areas of attention in the input image. Here, we have generated Grad CAMs for the

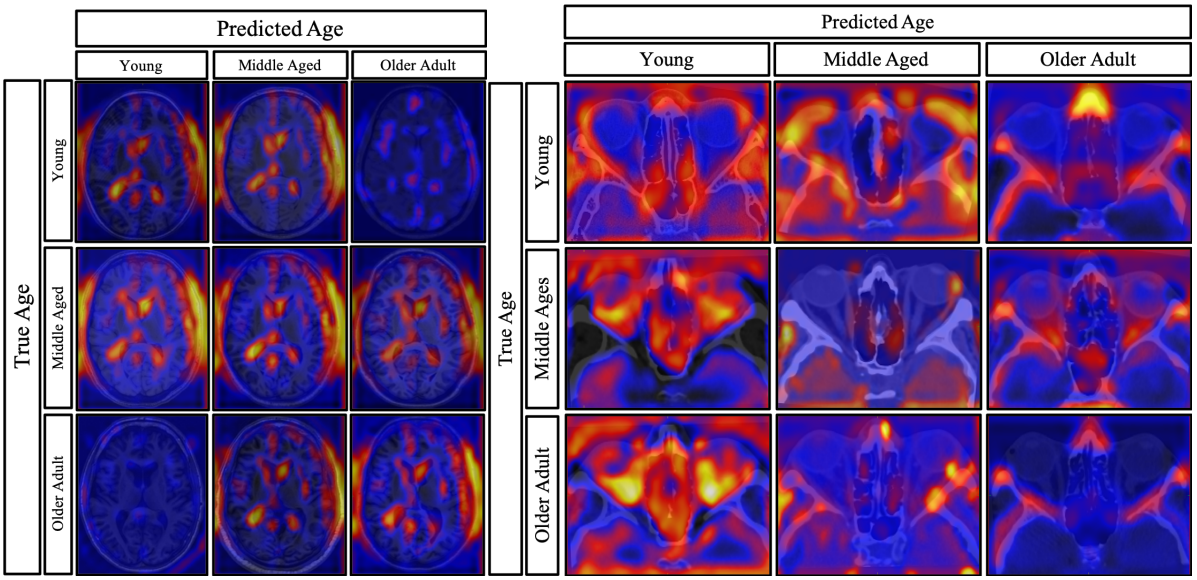

Table II.5 Gradient Class Activation Maps (Grad CAM) visualization for raw MRI (left) and raw CT (right) used in the combined networks. Visualizations were binned according to true and predicted age. Ten random subjects were chosen from each category to compute the Grad CAM and the maps were averaged. Activation maps are overlaid over a representative subject from the sample.

subjects in the testing set for both the MRI and CT task (Figure II.5). In the case of MRI, we see wide activation throughout the brain, but particularly centered around head size, the cerebral cortex, and the size of the ventricles. Although volume measurements of the ventricles are included as features in the combined model, the attention maps simultaneously encompass other areas of the brain, suggesting a complex interaction between areas of the brain. Importantly, we do not observe a clear and consistent segmentation of the brain structures which were included as volumes. Interestingly, the subjects off the diagonal, which show the worst prediction of age, have the least activation compared to the other subgroups. In the case of Orbital CT, there is a focus on the skull bones and the nose, which were not included as features, but may be indicative of aging. Interestingly, the activation in younger subjects was more wide-spread than in older subjects, suggesting higher variation across younger subjects, while bone structures may be a stronger indicator of age in older subjects.

Overall, Grad CAMs offer an interesting heuristic for visualization of attention from convolutional neural networks. In both CT and MRI, the attention maps seem to reiterate common changes associated with age such as head size or bone structures. It is reassuring that none of the activation maps uniquely highlight a segmentation corresponding to the volumetric features. However, it is still an open research question on how to interpret these maps and a deeper study is needed.

5. Conclusions

Applications in deep learning have focused on raw images due to the power in automatic feature recognition. However, we show that there remains valuable information in the features derived from image processing. These features with anatomical context can be used to complement deep learning tasks, especially, when there is finite training data. This work has significant implications in the field of medical

image processing, as decades of work on feature optimization can be used to improve on already groundbreaking deep learning breakthroughs.

III. Using Anatomical Features to Protect Against Adversarial Attacks in Brain Age Prediction

1. Preface

The work done in this chapter emerged from a collaboration with Dr. Yevgeniy Vorobeychik's group while at Vanderbilt University. I was closely involved in the conception, design, and improvement during the duration of this project, but I did not perform the experiments of adversarial training. I wrote drafts for the introduction, methods, and discussion of the work in this chapter. This chapter was included in this dissertation not only for my involvement in the project, but because we believe it is a relevant and innovative application of the work presented in Chapter II. The work presented here was published at *Neurocomputing* with Camilo Bermudez as the third author. Here, I present my contributions and a summary of the published work.

2. Introduction

Deep learning methods are transforming the way scientists approach data with machine learning across disciplines, and have recently become prominent in medical imaging. For example, Esteva et al. [149] showed the same accuracy as a dermatologist in the detection of malignant skin lesions using deep learning techniques. Gulshan et al. [150] used similar methods for the detection of diabetic retinopathy in retinal fundus photographs with great accuracy, while Bejnordi et al. [151] were able to accurately detect lymph node metastasis in patients with breast cancer, and Cole et al. [83] used deep learning to produce a more accurate brain age prediction, which has been shown to correlate with neurodegenerative diseases. Given the remarkable results that deep learning methods have shown compared to standard clinical practice, these have

started to be implemented into clinical practice, with several deep learning applications already approved by the United States Food and Drug Administration (FDA) [152-154].

Concerns regarding the clinical safety of deep learning models persist due to high-complexity of the model and the consequent difficulty in explainability of the features driving model prediction. This "black box" effect may hide the context being learned, such as learning the difference between a dog and a wolf based on the snow in the background [155]. Specifically, potential adverse consequences have been recently explored by introducing imperceptible perturbations to the image, resulting in large deviations in the predicted output [143-145, 156-158]. Relevant to medical imaging, these adversarial perturbations to the image capable of altering predictions can be indistinguishable to human vision and could have great implications in computer-assisted radiologic reading studies. This effect has been previously observed in medical image analysis systems [159].

Meanwhile, efforts of integrating engineered imaging features as contextual information into deep learning models have demonstrated an improvement in performance while driving model interpretability [160] [161]. However, it is yet unknown if introduction of contextual features offers a protective effect against adversarial perturbations.

Herein, we show that the medical imaging applications are susceptible to adversarial attack and that this susceptibility can be partially mitigated by integrating contextual anatomical features in a deep learning model. Specifically, We use Brain Age prediction as an application to show the following: i) A single visibly imperceptible noise field will reduce task accuracy ii) A general visibly imperceptible noise field can be generated which will reduce task accuracy on any subject iii) Adversarial attacks are less effective on deep learning models that have anatomical context.

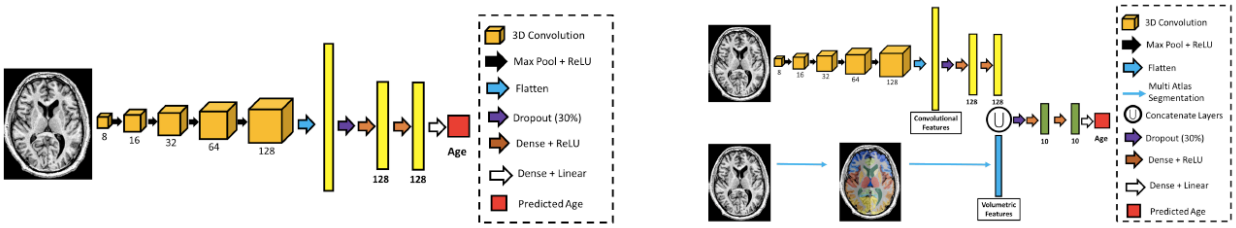

Table III.1 Models for brain age prediction. (left) Shows the model with raw MRI alone as input. (right) Shows the context-aware model with raw MRI and volumetric features.

3. Methods

3.1. Dataset

Our imaging dataset is an aggregate of 7 datasets with a total 3921 T1w 3D images from normal, healthy subjects. For a full description of the datasets used, these are the same publicly-available datasets used in Chapter II. The data include subjects with ages ranging between 4 and 94 years old, with a mean age and standard deviation of 25.5 ± 18.6 years. Of the 3921 subjects, 54.2% were male and 45.8% were female. Data were also acquired from different sites so there is a difference in field strength, of which 71.5% of scans were acquired at 3 Tesla and 28.5% were acquired at 1.5 Tesla. ROI volumes, gender, and field strength were all used as input features for age prediction. For these experiments, all brains were co-registered using an affine transformation to the MNI-305. All image intensities were normalized to [0,1] during training and testing of Brain Age algorithms and for the adversarial learning experiments.

3.2. Deep Learning Models

We considered two models for predicting age: 1) a conventional deep neural network (DNN), and 2) a context-aware model which combines deep learning with image segmentation techniques as shown in Figure III.1. These networks are based on the work of Bermudez et al. [125] The conventional deep neural

network model takes a 3D brain MRI image as input and produces a subject's age as output. The architecture consists of five 3D convolution layers of increasing size followed by two densely connected layers and one output layer. The ReLU activation function was used for all hidden layers. The neural network was trained using a learning rate of 0.001. The structure of this model is shown in Figure III.1. The context-aware model has a similar structure to the conventional deep neural network model, with the exception that 132 volumetric features are introduced after the convolutional layers followed by two densely connected layers and, finally, the output layer. Volumetric estimates for 132 regions of interest in the brain were obtained using multi-atlas segmentation [136, 162].

3.3. Adversarial Noise

The methods for adversarial attacks will be summarized here. For a detailed overview and rationale of adversarial learning in the context of Brain Age, please see [126]. The goal in this work is to introduce imperceptible noise to the input image such that the change in predicted age is maximized. Three kinds of noise patters are used: 1) modifies each pixel independently, putting a limit on the maximum change to any pixel intensity, 2) limits the total number of pixels that can be modified, and 3) limits the Euclidian norm of the intensity of the noise field inserted in the image. In this experiments, we also test different levels of imperceptibility limits to the noise for each method. The measure of success for adversarial learning will be the deviation, or absolute difference, between the original predicted age and the perturbed predicted age. The noise field is learned by following the gradient of the brain age prediction function in the direction of maximizing or minimizing predicted age in order to design the noise field that best maximizes deviance predicted age.

4. Results

In order to establish a baseline, we tested the impact of adversarial noise on the conventional DNN for a single image. All three methods of adversarial noise showed a large deviance in predicted age of 20 to 80 years, with increasing deviance as the noise limit increased. Age can be amplified by at least 30 years by introducing noise up to 0.001 in any noise pattern. Similar results are observed when minimizing predicted age.

Next, we tried to learn a noise field that can maximize deviance for any one image. While adversarial attacks are best suited to affect performance in a single image, a general adversarial noise field would be able to impact any image without having to learn subject-specific patterns. Using noise pattern (1), bounding the most a single pixel can be changed, we showed a predicted age deviation of over 10 years. We noticed that this effect was not dramatically affected by the number of samples used, as the average deviation was stable at 10 years for 300, 500, and 1,500 subjects.

We then tested adversarial attacks on the Integrated DNN, which uses anatomical context features to improve predictive performance of brain age. We showed that using anatomical features significantly reduced the adversarial perturbations by an average factor of two throughout all noise patterns at increasing noise thresholds. However, the adversarial effect was still significant enough to cause an average deviation of 20 years while maximizing age or 10 years while minimizing age. This effect was maintained when learning an adversarial noise field for a large number of images.

5. Discussion

Deep learning has revolutionized the field of medical image analysis with unprecedented accuracy and speed. However, here we show possible concerns with robustness and safety of these algorithms when presented with adversarial noise. Thus, adversarial attacks are a reasonable concern that should be

addressed when translation algorithms to assist in clinical decision making in the future. These kinds of perturbations may appear naturally in unanticipated situations such as scanner noise or artifact, or they may appear in the context of malicious tampering with the goal of introducing bias in computer-assisted diagnosis or driving healthcare costs through false positive outputs. Current radiology pipelines have robust security protocols but these are not uniformly implemented or applied [163]. While breach of healthcare security systems is difficult, our results show that we can generate a standard perturbation to introduce error in a large number of images, thus making the introduction of imperceptible noise relatively easier. Lastly, the lack of explainability in "black box" deep learning models make it difficult to detect and understand where erroneous predictions are originating.

In this work, we proposed a method to mitigate the effect of adversarial attacks by integrating contextual knowledge into a DNN via anatomical measurements. This high-level information introduced in the model may be less susceptible to voxel-level changes. Unfortunately, the absolute effect on the integrated DNN, despite reducing deviation by half, would be enough to significantly alter predictions, thus introducing concerns of the possible applicability during clinical workflows. Possible alternate approaches include introducing adversarial noise to the input images during training to reduce susceptibility to future adversarial attacks [143, 164]. Overall, we showed that while deep learning methods have shown significant advances in the field of medical image analysis, there are still robustness issues that need to be explored and understood before these tools can be deployed in clinical practice. It is likely that a combination of tools like anatomical context and learning with adversarial noise may be enough for deep learning algorithms to translate safely and assist during clinical decision-making.

IV. Accelerated Brain Aging Predicts Impaired Cognitive Performance and Greater Disability in Late but not Midlife Depression

1. Introduction

Aging has an inevitable effect at molecular, cellular, and organ levels, with biological aging resulting in degeneration or reduction in the organ's reparative or regenerative potential [165]. "Accelerated aging" refers to biological aging processes that occur more rapidly than expected, resulting in biological characteristics appearing older than expected based on the individual's chronological age. Accelerated aging may result from numerous disease processes and can be quantified using a variety of markers such as oxidative stress measures [166], telomere length [167] or epigenetic measures of methylation such as Horvath's epigenetic clock [168]. Differences in these markers of accelerated aging are reported in neuropsychiatric disorders including schizophrenia, post-traumatic stress disorder, anxiety disorders, and depression [169]. Depression is specifically associated with decreased telomere length [167], while a multi-biomarker index of aging derived from measures of inflammation, metabolism, and organ function predicted greater depression severity in older adults [170]. Importantly, rates of aging as measured by biomarkers differ across organ systems. As the central nervous system may age differently than the rest of the body [171], brain-specific markers of biological aging may be particularly germane to neuropsychiatric disorders.

By building on large databases of normative aging, examination of structural MRI data allows for examination of accelerated aging in the brain itself. One approach, the Brain Age Gap Estimation (BrainAGE) method [125], utilizes a machine-learning technique for identifying individual-level variability in brain aging. Using standard structural MRI sequences, a prediction model generated from a learning sample of neurologically healthy adults can be applied to a new individual brain MRI to estimate that individual's apparent biological age. The difference between this estimated biological age and the subject's chronological age yields the brain-age "gap" (BAG), a marker of how much "older" or "younger" a given

brain appears relative to the individual's chronological age. This technique has been applied in psychiatric populations including schizophrenia, where particularly obese individuals exhibit older-appearing brains, although differences in estimated brain age are not seen in patients with bipolar disorder [79, 172, 173]. Past studies using brain age estimation techniques examining adult major depressive disorder (MDD) report that individuals with MDD tend to exhibit older estimated brain ages than expected [174], although this finding is not universal [175]. If MDD is indeed associated with accelerated brain aging, this may be related to the fact that development of depressive episodes often involve environmental stress exposures [176], resulting in increases in allostatic load [177] that in turn contribute to accelerated aging processes. Thus, if MDD is associated with accelerated aging, that effect may be time dependent and require repeated stressors or depressive episodes. Such a hypothesis is supported by past work associating greater chronicity or duration of depression is associated with volumetric differences in key regions such as the hippocampus [178-180].

If this theory is correct, examination of brain aging may have greater utility in older populations and late-life depression, or geriatric MDD. Geriatric MDD is associated with cerebrovascular pathology [180], higher risk for dementia [181], and greater medical morbidity [182] that may contribute to impairment in multiple cognitive domains [183] and accelerated brain aging. For example, diabetes mellitus, a risk factor both for cerebrovascular disease, dementia, and depression, is also associated with an increased brain-age gap [184]. The potential utility of brain-age estimation is supported by work in Alzheimer's disease, where patients exhibit a greater brain-age gap than seen in cognitively intact elders [185] and a greater discrepancy between estimated biological brain age and chronological age predicts conversion from mild cognitive impairment to dementia [186]. Structural MRI as a marker of accelerated aging may therefore provide new insights in the interactions between aging, depression, cognition, and disability.

In this study, we calculated the BAG in two separate MDD cohorts, one of young- to midlife-adults and one of older adults. The goal of this study was to determine, using BAG as a cross-sectional marker of accelerated aging, whether depressed individuals exhibited accelerated brain aging and if this measure is

related to clinical outcomes, specifically cognitive performance and disability. We hypothesized that the estimated biological brain age of depressed participants would be older than their chronological age, and that this effect would be greater in the older cohort. We further hypothesized that a greater BAG indicating older biological age would be associated with poorer cognitive performance and greater disability, and that this effect would be particularly pronounced in depressed patients.The work presented in this chapter us currently under review at *Translational Neuropsychiatry* with Camilo Bermudez as co-first author.

2. Methods

2.1. Participants

The two cohorts included a group of young-to-midlife adults with and without MDD ("adult cohort" or "adult MDD") and a group of older adults with and without MDD ("geriatric cohort" or "geriatric MDD"). Other than the age criterion, these studies had similar entry criteria, with depressed participants being required to have a current diagnosis of MDD (DSM-IV-TR) and a Montgomery-Asberg Depression Rating Scale (MADRS) [187] score of 15 or greater. The studies shared common exclusion criteria including acute suicidality, current or past psychosis, current psychotherapy, electroconvulsive therapy in the previous 6 months, presence of central neurological disease, diagnosis of dementia, or unstable medical conditions, developmental disorders, and MRI contraindications. Never-depressed participants had neither a history of psychiatric diagnoses nor a history of mental health treatment. Both cohorts were outpatients recruited from clinical referrals and community advertisements.

The adult MDD cohort was enrolled at Duke University Medical Center. The eligible age range was 20 to 50 years. For depressed participants, entry criteria further specified a diagnosis of recurrent MDD with the onset of a first depressive episode prior to age 35 years and no antidepressant medication use in the last month. Exclusion criteria included other lifetime DSM-IV Axis I disorders including substance abuse or dependence, Axis II disorders identified by the SCID-II [188], use of illicit substances in the last month, a first-degree relative family history of bipolar disorder, or history of clinically relevant head injury.

Geriatric MDD participants were recruited at Vanderbilt University Medical Center as part of three separate studies with common entry criteria. Participants were age 60 years or older without a diagnosis of dementia or significant cognitive impairment assessed by a Montreal Cognitive Assessment (MoCA)[189] score greater than 24 or a Mini Mental State Exam (MMSE)[190] score greater than 24. For depressed participants, exclusion criteria included current or past Axis I disorder diagnoses, except for anxiety symptoms occurring during a depressive episode, history of substance abuse or dependence over the prior three years, and acute grief. Antidepressant medications were allowed in one geriatric study, with 9 of 14 depressed participants taking stable-dose antidepressant monotherapy at the time of MRI. The other two studies mandated no antidepressant use in the two weeks prior to MRI.

The Duke University Medical Center Institutional Review Board and the Vanderbilt University Institutional Review Board approved the studies conducted at each institution. All study participants provided written informed consent. Data from the adult MDD cohort has previously been reported [191, 192], while we have also reported cognitive data from the geriatric MDD cohort [183].

2.2. Clinical Assessments

For both studies, the DSM-IV-TR diagnosis of MDD was made using the Mini-International Neuropsychiatric Interview (MINI, version 5.0) [193] and confirmed by interview with a study psychiatrist. In all studies, participants were assessed for depression severity with the MADRS and medical burden with the geriatric Cumulative Illness Rating Scale (CIRS). In one but not the other two geriatric MDD studies, disability burden was measured using the World Health Organization Disability Schedule 2.0 (WHODAS 2.0) [194].

Using procedures similar to past reports [179, 192], we quantified age of initial depression onset and duration of depression using a life-charting approach in a detailed clinical interview, supplemented by acquisition of medical records. For the adult MDD cohort, this was for lifetime duration of depression. For the geriatric MDD cohort, this was limited to the current episode.

2.3. Cognitive Assessments

Participants completed a broad battery of neuropsychological tests that assessed cognitive domains relevant to depression or aging. As previously detailed [191, 195], we combined tests to create composite domain variables. We created z-scores for each measure based on the performance of all participants within each age cohort and averaged the z-scores for all tests within that domain. This resulted in a z-score for each domain for each participant. As previously published [191], for the adult cohort, tests in each domain included:

- Episodic Memory: Logical Memory 1 and 2 from the Wechsler Memory Scale, Benton Visual Retention Test, Rey's Verbal Learning Test (total I-V and total VII);
- Executive Function: Controlled Oral Word Association (COWA) test (letters: C, F, L), Trail Making B time (reverse scored), semantic fluency (Animal Naming), Stroop Color-Word interference condition;
- Processing Speed: Symbol-Digit Modality, Trail Making A time (reverse scored), Stroop Color Naming condition;
- Working Memory: Digit Span forward and Digit Span backward from the Wechsler Memory Scale.

Two of the geriatric studies used identical neuropsychological test batteries and so were included in analyses examining cognitive performance. As previously published [183], for the geriatric cohort, tests in each domain included:

- Episodic Memory: Word List Memory Recall (immediate and delayed), Paragraph Recall test, Constructional Praxis test, Benton Visual Retention Test;

- Executive Function: COWA test (letters: C, F, L), Trail Making B time (reverse scored), Stroop test color-word interference condition, Mattis Dementia Rating Scale, Initiation-Perseveration subscale;
- Processing Speed: Symbol-Digit Modality Test, Trail-Making A time (reverse scored), Stroop color naming condition;
- Working Memory: Digit Span forward, Digit Span backward, and Ascending Digits from the Wechsler Memory Scale;
- Language Processing: Stroop word reading condition, Boston Naming Test, Shipley vocabulary test.

2.4. MRI Acquisition

The adult cohort was imaged on a research-dedicated whole-body Siemens 3.0 T Trio Tim scanner at Duke University Medical Center using an 8-channel head coil. Parallel imaging was employed with an acceleration factor of 2. Duplicate sagittal MPRAGE sequences were obtained using a repetition time (TR) of 2300 ms, echo time (TE) of 3.46 ms, a flip angle of 9°, a 256 × 256 matrix, FOV 240 mm, 160 slices with a 1.2 mm slice thickness for voxel size of 0.9 × 0.9 × 1.2 mm.

The geriatric cohort was imaged on a research-dedicated 3.0T Philips Achieva whole-body scanner at Vanderbilt University Medical Center using a 32-channel head coil. The MPRAGE images were obtained using TR = 8.75ms, TE = 4.6ms, flip angle=9 degrees, and spatial resolution = 0.89 × 0.89 × 1.2 mm3 plus a FLAIR T2-weighted imaging conducted with TR = 10,000ms, TE = 125ms, TI = 2700ms, flip angle = 90 degrees, and spatial resolution = 0.7 x 0.7 x 2.0mm^3. FLAIR T2-weighted imaging was also conducted using TR = 10,000ms, TE = 125ms, TI = 2,700ms, flip angle = 90 degrees, and spatial resolution = 0.7 x 0.7 x 2.0mm^3.

2.5. MRI Analyses and Calculation of Brain Age

The brain age estimator is an automated deep learning tool used to predict or estimate age from a T1-weighted brain MRI. The first step in the brain age biomarker pipeline is to align the subject T1-weighted brain MRI with the MNI-305 template [196] using the affine registration from the NiftiReg library [197]. Images also undergo N4 bias field correction [198] to alleviate bias from acquisition. The input to the brain age estimation algorithm consists of the preprocessed brain MRI described above as well as the volume of 132 distinct regions of interest in the brain, obtained from a whole-brain segmentation using a multi-atlas technique [162]. The BAG algorithm described by Bermudez et. al, uses a deep convolutional neural network regression model trained on over 5,000 healthy controls ages 4 to 96 to predict age with high accuracy [125]. The innovation presented from this work is the addition of anatomical context in the form of volumetric estimates of regions of interest throughout the brain, which resulted in a more accurate prediction of age. The output is the estimated age for that subject, with the BAG biomarker being the difference between chronological true age and algorithm estimated age. For this study, we conducted model inference in our two cohorts using the BAG algorithm without any further model optimization or changes. Brain age calculations were performed on an NVIDIA GeForce Titan GPU with 12 GB memory and all deep learning algorithms were implemented and tested using Tensorflow v1.4 with a Keras backend v2.2. In order to analyze this cohort, we used a large-scale medical image processing infrastructure [199] and high performance computing cluster at Vanderbilt University.

The Lesion Segmentation Toolbox [200] was used to measure white matter hyperintensity (WMH) volumes, findings on T2- weighted or FLAIR images related to cerebral ischemia. These analyses were implemented through the VBM8 toolbox in SPM8 and have been previously described [183, 201]. This lesion map is then used to calculate total cerebral WMH volume.

2.6. Analytic Plan

Statistical analyses were conducted using SAS Studio 3.8 (SAS Institute, Cary, NC). Participant demographics within each cohort were summarized and univariate comparisons conducted using pooled,

two-tailed t-tests for continuous variables and chi-square tests for categorical variables. Data were graphed to facilitate identification of outliers and one geriatric participant exhibited cognitive domain z-scores several standard deviations lower than the rest of the geriatric cohort. This individual was excluded from analyses.

The primary imaging measure was the BAG, calculated as the difference between the algorithm-determined estimated age and the chronological age. A negative BAG indicated that the brain appeared younger than anticipated based on chronological age, while a positive BAG indicated a brain appearing older than anticipated (Figure IV.1). As we anticipated that the BAG might exhibit more variability in the geriatric cohort than the midlife cohort, we included chronological age as a covariate in statistical models.

Statistical analyses used general linear models (PROC GENMOD) using a similar approach for both age cohorts. Initial models tested for diagnostic group differences in BAG, including covariates of chronological age, sex, education, and medical morbidity by CIRS. This was followed by examining the effect of BAG on z-scored cognitive domains, including covariates of diagnostic group, chronological age, sex, education, and medical morbidity measured by CIRS. As part of these analyses, we tested for a

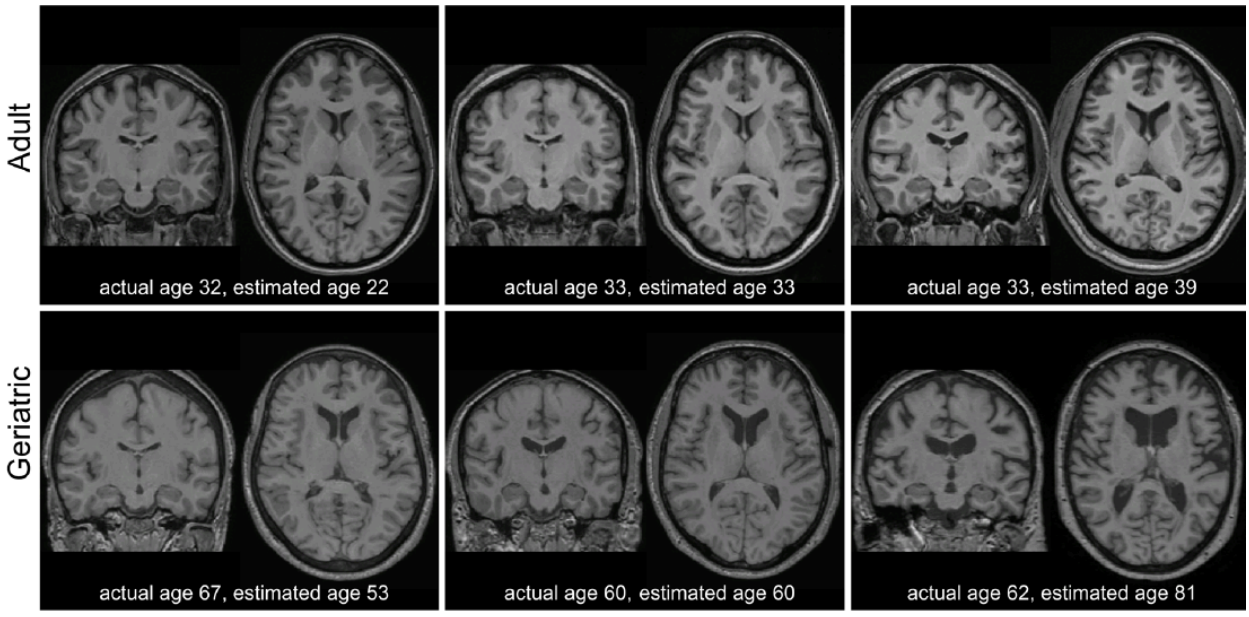

Figure IV.1 Comparison of structural MRI of participant brains in mid-life and older adult cohorts. Each image is a separate participant, displaying coronal and axial images and actual (chronological) age and estimated (calculated) age. The top row is from the midlife adult cohort and the bottom row is from the geriatric cohort.

statistical interaction between diagnostic group and BAG affecting cognitive domain score. Similar approaches were used to assess the effects of BAG on disability measured by the WHODAS 2.0.

Subsequent models focused on effects of depression history or depression severity, so included only depressed subjects. These models included covariates of chronological age, sex, education, CIRS, and depression severity by MADRS. We first tested for the relationship between BAG and depression exposure, examined both as age of onset of the initial depressive episode and duration of depression, calculated as lifetime exposure for the adult MDD group and duration of current episode for the geriatric MDD group. We next examined the effect of BAG on z-scored cognitive domain performance in the depressed groups alone, also testing for a statistical interaction between depression severity and BAG.

As WMHs are common in geriatric MDD, associated with aging [202, 203], and a potential marker of accelerated cerebrovascular aging, we examined the relationship between WMH volume, BAG, and clinical measures. As WMH volumes are often not normally distributed, our primary measure was log-transformed WMH volume. These exploratory analyses examined the same models as detailed above but included transformed WMH volume as an additional covariate.

3. Results

3.1. Brain Age Analyses in the Midlife Adult Sample

The adult cohort included 76 depressed and 94 non-depressed adults (Table IV.1). Depressed participants were significantly older than never-depressed participants in chronological age and estimated age but exhibited a comparable BAG. Depressed participants exhibited significantly higher medical comorbidity severity (via CIRS) and, in univariate analyses, poorer episodic memory and processing speed performance. The depressed cohort's mean age of onset was 20.8 years (range 6-35), with mean lifetime depression duration of 2,115 days (5.8 years; range 90-7500 days).

After controlling for covariates (chronological age, sex, education, CIRS) BAG did not differ between depressed and non-depressed participants (Wald X^2=0.40, 164 df, p=0.5294). BAG was also not

	Adult Sample				Geriatric Sample			
	Depressed (N=76)	**Never-Depressed (N=94)**	**test value**	**p-value**	**Depressed (N=118)**	**Never-Depressed (N=36)**	**test value**	**p-value**
Age, years (chronologic)	36.21 (9.04)	30.14 (9.2)	-4.3	<0.0001	66.41(5.45)	70.06(6.65)	3.33	0.0011
Age, years (calculated)	43.67 (11.27)	37.48 (10.09)	-3.77	0.0002	70.09(8.12)	68.83(10.59)	-0.66	0.5117
Brain-age gap (BAG)	7.46 (7.56)	7.33 (5.54)	-0.12	0.9027	3.69 (7.16)	-1.23 (7.62)	-3.55	0.0005
Sex, % female (N)	68.42 (52/76)	61.7 (58/94)	0.83	0.3621	61.0 (72/118)	55.6 (20/36)	0.34	0.5586
Education, years	15.34 (2.43)	15.69 (2.06)	1.01	0.3123	16.79 (2.21)	17.17 (1.93)	0.92	0.3573
CIRS	0.67 (1.16)	0.31 (.75)	-2.36	0.0199	5.45 (3.22)	4.72 (2.77)	-1.22	0.2234
MADRS	23.62 (4.33)	0.80 (1.16)	-44.7	<0.0001	26.21 (5.20)	0.75 (1.02)	-50.11	<0.0001
					N=103	N=36		
MMSE					29.20 (1.1)	29.3 (1.1)	0.23	0.8186
Episodic memory	-0.68 (4.24)	0.85 (3.68)	2.49	0.0137	-0.03 (0.73)	0.15 (0.74)	1.27	0.21
Executive function	-0.36 (3.05)	0.46 (2.92)	1.79	0.0754	-0.08 (0.69)	0.29 (0.51)	3.33	0.0013
Processing speed	-0.51 (2.49)	0.51 (2.20)	2.82	0.0053	-0.05 (0.72)	0.30 (0.54)	3.03	0.0033
Working memory	-0.21 (1.67)	0.20 (1.88)	1.49	0.1370	-0.02 (0.82)	0.15 (0.81)	1.12	0.2632
Language Function					-0.004 (0.65)	0.06 (0.50)	0.49	0.6246
					N=85	N=15		
WHODAS					23.91 (14.82)	4.51 (4.05)	-10.21	<0.0001

Table IV.1 Demographic and clinical differences across samples. Data presented as mean (standard deviation) for continuous variables and percent (N) for categorical variables. Analyses used pooled, two-tailed t-tests for continuous variables and chi-square tests for categorical variables. The exceptions requiring the use of Satterthwaite t-tests due to unequal variances for the adult sample included analyses of CIRS (122.8 df) and MADRS (83.7 df) and for the geriatric sample MADRS (141.05 df), calculated age (48.2 df), processing speed (87.7 df), and WHODAS score (82.0 df). Adult sample pooled t-tests had 168 degrees of freedom. For the geriatric sample, for the overall demographics df=152, for the cognition sample df= 137, and for the disability sample df=98. CIRS = Cumulative Illness Rating Scale; MADRS = Montgomery-Asberg Depression Rating Scale; MMSE = Mini-Mental State Examination; WHODAS = World Health Organization Disability Assessment Schedule (Version 2.0).

significantly associated with episodic memory (Wald X^2=1.56, 163 df, p=0.2112), executive function (Wald X^2=1.94, 163 df, p=0.1637), processing speed (Wald X^2=0.01, 163 df, p=0.9210), or working memory (Wald X^2=0.02, 163 df, p=0.8837.) Tests for an interactive effect between MDD diagnosis and BAG on cognitive performance were not statistically significant, thus the relationship between BAG and cognitive performance did not appear to differ based on a diagnosis of MDD (data not shown).

In analyses of depressed participants only, depression severity by MADRS was not significantly associated with BAG (Wald X^2=0.25, 70 df, p=0.6141). We further did not observe significant relationships between BAG and age of onset (Wald X^2=0.06, 68 df, p=0.8098) or lifetime duration of depression (Wald X^2=0.35, 68 df, p=0.5515). In depressed participants alone, there were neither significant direct effects of

BAG nor interactive effects between MADRS and BAG on cognitive domain performance (data not shown).

3.2. Brain Age Analyses in the Geriatric Sample

The geriatric cohort included 118 depressed and 36 never-depressed elders (Table IV.1). Compared to non-depressed participants, depressed elders were younger with a lower mean chronological age but exhibiting comparable estimated age. This resulted in depressed elders exhibiting a significantly higher BAG. Depressed elders exhibited poorer performance on unadjusted measures of executive function and processing speed. For depressed elders, the mean age of onset was 34.7 years (range 5-84 years), with a mean current episode duration of 953 days (2.6 years, range 15-5141 days).

After adjusting for chronological age, sex, education, and CIRS, BAG was significantly higher in depressed elders (Wald X^2=8.84, 148 df, p=0.0029) indicating that brains of depressed participants appeared older than expected by chronological age alone. After adjusting for covariates including diagnosis, a higher BAG was associated with poorer episodic memory performance (Wald X^2=4.10, 132df, p=0.0430; Figure IV.2a) but not executive function (Wald X^2=0.03, 132df, p=0.8643), processing speed (Wald X^2=2.78, 132df, p=0.0957), working memory (Wald X^2=0.00, 132df, p=0.9974), or language function (Wald X^2=0.02, 132df, p=0.8877). Tests for an interactive effect between MDD diagnosis and BAG on cognitive performance were not statistically significant (data not shown).

Examining depressed elders only, there was no statistically significant relationship between BAG and depression severity by MADRS (Wald X^2=0.96, 112 df, p=0.3271). We also did not observe significant relationships between BAG and either age of onset (Wald X^2=0.31, 110 df, p=0.5769), or duration of current episode (Wald X^2=0.05, 110 df, p=0.8225). In models examining depressed elders only (N=103), controlling for age, sex, education, CIRS and MADRS, there was a primary effect of BAG on processing speed (Wald X^2=4.43, 96 df, p=0.0354; Figure IV.2b) but not episodic memory (Wald X^2=1.07, 96 df, p=0.2999), executive function (Wald X^2=0.30, 96 df, p=0.5836), language (Wald X^2=1.05, 96 df,

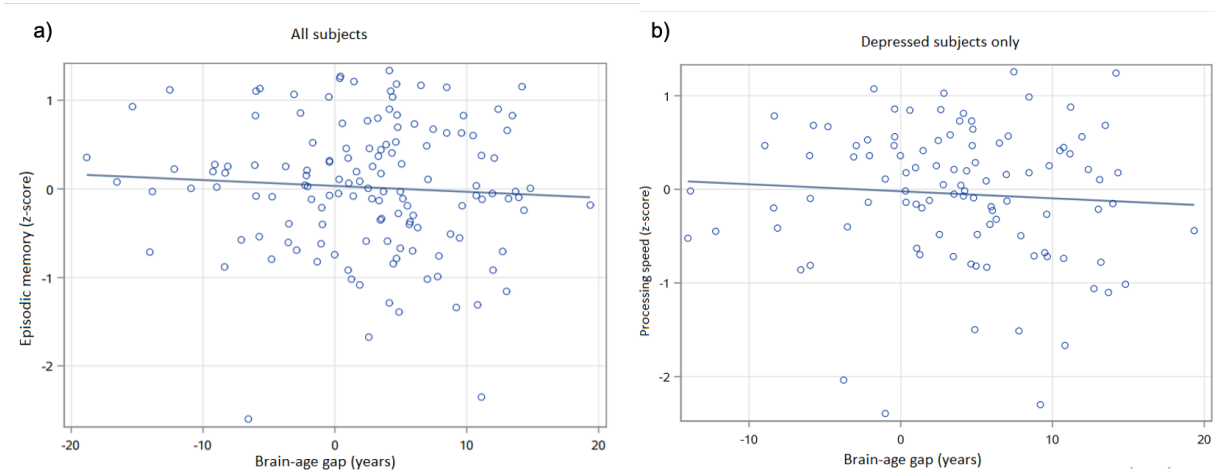

Figure IV.2 Association between Brain Age Gap and cognitive testing. a) Association between brain-age gap and episodic memory in geriatric subjects. Episodic memory has no units, presented as an average z-score across tests. Brain-age gap (BAG) is in years, calculated as the difference between the calculated estimated age and the chronological age. **b) Association between brain-age gap and processing speed in older depressed subjects.** Processing speed has no units, presented as an average z-score across tests. Brain-age gap (BAG) is in years, calculated as the difference between the calculated estimated age and the chronological age.

p=0.3056), or working memory (Wald X^2=1.00, 96 df, p=0.3164). We further observed statistically significant interactive effects between MADRS and BAG on executive function (Wald X^2=5.89, 95df, p=0.0152) and working memory (Wald X^2=4.47, 95 df, p=0.0346). In these analyses, a greater BAG had an increasingly negative effect on executive function and working memory in context of worsening MADRS score. In other words, the effect of a higher BAG (or having an older-appearing brain than expected) on executive function and working memory performance is greater in context of more severe depressive symptoms. No significant interaction effects were observed between MADRS and BAG on episodic memory (Wald X^2=0.98, 95 df, p=0.3213), processing speed (Wald X^2=0.04, 95 df, p=0.8448) or language function (Wald X^2=0.04, 95df, p=0.8346).

Disability data measured by WHODAS was available from one study, consisting of data from 85 depressed and 15 never-depressed older participants. After adjusting for covariates, BAG was associated with greater disability (Wald X^2=6.00, 93 df, p=0.0143; Figure IV.3). We did not observe a statistically significant interactive effect between BAG and MDD diagnosis on disability.

3.3. Effect of EMH on the BAG and its Clinical Correlates in Geriatric MDD

As greater severity of WMH is also an aging-related marker of vascular damage in geriatric MDD, in secondary analyses we examined the relationship between log-transformed WMH volume, BAG, and diagnosis. When adding log-transformed WMH volume to the models described above, WMH was positively associated with BAG (Wald X^2=7.01, 147 df, p=0.0081), but this did not appreciably change the relationship between MDD diagnosis and BAG (Wald X^2=7.00, 147df, p=0.0082).

We next examined whether the addition of log-transformed WMH volume to models examining cognitive performance and disability changed the results described above. The addition of WMH did not appreciably change our observed associations between BAG, cognitive performance, or disability (data not shown), except for the interactive effect observed in the geriatric depressed cohort between BAG and MADRS on executive function. Despite WMH not being significantly associated with executive function

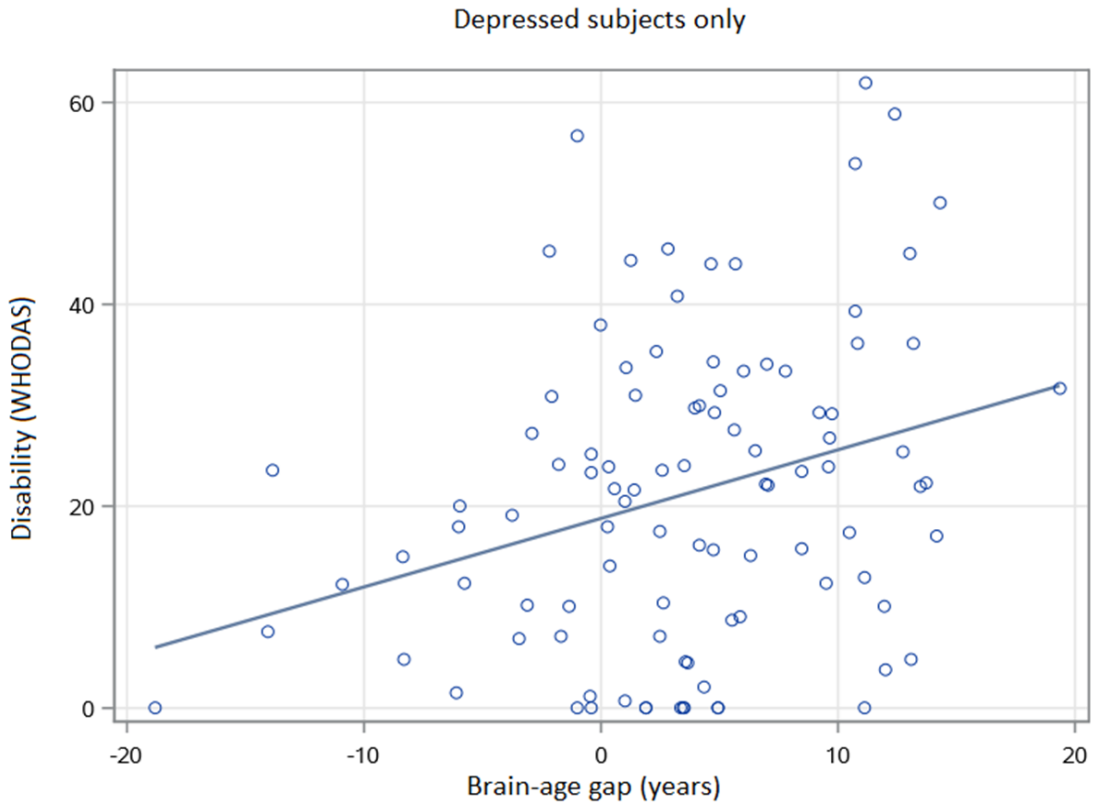

Figure IV.3 Association between brain-age gap and disability (WHODAS) in older adults. Disability measured by the WHODAS 2.0, calculated as percent disabled. Brain-age gap (BAG) is in years, calculated as the difference between the calculated estimated age and the chronological age.

in this model (Wald X^2=0.99, 94 df, p=0.3202), the interaction between BAG and MADRS was also no longer statistically significant (Wald X^2=0.28, 94 df, p=0.5979).

4. Discussion

Differences in brain aging can be observed comparing chronological age with estimated calculated age using the BAG metric (Figure IV.1, video displaying geriatric MRI data with both chronological and estimated ages at: https://vimeo.com/393048773). We observed marked differences in the clinical implications of the difference between estimated brain age and chronological age between younger and older cohorts. In the young to middle-aged adult cohort, BAG did not differ by diagnosis and we observed no significant relationships between BAG and cognition. In contrast, depressed elders exhibited a higher BAG, indicating that the estimated age based on structural MRI was higher than expected. This higher BAG was associated with poorer episodic memory performance and greater disability. In depressed elders only, higher BAG was further associated with slower processing speed while a higher BAG exhibited an interactive effect with depression severity on executive function and working memory performance. Greater WMH volume, another MRI finding associated with geriatric depression and aging [202, 203], was associated with a higher BAG, but largely did not change observed relationships between BAG and cognition or disability.

Our findings in adult MDD are generally concordant with some past work in this population [175], suggesting that accelerated brain aging is not prominent in midlife adults even in context of recurrent depressive episodes. We now extend this approach to older adults. To our knowledge, this is among the first reports to use a deep learning approach to examine a brain-based biomarker of accelerated aging in geriatric MDD. The difference in findings between younger and older cohorts suggests that over the lifetime, at some point depressed subjects may diverge from never-depressed individuals and the aging process accelerates. Although we did not observe any diagnostic group differences in the BAG in the younger cohort, the difference we observe in older depressed subjects must have started at a younger age. It remains unclear whether such a process might be linear or if depressed and nondepressed groups diverge

in a specific midlife window. We may also not have had sufficient statistical power to detect smaller between-group differences earlier in the aging process. Notably, despite the implication that the depressed group exhibits accelerated brain aging over time, we did not find support for the hypothesis that longer depression exposure is associated with increased BAG. However, our findings may be constrained by using retrospective measures to assess duration of depression that are based primarily on remote patient recall.

Beyond diagnostic group differences, in geriatric MDD BAG is associated with cognition and disability. The association between increased BAG and poorer episodic memory is consistent with work showing episodic memory to is particularly vulnerable even to normal aging [204]. Episodic memory is the hallmark domain affected in mild cognitive impairment and shows steeper decline in pathological aging such as in Alzheimer's dementia. Similarly, other biomarkers of accelerated aging are associated with worse episodic memory performance in neuropsychiatric populations, including decreased telomere length and gray matter volume [205]. Conversely, in older adults, lower epigenetic age calculated from DNA methylation was predictive of intact episodic memory [206]. The primary effect of BAG on processing speed in depressed elders is also consistent with past reports of decreased processing speed in geriatric MDD [207, 208] that may serve as a core cognitive impairment, mediating deficits in working memory and verbal capabilities [207, 208]. Intriguingly, we found an interaction between increased BAG and depression severity on executive function and working memory, although the effect on executive function did not persist after adjusting for WMH volume. It is well established that active depressive symptoms can worsen cognitive performance, even with treatment [209]. Our current findings suggest that the effect of accelerated brain aging on executive function and working memory may be mediated by depression severity, or that an individual with an older-appearing brain may exhibit greater decline in these cognitive domains as their depression severity worsens. This may help explain variability in executive processes in late-life depression, and furthermore why performance may improve with successful antidepressant treatment. Such a "two-hit" hypothesis of an older-looking brain becoming increasingly vulnerable to the cognitive effects of a depressive episode is consistent with clinical experience.

Our observed association of higher BAG with greater disability is supported by past work in older adults. Repeatedly, biomarkers of accelerated brain aging in both gray matter and the white matter are associated with increased disability and impairment in activities of daily livings. These findings include associations between disability and hippocampal volume loss, cortical gray matter changes, WMH, infarcts, and other measures of white matter microstructure [210-212]. Our observation utilizing a BAG measure derived from structural differences, is concordant with this past work. It also identifies biological contributors to disability in geriatric MDD that extend beyond severity of depressive symptoms.

A strength of the study includes two large cohorts across the adult lifespan. Limitations include the examination of cross-sectional rather than longitudinal data, limiting our ability to make causal inferences or determine whether BAG measures may have prognostic utility in predicting cognitive decline or worsening disability. We additionally did not have the ability to look at potentially protective behaviors that may be positively associated with a negative BAG, where brain age is younger than expected, such as physical activity or aerobic exercise.

Although the size of our cohorts and the age range examined is a strength, there are additional limitations to our approach. First, we cannot combine neuroimaging data across cohorts. MRI data from each study were gathered at different sites using different scanners, so cohort age is confounded with scanner parameters. Additionally, while our participants include younger and older adults, we have an age gap between the studies, with no participants between the ages 50-59 years. These factors make it difficult to suggest where a divergence in estimated brain age and chronological age may be observed between younger and older depressed adults. These issues are highlighted by differences in the accuracy in age prediction between the two cohorts. We saw that the average BAG in the midlife diagnostic groups about 7 years, whereas the average BAG in the geriatric cohort was 3.7 years for depressed and -1.2 years for never-depressed subjects (Table IV.1). The model of age prediction used in this work relies on patterns in image intensity from the T1-weighted brain MRI to predict age. Deep neural networks are susceptible to small, systemic changes in image intensity [126], so site or scanner effects, such as the differences between the two cohorts in this study, may bias age prediction. However, true age and predicted age are highly

correlated, suggesting strong associations between these values. For these reasons, we analyze the BAG biomarker in the context of each study separately as a comparative biomarker between patients with depression and healthy controls. Observed differences in brain age prediction between diagnostic groups will thus be due to clinical differences instead of bias due to site effect, despite reducing the possible sample size if both cohorts were analyzed together.

Future work should continue to examine the clinical significance of measuring BAG and whether BAG may be predictive of short- or long-term outcomes such as acute antidepressant response or risk of recurrence following remission [177]. It should also be determined whether it has long-term prognostic value to identify individuals at increased risk of cognitive decline. Such work would be valuable in geriatric MDD but also early in the course of neurodegenerative disorders such as Alzheimer's disease. This longitudinal work should examine BAG not only as a cross-sectional predictor, but also how BAG changes with aging, and whether such change trajectories may be more informative than cross-sectional assessments alone. Although this study does not support brain age as a clinically useful marker of accelerated aging in midlife adults with MDD, it does support a potential role for disorders of aging.

V. Automated Dentate Nucleus Segmentation and Its Clinical Application in Movement Disorders

1. Introduction

The dentate nucleus (DN) is a gray matter structure deep in the cerebellum, which is involved in motor coordination, sensory input integration, executive planning, language, and visuospatial function [90]. The DN is believed to be an important target for network-based functional and structural connectivity studies [90]. DN segmentations have been previously used to assist in planning of deep brain stimulation surgical approaches that stimulate the DN directly [213, 214]. In these applications, stimulation of related white matter tracts, like the dento-rubro-thalamic tract (DRTT), may alleviate tremor, or improve post-stroke neuroplasticity and remodeling [166, 213, 215]. Changes in the DN and the DRTT resulting in cerebellar dysfunction have been previously linked to the pathophysiology of essential tremor (ET) and PD [91-95]. These two disorders of movement affect over 4% and 1% of the population over 60 years old, respectively [216]. Nicoletti et al. used magnetic resonance imaging (MRI) methods to detect reductions in fractional anisotropy and elevated mean diffusivity in the DN of patients with familial ET relative to PD patients and controls [217]. While existing literature has demonstrated the importance of the DN in neurological disease, no highly reproducible multimodal method of automatic segmentation exists for this key structure. Establishing an accurate and reliable method to define the DN will allow for large-scale studies and a better understanding of this structure in the evolution of disease.

Previous work has used manual or automatic segmentation of the DN in MRI as the seed region to initiate MRI processing pipelines such as white matter tractography [96-98] or resting state functional connectivity [218-220]. DN segmentation can also be used for volumetric analysis. Solbach et al. found a decrease in DN and cerebellum volumes of patients with Friedreich's ataxia compared to healthy controls [221]. This study used susceptibility-weighted MRI acquired in a 7 Tesla scanner to visualize and delineate the DN. Unfortunately, manual segmentations of the dentate on large datasets can be resource-prohibitive,

making automatic segmentations an attractive alternative in the field of medical image processing and clinical imaging biomarkers [222-224]. A critical impediment in imaging is the lack of accurate and broadly available methods to quantify the volume of deep cerebellar structures, such as the DN.

Prior attempts at automatic segmentation of the DN have largely applied single atlas [90, 100, 225, 226]. Atlas-based segmentation relies on a reference anatomical image and a set of labels, where every voxel in the anatomic image has been manually labelled [227]. Unfortunately, single-atlas segmentation has a limited ability to capture the wide variability between subjects [228]. Moreover, the DN and other deep cerebellar nuclei are difficult to visualize with MRI due to their small size and convex structures [90]. High magnetic fields (7 Tesla) or specific MRI sequences such as T2*-weighted or susceptibility weighted imaging (SWI), are required to provide enough contrast to visualize these structures with millimetric resolution. High field strength and SWI methods have achieved a mean overlap of only 30-50% between the automatic segmentation of the DN and a manual tracing [2, 99, 100].

Deep learning methods have shown remarkable improvements on image processing tasks including segmentation, lesion detection, and computer-assisted diagnosis [38]. In particular, the U-Net [48] has been shown among the highest accuracy in 2D and 3D segmentation tasks [37]. Deep learning networks tend to require large sets of labelled training data to achieve sufficient performance and avoid overtraining. However, increasing labeled dataset size is a key challenge for medical imaging applications, since producing high-quality labels is resource-intensive. Creating additional training data is a problem for structures that are difficult to visualize, like the DN, since these require acquisition parameters that are often specific to a given research study. The development of a reliable segmentation pipeline using common clinical MRI modalities such as T1, T2, and diffusion-weighted imaging (DWI), would provide a useful alternative strategy for post-processing studies. The goal of this work is to produce an accurate and reliable segmentation of the DN across common clinical imaging modalities. The work presented in this chapter was published in the *Journal of Medical Imaging* with Camilo Bermudez as first author.

2. Methods

2.1. Imaging Datasets and Manual Labelling

We used two different multimodal datasets obtained under IRB approval: one of healthy subjects and one of subjects diagnosed with PD or ET. The demographics of the three cohorts are shown in Table V.1.

The first dataset consists of 83 healthy controls across a broad adult age range with T1-weighted, fluid-attenuated inversion recovery (FLAIR), and diffusion tensor imaging (DTI). All participants provided written informed consent [229]. All modalities were acquired on a 3T Phillips Scanner (Phillips Healthcare Inc.). T1-weighted imaging was acquired with a magnetization-prepared sequence and rapid gradient echo sampling (MP-RAGE) (TR/TE = 8.9/4.6 ms, FOV=256x256x170) at 1 mm isotropic spatial resolution. FLAIR imaging (TR/TE = 4000/120 ms; FOV = 256x156x176) was acquired at 1 mm isotropic spatial resolution. Diffusion weighted imaging was acquired with 32 diffusion directions (TR/TE=10000/60 ms; b-val=1000 s/mm2; FOV:240x240x50) at a 2.5 mm isotropic spatial resolution.

	Controls (N = 83)	Essential tremor (N =38)	Parkinson's disease (N = 60)
Age (years)	48.71 +/- 17.47	**67.29 +/- 5.54**	**63.61 +/- 8.69**
Gender ratio (M:F)	32:51	18:20	37:23
Dentate nuclei volume (mm^3)	716.5 +/- 203.3	**896.9 +/- 202.3**	**904.7 +/- 195.2**
Posterior fossa volume (cm^3)	171.3 +/- 18.1	**182.5 +/- 15.0**	**185.9 +/- 20.4**

Table V.1 Demographic data for all three cohorts. Age and volumes are presented as mean +/- standard deviation. Bold represents a statistically significant difference between the disease cohorts and healthy controls, $p<0.025$ using Mann-Whitney test with Bonferroni corrections.

The second dataset consists of 98 subjects diagnosed with either PD or ET who provided written informed consent. These subjects were recruited from the Vanderbilt University Medical Center (VUMC) patient population under consideration for deep brain stimulation surgery. All patients underwent evaluation by a neurologist and a neuropsychiatrist at VUMC as part of preoperative to rule out other movement or psychiatric disorders. All subjects underwent T1-weighted imaging, T2-weighted imaging, and DTI, which were acquired on a 3T Phillips Scanner. (Phillips Healthcare Inc.). T1-weighted imaging (TR/TE = 7.9/3.8 ms, 256x256x170 voxels) was acquired with a SENSE parallel imaging sequence (T1W/3D/TFE) at 1 mm isotropic spatial resolution. T2-weighted imaging (TR/TE = 3000/80 ms, FOV = 512x512x45) was acquired

using a SENSE parallel imaging technique (T2W/TSE) at 0.47-by-0.47-by-2 mm spatial resolution. Diffusion tensor imaging was acquired with 32 diffusion directions (TR/TE=10000/60 ms; b-val=1000 s/mm2; FOV = 128x128x60) at a 2 mm isotropic spatial resolution.

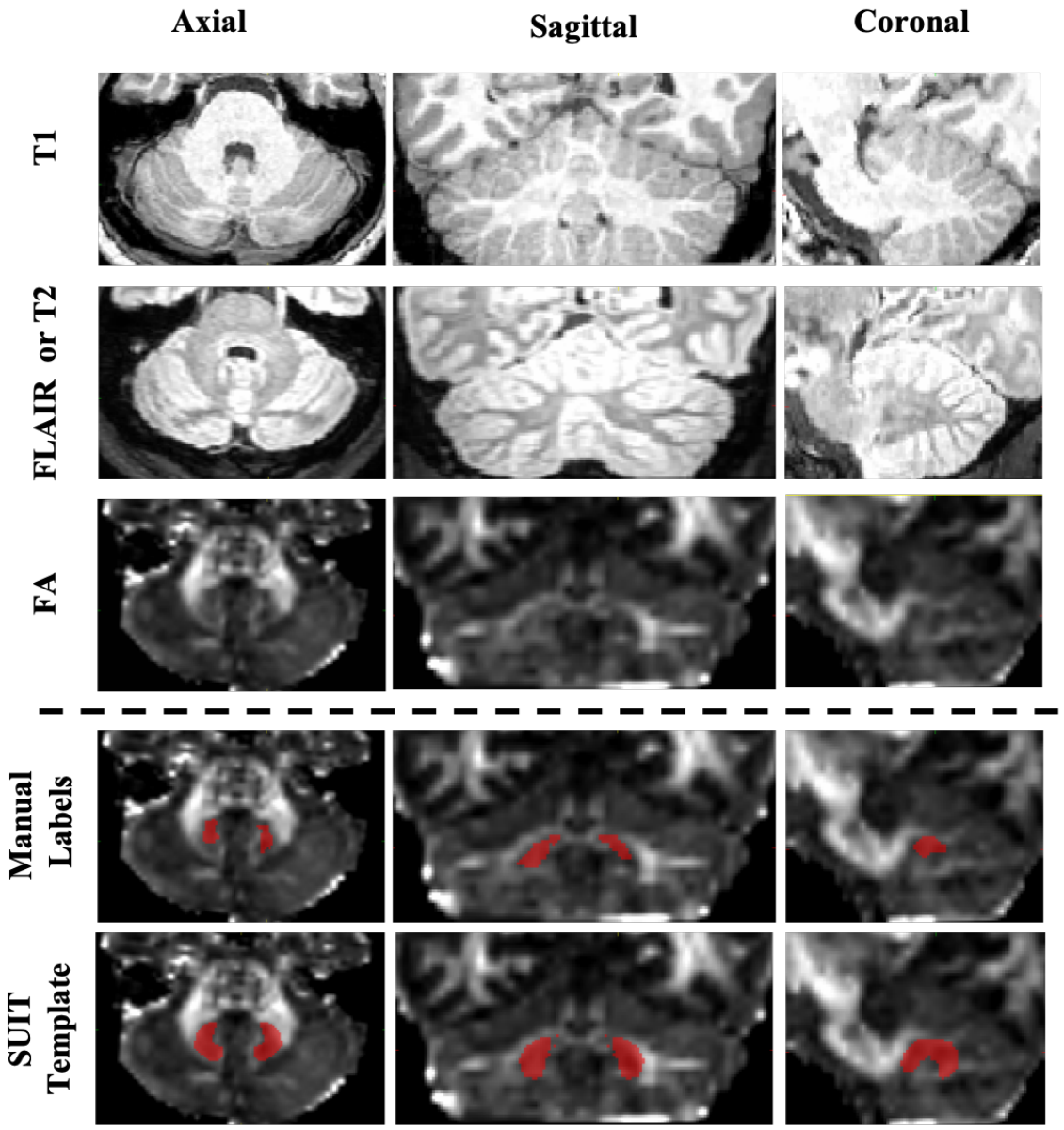

Figure V.1 Visualization of the dentate nucleus in the cerebellum across several modalities. MRI modalities used in this study, including T1-weighted, T2-weighted (T2) or FLAIR (T2-F), and FA maps derived from diffusion tensor imaging. All images are shown registered to the SUIT template. Dentate labels were manually traced on the FA maps due to ease of visualization. For comparison, we also show the SUIT template labels for the dentate nucleus on the bottom row [2].

2.2. Image Preprocessing

Diffusion imaging underwent eddy current and motion corrections using the FDT toolbox in the FMRIB software library (FSL). This toolbox was also used to calculate diffusion tensors and fractional anisotropy (FA) maps were computed for manual labeling of the DN [230]. The DN appears as a hypointense area on FA maps surrounded by cerebellar white matter. The left and right DN were manually labelled by the same expert rater on all subjects in both datasets using the FA maps due to easier visualization on both datasets (Figure V.1). The rater labelled any contiguous hypointense voxels surrounded by hyperintense cerebellar white matter. All algorithms were trained on these manual labels as truth. The volumes of the DN was calculated by multiplying the number of marked voxels in the DN mask by the resolution of FA maps. The average volume of the dentate nuclei was 716.5 +/- 203.3 mm3 for the healthy controls, 896.9 +/- 202.3 for the ET cohort, and 904.7 +/- 195.2 for the PD cohort (Table V.1). These measurements are consistent with volumetric estimates in the literature, which report similar volumes for healthy controls and PD subjects(35). There was a significant difference in age between both disease cohorts and the healthy controls ($p < 0.001$ on both, Mann-Whitney test with Bonferroni corrections). The manual labels for the DN in both ET ($p<0.001$) and PD ($p = 0.002$) are significantly larger than those of healthy controls. This difference persisted after normalizing the DN volume by the posterior fossa volume for ET ($p = 0.0012$) and PD cohorts ($p = 0.0029$).

The initial step in preprocessing the images was to co-register each modality to a single modality in subject space. The FA maps and the corresponding labels were up-sampled to 0.83 mm isotropic spatial resolution in the healthy dataset and to 1 mm isotropic spatial resolution in the PD/ET dataset in order to approximate the resolution to that of the T1 and T2 modalities and improve registration accuracy. Afterwards, T1 and FLAIR/T2 were skull-stripped and registered to the FA maps using an affine transformation in order to align all modalities within each subject using the publicly available software NiftiReg. The affine transformation algorithm reg_aladin is based on the work by Ourselin et al (36), which uses normalized cross-correlation as a loss function for registration accuracy. Lastly, all imaging modalities and labels were registered to the SUIT cerebellar template space using an affine transformation, so all

subjects would be aligned (1). The SUIT template from the Neuroimaging & Surgical Technologies Lab was used at a resolution of 1 mm3 isotropic voxels and a matrix size of 172 x 220 x 156 mm. Visual QA of the registered image of the SUIT template was done to ensure accurate registration. Each modality was intensity-standardized to have a zero mean and variance of one. For multimodal experiments, the 3D volumes were stacked along the fourth dimension to produce a single image for each subject.

2.3. Segmentation Network Architecture

The segmentation network consists of a four-layer 3D U-Net [48] with increasing filter size and feature concatenation at each layer. The convolutional kernel had a size of 5 voxels. The number of filters at each level of the U-Net was 8, 16, 32, 64, and 128 with Batch Normalization and ReLU activation. Max Pooling of 2 voxels was used in the encoding portion and 3D upsampling of 2 voxels was used in the decoding portion. A graphical representation of the network architecture can be seen in Figure V.2. All networks were trained at a learning rate was 1x10-5 using the Adam optimizer (37). Four separate networks with identical architecture were trained: 3 for each modality (T1, T2 or FLAIR, FA) and one multimodal network (T1 + T2 or FLAIR + FA). The network takes the image volume as input and produces the label as output. All networks were trained using the Dice Similarity Coefficient (DSC) between the predicted label and the manual label of a single rater as the cost function [37]. DSC is the proportion between twice the intersection between a manual label M and an automatic label A over the total number of voxels between both labels, defined mathematically as:

$$DSC = \frac{2\,|A \cap M|}{|A| + |M|} \qquad \text{(Equation V.1)}$$

All models were trained with an Adam optimizer for 500 epochs, when the DSC on the validation set was no longer changing more than 1% over 20 epochs. All networks were written using Tensorflow version 1.5 and Keras version 2.2.4, and were trained on an NVIDIA Titan Graphics Processing Unit (GPU). All trained models and analysis code will be made publicly available on (https://github.com/MASILab). The raw

imaging datasets will not be made available due to the data usage agreement that precludes data redistribution.

To further validate that the networks are learning the appropriate features, we use the Gradient Class Activation Maps (Grad CAM) method described by [55] and implemented by the publicly available library *keras-vis* [139] to visualize the areas of the raw MRI with higher attention in the combined model.

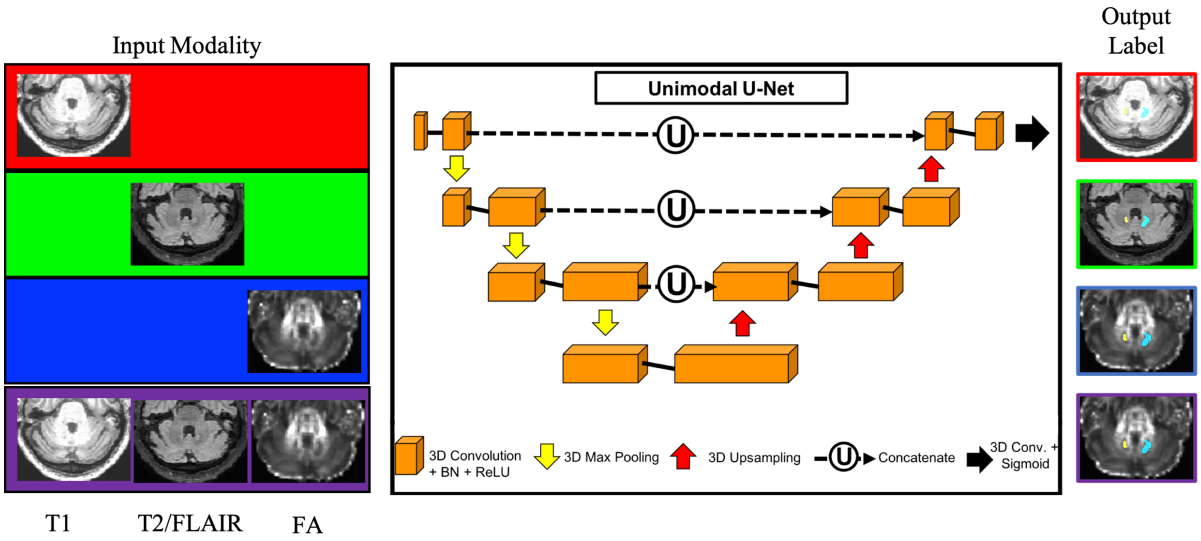

Figure V.2 Pipeline for multimodal segmentation of the dentate nucleus. A modified U-Net was used to segment the dentate nucleus. A total of four networks were trained: three unimodal (T1, T2/T2-FLAIR, or FA) and one multimodal (T1 + T2/T2-FLAIR + FA).

2.4. Statistical Analysis for Segmentation Models

All models were trained with 80% of each of the subjects as the training set, while the remaining 20% was evenly split between validation and testing. A five-fold cross-validation scheme was employed in independently trained models. Model accuracy was evaluated using three metrics of label quality: DSC, Mean Surface Distance (MSD), and Hausdorff Distance (HD). Surface error metrics are used as complementary measurements for the accuracy of the segmentation. All models were compared using Wilcoxon signed-rank test with Bonferroni corrections. Significance was established at $p < 0.0125$ due to four comparisons.

3. Results

We wish to find which input imaging modality can generate the best automatic segmentation. The automatic segmentation models for each input were evaluated on the testing set of each cross-validation fold. We also perform a comparison against single atlas registration using the SUIT template. Table V.2 shows that there was small agreement between the SUIT template and the manual labels drawn on the FA maps. Of the unimodal U-Net models, the FA maps performed best, with a mean DSC of 0.83 +/- 0.22. Including all modalities as separate channels in a multimodal network, did not improve the DN segmentation, achieving a lower DSC of 0.80 +/- 0.23 ($p < 0.001$). A representative example of a segmentation for each method is shown in Figure V.3.

	Dice Coefficient		Mean Surface Distance (mm)		Hausdorff Distance (mm)	
	Mean +/- SD	Median	Mean +/- SD	Median	Mean +/- SD	Median
Unimodal T1 U-Net	0.76 +/- 0.26 (p<0.001)	0.89	0.85 +/- 2.91 (p<0.001)	0.35	3.00 +/-3.48 (p<0.001)	2.20
Unimodal T2 U-Net	0.79+/- 0.26 (p<0.001)	0.93	0.84 +/- 3.63 (p<0.001)	0.24	3.05 +/- 3.63 (p=0.0328)	2.00
Unimodal FA U-Net	0.83 +/- 0.22 (Ref.)	0.95	0.66 +/- 2.71 (Ref.)	0.18	2.58 +/- 3.24 (Ref.)	1.41
Multimodal U-Net[+]	0.80 +/- 0.23 (p<0.001)	0.93	0.78 +/- 2.84 (p<0.001)	0.25	2.88 +/- 3.51 (p<0.001)	1.87
SUIT Atlas Registration	0.23 +/- 0.09 (p<0.001)	0.23	1.92 +/- 2.14 (p<0.001)	1.71	5.12 +/- 2.45 (p<0.001)	4.47

Table V.2 Mean and median metrics of segmentation quality for segmentation networks with different inputs.
[+]The unimodal FA network is used as reference for statistical comparisons. Wilcoxon rank sum test corrected for multiple comparisons using Bonferroni corrections. Significance was established as $p < 0.0125$.

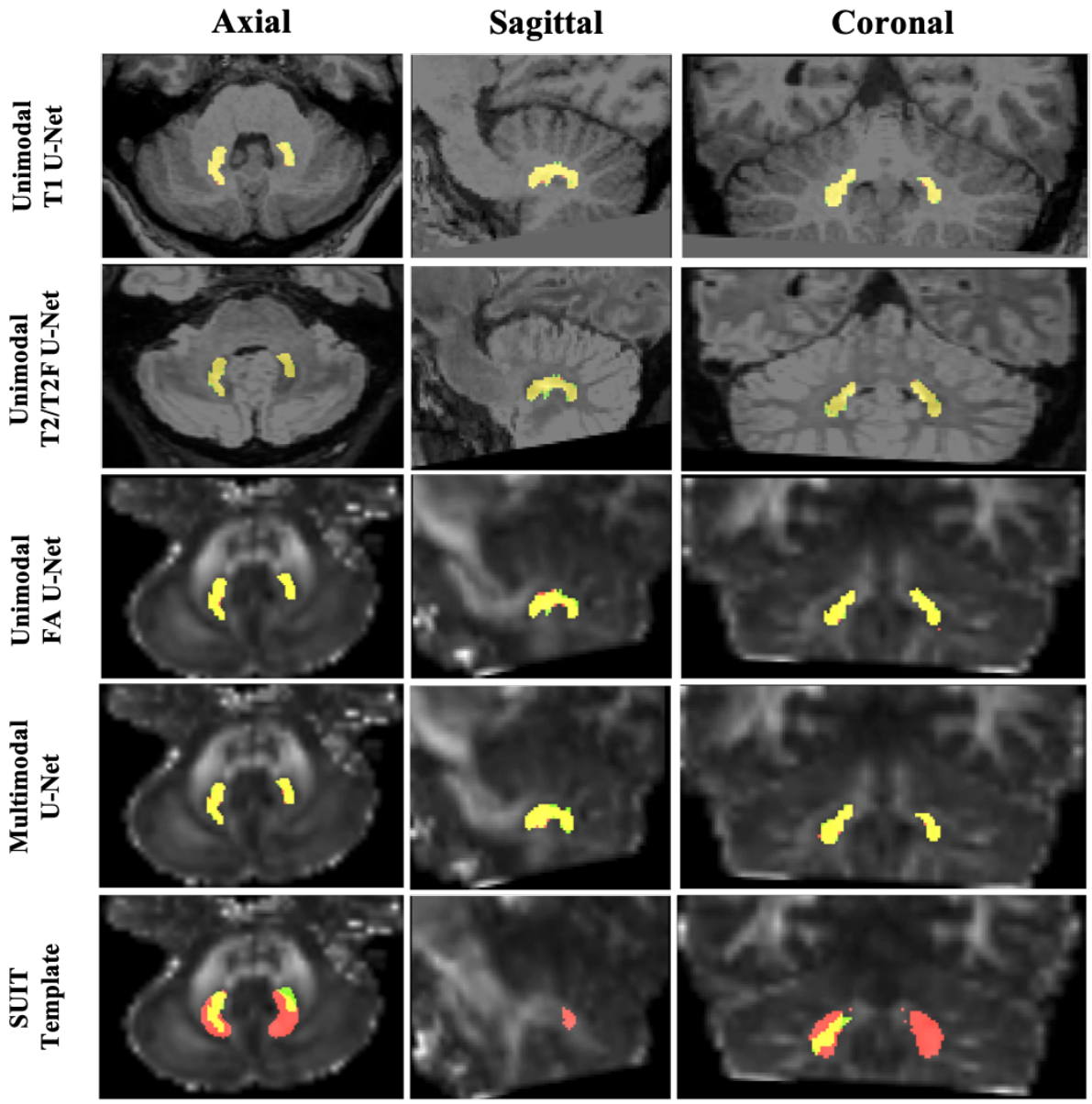

Figure V.3 Qualitative results of the DN segmentation. Here, we show the output of the automatic segmentation using unimodal U-Net, a multimodal U-Net, and single atlas registration of the SUIT template in the same subject. The manual label is shown in green while the predicted label is shown in red. Overlap between the two structures is shown in yellow. The dice coefficient is 0.95 for the unimodal T1 network, 0.95 for the unimodal T2 network, 0.97 for the unimodal FA network, 0.93 for the multimodal network, and 0.41 for the SUIT atlas.

In Figure V.4, we show an example of an outlier automatic segmentation with DSC outside of the range of the standard deviation. Here, the DSC was 0.48 for the unimodal T1 network, 0.27 for the unimodal T2 network, 0.54 for the FA network, 0.45 for the multimodal network, and 0.14 for the SUIT template.

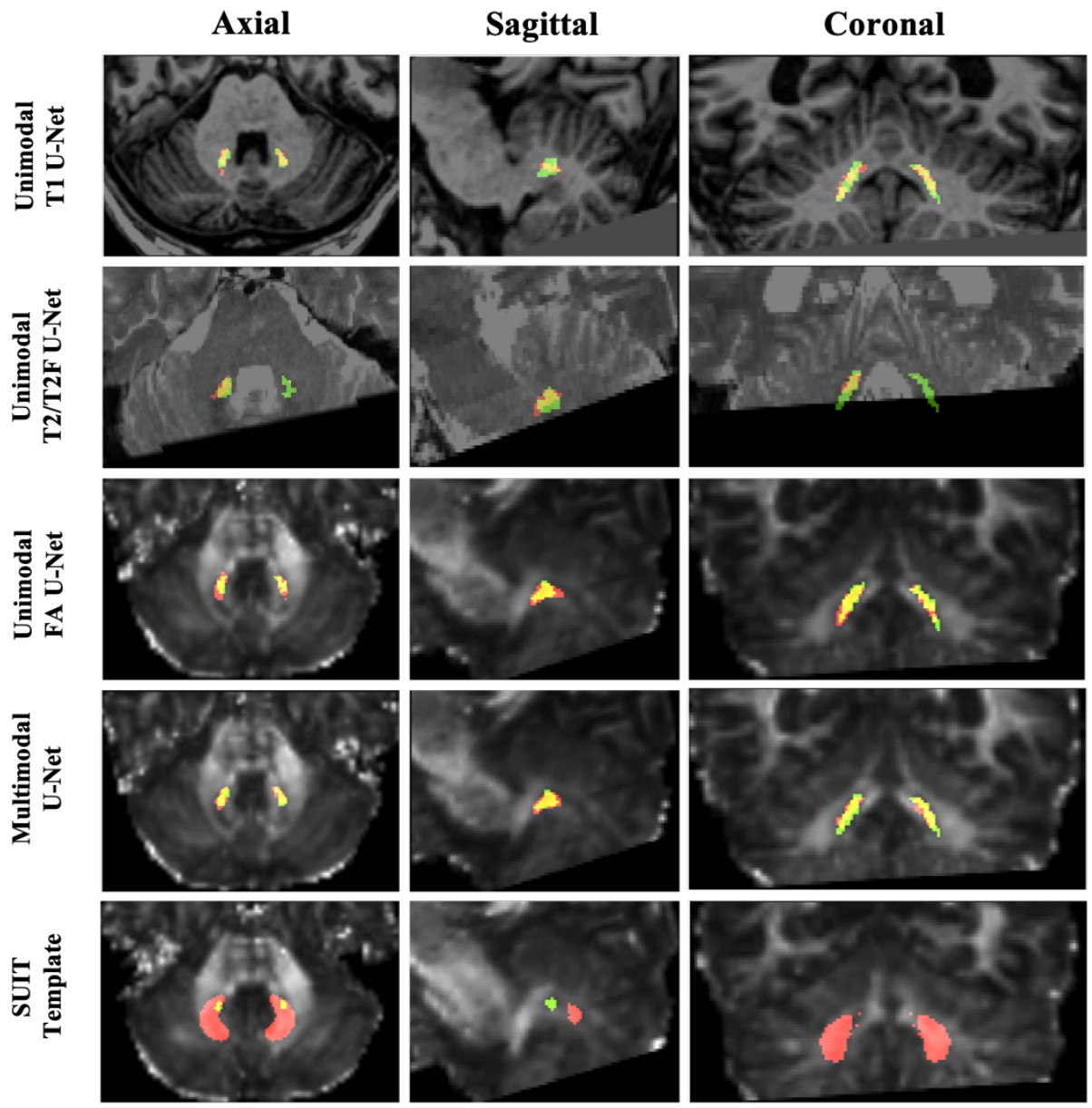

Figure V.4 Results of the DN segmentation for an outlier subject. Output segmentations from a representative subject with a dice coefficient over one standard deviation below average across all network modalities. The manual label is shown in green while the predicted label is shown in red. Overlap between the two structures is shown in yellow.

4. Discussion

Here, we applied a deep neural network for automatic DN segmentation using single imaging modalities as well as a combined multi-modal approach. Like many other deep gray matter structures, the DN is difficult to visualize on T1 images and therefore has required time and resource-intensive multimodal specialized imaging sequences in previous studies. We provide a method to translate DN labels traced in FA maps derived from diffusion imaging in order to achieve reliable automatic delineation on other modalities such as T1, T2, or FLAIR. We find that a deep learning-based DN segmentation can better reproduce FA labels than single-atlas registration using the SUIT template. The proposed automatic method shows a substantial increase in performance compared to available methods in the literature, which showed agreement in the range of 30-50% using a probabilistic atlas [2, 99, 100]. This work can be further extended to other deep gray matter structures to translate information across modalities. While, a multi-modal approach does not substantially increase performance, high agreement can be achieved with FA maps and even T1 and T2-weighted MRI. This is not entirely surprising, since the labels were drawn on the FA maps, and may contain all necessary information to successfully delineate the DN. Including additional modalities may introduce errors in registration between modalities that decrease accuracy. It is encouraging that each unimodal U-Net achieves superior performance compared to the literature. This will make such tools widely applicable across research and clinical domains.

We use FA maps to delineate the DN and train the proposed algorithm. FA measurements act as a surrogate for water diffusivity molecules, contrasting isotropic gray matter structures with adjacent white matter tracts. The DN, given its larger size relative to other nuclei and location surrounded by cerebellar white matter, is an apt candidate for detection using FA maps. Unfortunately, the resolution of typical diffusion studies is 2.5 x 2.5 x 2.5 mm3, resulting in a coarser automatic delineation, even when transferred to higher resolution scans like T1-weighted MRI. Notably, it is not possible to visualize the characteristic gyri of the DN at this resolution. It is exciting to see that we can transfer structural information from the manual labels in FA maps to learn the delineation of the DN in other modalities like T1- or T2-weighted. If a large number of high-quality manual segmentations can be obtained from sophisticated imaging

techniques that better visualize the DN, then it is possible to apply these to common clinical imaging sequences.

A benefit of using FA maps over T1- or T2-dependent sequences is the effect of age-dependent iron deposition in the DN. Previous studies have shown that iron deposition is not only associated with age, but is also increased in PD [225, 231]. A higher concentration of iron has been shown to shorten the longitudinal relaxation time in MRI and cause local field inhomogeneities. These inhomogeneities may artificially increase volumetric measurements of the DN compared to a diffusion-based measurement. Iron-related changes have been exploited for better visualization of the DN in proton density [232] or through quantitative susceptibility imaging [233]. This may explain the discrepancy between the SUIT atlas and the manual segmentation seen in Figure V.3 and V.4. The SUIT template was created by visualizing deep cerebellar structures in SWI at 7T (1), which is affected by iron content. We provide an accurate and reliable method of segmenting the DN that may be less correlated with iron content. It is still an open question of how iron content may change in the deep structures of the cerebellum with age and with disease, particularly movement disorders. Further work may explore how a diffusion-based measurement of the DN differs from a iron-dependent measurement such as the SUIT atlas. A longitudinal study may provide some valuable insights about structural changes in the cerebellum that occur with age.

The diffusion-based quantification of DN structure proposed here offers a reliable and accurate method of automatic delineation. Deep learning models have shown a dramatic improvement in accuracy over conventional methods in many medical imaging applications, but frequently require large training datasets in order to generalize well to new test cases [17]. Our model can generalize well within a subset of imaging data and reproduce manual labels with high fidelity to manual labels. Inevitably, the method is limited by the quality of the manual labels provided by the rater. Figure V.3 shows that we can reproduce the manual labels with high fidelity. This provides a rough regional estimate to delineate and detect the DN. Future applications of this work may include using the DN segmentation as a region of interest for connectivity studies in functional and structural imaging or as transfer learning for other deep cerebellar structures.

We probe the accuracy of our method by observing the output of a segmentation over one standard deviation below the mean DSC for each modality. In Figure V.4, the output segmentation for this subject does not fail entirely, but a difference of a few pixels can largely affect accuracy as measured by the dice coefficient. The low performance in this subject may be due to errors in registration, since we see the inferior part of the image is cut off. Deep convolutional networks can be affected by heterogeneity in the data [141-146], so irregular boundaries in the inferior portion of the image could affect the output. Regardless, the segmentations from the convolutional networks better reproduce the manual labels than a single SUIT atlas. We further probe the accuracy of our model by looking at the Grad CAMs for each

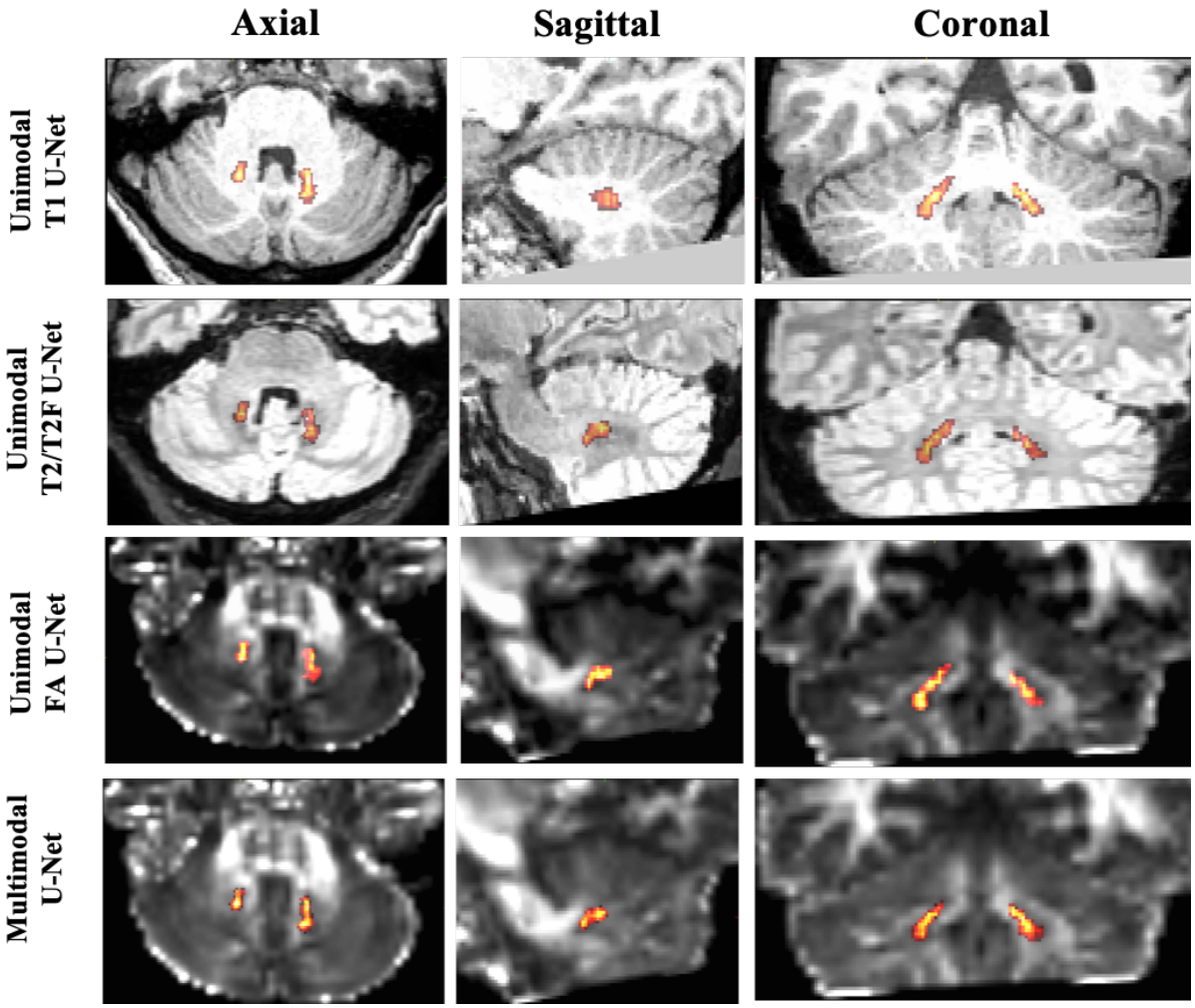

Figure V.5 Grad CAMs for each deep convolutional network with unimodal or multimodal inputs. We see that the attention is focused around the areas of the DN, which correspond to the output segmentations.

network (Figure V.5). While interpreting these maps remains an open problem in the field of image processing, these maps provide a valuable heuristic for locating where the network is "paying attention" or if there is additional off-target data informing the segmentation. Figure V.5 shows that the attention across all networks is highly concentrated around areas corresponding to the DN. These maps may even be used to inform a probabilistic segmentation of new test subjects to improve generalizability.

One limitation of this study is that labels were drawn at different image resolution for both the healthy controls and the disease cohorts. A direct comparison of DN volume between the healthy controls and either of the disease populations will be impossible due to partial volume effects, which make the DN tracing in the disease population artifactually larger than the healthy controls. We observe this in Table V.1, where the DN volumes are significantly larger in the disease cohorts, even after normalization by posterior fossa volume. We believe this to be an artifact of the image resolution, since evidence in the literature supports no difference in the size of the DN between healthy controls and PD patients [234].

The methods proposed here demonstrate that automatic segmentation of MRI can achieve a reliable and accurate delineation of structures difficult to visualize on common imaging modalities. Moreover, we test the applicability of this volumetric method to discriminate between cohorts with movement disorders and controls. This work will assist volumetric and functional studies as the role of the cerebellum in motor function, sensation, and cognition continues to be elucidated.

VI. Generalizing Deep Whole Brain Segmentation for Pediatric and Post-Contrast MRI with Transfer Learning

1. Introduction

Whole brain segmentation is an important task in medical image analysis in order to provide a quantitative, noninvasive measurement of neuroanatomy from magnetic resonance imaging (MRI). Although a manual delineation of brain structures is considered the gold standard, this process can be largely time- and resource-intensive [235]. For many years, atlas-based segmentation was considered the standard for automatic whole brain segmentation due to high accuracy across over 100 labels [87, 88, 235]. However, a key limitation of atlas-based segmentation is the extensive computational time required, which limited the ability to analyze large-scale cohorts [228, 235]. Recently, Huo et al. proposed a complete 3D deep learning pipeline to produce fast, full-resolution, whole brain segmentation with over 100 labels. This pipeline, called spatially localized atlas network tiles (SLANT), uses a tile-based method to divide a T1-weighted brain MRI into overlapping tiles and train independent networks, which are then fused via majority voting [147, 236]. This method exhibited better performance than multi-atlas and other deep learning-based methods on brain MRI acquired for research[237]. Since then, SLANT has been used as a benchmark for whole-brain segmentation [238, 239] and the tile-based method has been adapted by other groups to solve segmentation problems with a large number of labels [240-242].

When evaluating the SLANT segmentation on same-subject clinical data, investigators found a high coefficient of variation between volume estimates of several ROIs across clinical T1 MRI acquisition modalities within the same subject [243]. This observation highlights a key obstacle in deep learning-based image processing tools: most algorithms are trained and tested on high-quality research scans, due to high resolution and SNR and reported accuracy that reflects performance on research data. However, rarely is there a description of performance on clinical data, which can be heterogenous due to resolution, noise, artifact, or the presence of intravenous contrast. For instance, Figure VI.1 shows T1w brain MRI from the

same subject acquired under two different acquisition parameters, as well as with and without intravenous contrast. These images highlight the heterogeneity in clinical data due to differences in acquisition, tissue contrast, or noise. It is unclear how deep learning algorithms trained on research data will generalize to heterogeneous clinical datasets.

The challenges in harmonizing multi-site and multi-sequence imaging datasets have been described previously, recognizing that the variability in tissue contrast can come from scanner factors (number and

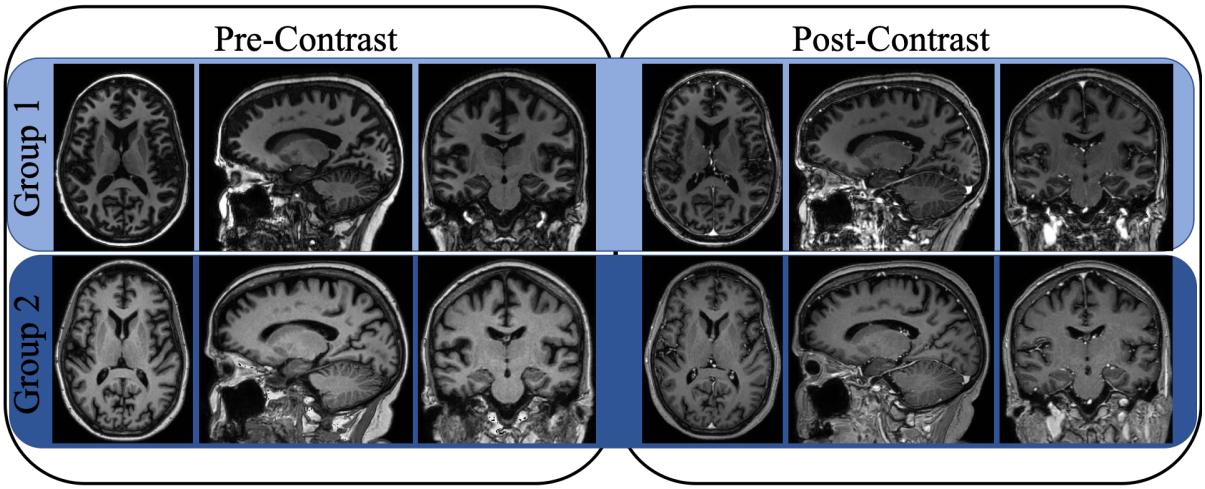

Figure VI.1 Representative clinical magnetic resonance imaging of the same subject acquired under different acquisition parameters (Groups 1 and 2), with and without intravenous contrast.

sensitivity of head coils, imaging gradients, magnetic field homogeneity, or scanner noise) or software factors (image reconstruction methods or software updates) [101]. Many of the efforts directed at image harmonization have sought to find a mapping of tissue contrast intensities to account for small changes in acquisition parameters [101-104], while recognizing that these methods may not be robust to large changes [101, 103], such as those found in clinical data like the presence of intravenous contrast or large variations in flip angle. These large variations result in a different representation of the underlying anatomy and harmonizing across them requires a robust technique that does not only map tissue intensities. Similarly, it has been shown that image processing algorithms have decreased performance in young children due to differences in gray matter to white matter contrast during development and limited training examples [244].

Consequently, most efforts dedicated at pediatric brain segmentation have focused on the development and application of age-specific atlases, which can be time- and resource-intensive [245, 246].

There is extensive work done on supervised, semi-supervised, and unsupervised domain adaptation in deep learning, as reviewed by Cheplygina et al. [247]. For example, van Opbroek et al used weights on the distribution of training images to improve voxel-wise lesion classification from different research MRI studies [248]. Kouw et al. used a Siamese network to learn the invariant representation of different research MR acquisitions to improve on brain tissue classification [249]. Other approaches, have aligned image manifolds to exploit correspondences between domains [250]. Similarly, spatial alignment between corresponding images would facilitate domain adaptation between corresponding spatial features. Improving algorithm generalizability on clinical data will allow for larger learning datasets and translation to the bedside.

One common approach towards generalizability across sites or scanners is to use transfer learning (TL) to update the neural network weights for local factors [251, 252]. For instance, if SLANT were applied at a new site, its performance could be refined by taking a number of training examples and updating the weights of the original network. However, this technique is commonly recognized to risk degradation of performance on the original validation/test cohorts, thus limiting its generalizability [251]. Nevertheless, it is often desirable to improve performance on datasets that deviate from the original training data. For instance, the presence of intravenous contrast may affect the performance of intensity-based methods such as deep neural networks [243]. Although most algorithms are trained on adult images exclusively with or without intravenous contrast, it would be ideal to generate reproducible results in image processing pipelines despite the presence of contrast, since contrast-enhanced data are common, but underrepresented in deep learning algorithms. Recent work in medical image processing has focused on analysis of contrast-enhanced MRI, particularly when using deep learning methods [253-257]. It is unclear whether commonly used algorithms, such as SLANT, can generalize to imaging data acquired in the clinical scenarios. TL offers an alternative to improve the generalizability of algorithms across imaging contrasts, but the degradation on the training set performance is unknown. This approach is superior to intensity

harmonization [102, 258, 259] or image hallucination [260-262] because it does not alter to the original image, but rather can adapt to differences in acquisitions. Achieving reproducible results on heterogeneous clinical imaging data would largely increase the data available for medical image analysis in patient populations and translate the tools from engineering to the patient bedside.

In this work, we explore how TL applied to unlabeled clinical data can address these concerns in the context of adapting SLANT to scanning protocol variations. We hypothesize that the existing algorithm SLANT can be refined using TL to improve performance across acquisition parameters and the presence of intravenous contrast with a minimal loss in performance on the original SLANT training dataset. We optimize for acquisition with 480 pairs of clinically acquired T1w MRI with and without intravenous contrast and the accuracy of the post-contrast segmentations assessed relative to the pre-contrast automated assessment. We compare our results against the original SLANT algorithm without fine-tuning as well as FreeSurfer v6.0.1. We perform five-fold cross-validation on paired clinical data and then evaluate on a completely withheld paired clinical dataset (N = 29 scan pairs), which has a different acquisition domain, for withheld validation on all algorithms. We show that we can leverage unlabeled clinical data to refine a pre-existing segmentation algorithm against unknown image heterogeneities using TL, while preserving segmentation performance on the original dataset. This work emphasizes the feasibility of sharing pre-existing algorithms and refining models for data of a different domain instead of generating new manual labels and retraining from scratch. The work extends our previous conference work [128] with the following new efforts: (1) the extension of the clinical dataset from 36 to 255 subjects with paired imaging, (2) a detailed characterization of acquisition parameters and subsequent formation of subgroups, (3) the inclusion of a withheld test set with different acquisition parameters in addition to the cross-validation efforts, (4) A comparison to another commonly used whole brain segmentation pipeline (i.e. FreeSurfer).

To further demonstrate the challenge of reproducible segmentations, we consider the automatic segmentation of the hippocampus. The hippocampus is a small gray matter structure located in the medial temporal lobe. There has been extensive work dedicated to the segmentation of the hippocampus [263-268] since it is an important structure involved in learning and memory [269] as well as spatial navigation [270].

The hippocampus has also been demonstrated as an important imaging biomarker for Alzheimer's disease [271], psychosis [272], and epilepsy [273], thus making it a desirable anatomical target for reproducible segmentation and volumetric segmentation across scans within the same subject [89, 263, 274, 275]. However, the location of the hippocampus makes it challenging to segment due to proximity of heterogenous tissue, such as cortical folds, white matter, and vasculature, each of which has different image texture and contrast [268]. Changes in any of the image intensity patterns in these structures due to aging, disease states, acquisition parameters, or contrast may affect segmentation reproducibility between subjects, in the case of cross-sectional studies, or within subjects, in the case of longitudinal studies [268, 275]. We tested the application of our pipeline on volumetric estimates of the hippocampus, where we show a decrease the RMSE between paired imaging of the same subject by 67%. Overall, this work proposes a solution to preserve generalizability of whole brain segmentation through a TL pipeline with unlabeled clinical data to translate algorithms optimized for research data towards heterogeneous clinical acquisitions. The work presented in this chapter was initially published in the conference proceedings of *SPIE Medical Imaging 2020.* Further expansion of this work focused on clinical, contrast MRI was done and submitted for review at the *Journal of Medical Image,* both works with Camilo Bermudez as first author.

2. Methods

2.1. Subjects and Imaging Data

The original SLANT whole brain segmentation algorithm consists of two stages: (1) Pretraining by leveraging a large dataset (N = 5,111 subjects) with automatically-generated labels from multi-atlas segmentation [147], (2) Transfer learning for refinement using a small dataset of manually labelled ground truth atlases from the Open Access Series on Imaging Studies (OASIS) dataset [147, 236]. In this work, we began with the pretrained model (https://github.com/MASILab/SLANTbrainSeg) and use a paired clinical dataset with pre- and post-contrast brain MRI without manual labels to refine the segmentation.

The original OASIS dataset used for TL in SLANT consists of 45 subjects with T1-weighted brain MRI with ages 18-96 years old. These scans were acquired at 3T field strength at a field of view which varies from 256x270x256 to 256x334x256 voxels all with 1mm isotropic resolution. Each scan has 133 manual labels traced according to the BrainCOLOR protocol used as ground truth [136]. Accuracy was measured using the Dice Similarity Coefficient (DSC) between the predicted whole brain segmentation and manual labels [37].

In this work, we use two datasets to perform TL on two different applications. The first dataset consists of 30 pediatric subjects (ages 2.34 – 4.31) with T1-weighted brain MRI. These datasets were acquired with the approval of the Institutional Review Board at the University of Calgary. Imaging was acquired at 3T field strength with field of view 256x210x256 at 0.9 mm isotropic resolution. Initial automatic whole brain segmentation was obtained into the same 133 regions using multi-atlas segmentation

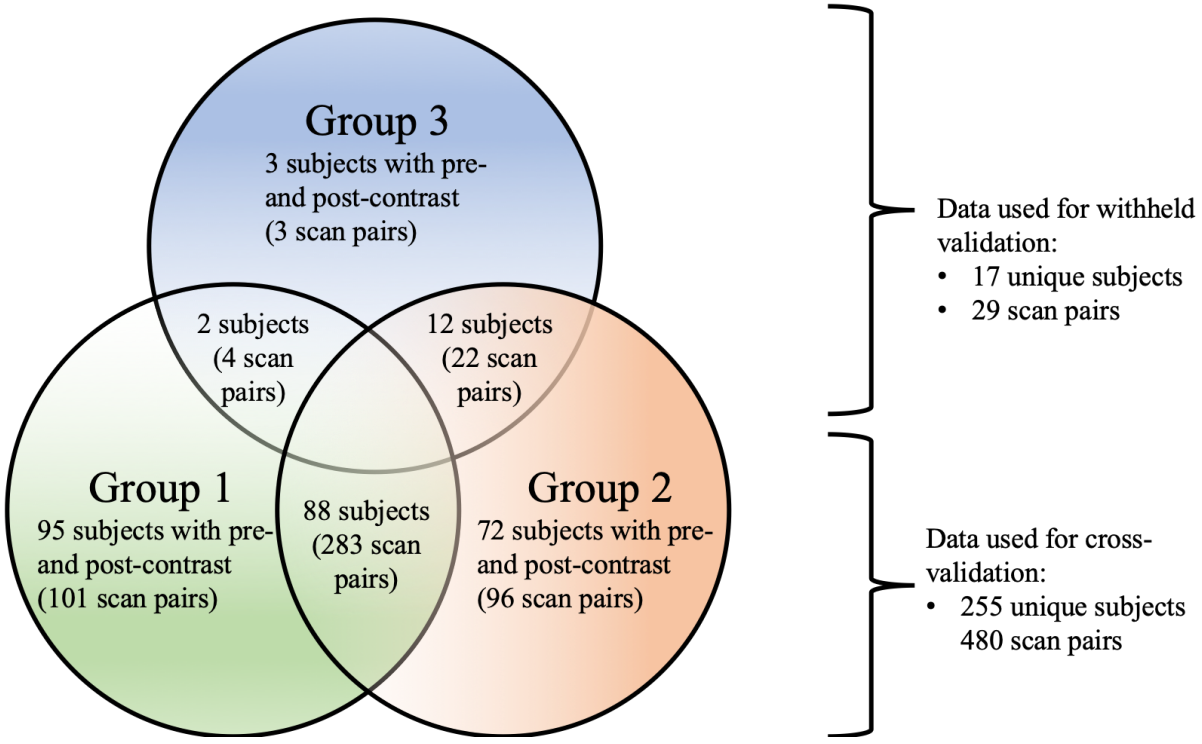

Figure VI.2 Clinical dataset used for transfer learning. We split the clinical data given the acquisition parameters TR, TE, and flip angle into three groups. The figure shows the total number of subjects and paired scans within and between groups. We used Groups 1 and 2 for training and cross-validation and Group 3 for withheld validation.

from 35 adult atlases as outlined by [89, 135, 276]. These labels were then manually corrected by a rater expert in pediatric imaging.

In the second dataset, we use unlabeled, paired clinical imaging to improve reproducibility across heterogeneous clinical datasets. All datasets were acquired in a deidentified manner under the approval of the Institutional Review Board (IRB) at Vanderbilt University Medical Center. We used a clinical dataset consisting of 272 subjects with clinically acquired, paired T1-weighted brain MRI with and without intravenous contrast for a total of 509 scanning sessions with paired imaging. These subjects were acquired with 3T field strength and a field of view of 256x170x256 at 1mm isotropic resolution. Paired images were registered to MNI-305 space using affine registration [197]. Image intensity normalization was performed using the method described by Huo et al. [147]. Since the paired clinical imaging is unlabeled, we use the original SLANT to generate labels on the pre-contrast image. Since these images are co-registered, we use these same labels to train on the post-contrast images.

Clinical imaging sequences may be acquired with a broad range of acquisition parameters. In order to assess the ability to generalize within and between groups, we found three clusters of scans when organized according to echo time (TE), repetition time (TR), and flip angle. We found the following three groups:

1. 538 total scans with a flip angle of 5°, TE < 4 ms (range 3.65 – 3.94 ms; median = 3.65), and TR in the range of 7.92 – 8.55 ms (median = 7.92 ms).
2. 603 total scans with a flip angle of 8° and TE < 4 ms (range 3.16 – 4.00 ms; median = 3.16) and a TR in the range of 6.98 – 8.76 ms (median = 6.99 ms).
3. 195 all other scans with a flip angle > 8° and a wide range of TE (range 8.06 – 25.0 ms; median = 9.80) and TR in the range 2.21 – 5.41 ms (median = 4.60).

We decided to use Group 3 as the withheld test set and use Groups 1 and 2 for training and testing. We further refine the groups by finding those subjects who have both pre-and post-contrast imaging within the

same group (intra-group harmonization) and those subjects who have imaging across groups (inter-group harmonization). Figure VI.2 shows the total numbers within each group having pre- and post-contrast imaging, as well as the count of subjects that span across groups. We pooled Group 1 and Group 2 for training, validation, and testing of the segmentation algorithm to harmonize across contrast and acquisition modalities. We then evaluated the resulting algorithm on the withheld test set (Group 3) subjects.

2.2. Transfer Learning Pipeline

Throughout this work, we begin all of our experiments with a SLANT model of whole brain segmentation that has been pretrained in 5,111 subjects using automatically generated manual labels, as described by Huo et al [147]. Therein, Huo performed TL using the dataset of 45 adults with manual labels for best results [147]. Here, we use this result from Huo et al. as the baseline comparison. Instead of refining on the 45 adults with manual labels, we refine on either a) the manually corrected pediatric labels or b) the training set of paired imaging using the labels generated on the pre-contrast image to train on the co-registered post-contrast image. We did not freeze any weights or alter the network in the learning process, since our goal was to use SLANT without alterations to improve local performance. Instead, we drive reproducibility by leveraging the pediatric or paired, unlabeled clinical data. We seek to train a network that achieves high performance in a new, unseen dataset while preserving performance on the original adult cohort.

Here, we aim to refine the whole brain segmentation in the context of different acquisition parameters and the presence of intravenous contrast. In the case of pediatric imaging we trained on the manually corrected labels obtained from multi-atlas segmentation to generate a model called pediatric SLANT (pSLANT). In the case of contrast imaging, we generated labels on the pre-contrast MRI in the case of paired contrast imaging, or on the image in Group 1 in the case of comparisons between Groups. We use TL to train a model using the contrast-enhanced images exclusively called contrast SLANT (cSLANT). We sought to maximize reproducibility DSC (rDSC), or the accuracy between the pre-contrast

and post-contrast image, which represents similarity of automatic segmentation between two images of the same subject. All models and TL code is available at (https://github.com/MASILab).

2.3. Performance Analysis

All models were trained with 70% of each of the subjects as the training set, with 10% of the subjects dedicated to model validation and hyperparameter tuning and the remaining 20% for testing. Model accuracy was evaluated on the validation set using rDSC, with the final model chosen as having the highest validation set rDSC after 30 epochs of TL as in [147, 236]. A five-fold cross-validation scheme was employed in five independently trained models such that each subject was included in the testing set only once. All the results reflect the performance on the testing set. All performance metrics were compared using Wilcoxon signed-rank test with Bonferroni corrections for multiple comparisons. All experiments were performed on an NVIDIA GeForce Titan GPU. The original SLANT algorithm as well as TL modifications were implemented in PyTorch v0.4 [277]. TL training was performed with the Adam optimizer using the Dice coefficient loss function at a learning rate of 0.0001 [37, 278].

We further evaluate the trained models on a withheld dataset consisting of subjects with imaging in Group 3. This cohort includes paired imaging pre- and post-intravenous contrast as well as subjects with paired imaging in Groups 1 and 2. Since we trained five different models in the initial TL phase, we evaluate all models and aggregate the results using majority vote. Last, in order to assess the performance against the original manual labels on the OASIS dataset, we evaluate the accuracy of segmentation first with the original SLANT and then with the fine-tuned cSLANT.

2.4. Hippocampal Volume Estimation

We tested an application of our method by measuring the volume of the hippocampus across paired images between intravenous contrast and between acquisition parameters. We used the full segmentation pipeline in Freesurfer 6.0 [279] to generate a whole-brain segmentation and extract the volume of the whole left and right hippocampus (Labels 17 and 53 in the FreeSurferColor LUT) ran on XNAT distributed

computing at Vanderbilt University [199]. In order to account for the differences in atlas delineations between FreeSurfer and the atlases used in this work, we only look at the root mean squared error (RMSE) between imaging pairs. Again, we evaluate the change in hippocampal volume within each Group for pre- and post-contrast imaging and between Groups of acquisition parameters. We evaluate in the five-fold testing set (Groups 1 and 2) as well as the withheld dataset (Group 3). We perform three analyses: volume changes using the original SLANT, the TL cSLANT, and FreeSurfer.

3. Results

3.1. Pediatric Imaging Reproducibility

The first set of experiments seeks to improve performance on the whole brain segmentation of a pediatric population. As a baseline, we perform a prospective evaluation of the original SLANT algorithm on the pediatric dataset, which achieves a mean rDSC of 0.82. After TL, we achieved a significant improvement on rDSC of 0.90 on the pediatric labels using pSLANT. Performance on the original OASIS dataset dropped from 0.72 to 0.70 ($p < 0.001$) (Table VI.1). An important goal of this task is to enforce the modifications introduced during manual correction of the automatic labels from multi-atlas segmentation in order to introduce expert context. We examined the bilateral occipital poles as a region of interest (ROI) that underwent extensive manual editing as a representative ROI for this task. Figure VI.3 shows the output segmentation overlaid on a representative subject with median performance.

Method	Pediatric Labels	OASIS Labels
Original SLANT	0.82 +/- 0.01	0.72 +/- 0.05
Pediatric SLANT	0.90 +/- 0.01	0.70 +/- 0.07

Table VI.1 Qualitative results of SLANT generalization on pediatric labels measured using reproducibility DSC. There is a significant increase in performance on the pediatric labels using pSLANT ($p<0.001$) with a smaller but yet significant decrease in performance on the original OASIS labels ($p<0.001$).

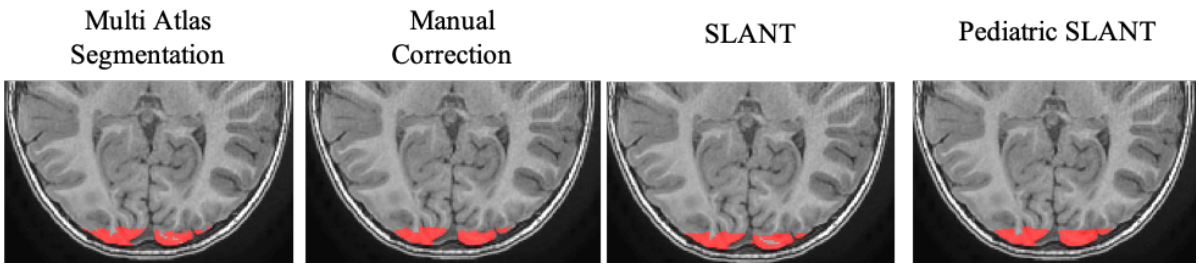

Figure VI.3 Qualitative results on pediatric SLANT on the occipital lobe labels. The original multi atlas labels are shown on the left, followed by the manual corrections by the neuroimaging expert. Next are the results from the original SLANT and the pediatric SLANT. Pediatric SLANT is able to reproduce many of the corrections made during manual editing.

3.2. Contrast Imaging Reproducibility

First, we test whether using TL on paired, unlabeled clinical data can improve same-subject reproducibility of whole-brain segmentation. Table VI.2 shows the results for inter- and intra-group comparisons for subjects in Groups 1 and 2. We observe that at baseline, the original SLANT (OS) has an average reproducibility of 0.72 DSC. Once we perform TL, we can compare the OS labels against the labels produced on the re-scan image of the subject that has either intravenous contrast or different acquisition parameters for a mean rDSC of 0.80 ($p < 0.001$). We note a larger increase in rDSC to 0.82 ($p< 0.001$) between the original labels and the labels after TL between the scan-rescan images. This effect was also observed when analyzing ROIs separately (Appendix Table IX.1). Furthermore, if we compare the TL labels on the pre- and post- images of the same subjects, we see even higher agreement across all groups. The FreeSurfer comparison shows lower reproducibility overall (DSC = 0.53, $p < 0.001$) as well as a larger variance across subjects. Representative segmentations on the pre- and post-contrast image are shown for all methods in Figure VI.4, highlighting areas of disagreement that are fixed via TL. In addition, Figure VI.5 shows the disagreement overlaid on imaging from subjects with median performance across experiments.

	OS:OS	OS:TL	TL:TL	FreeSurfer
G1c	0.71 +/- 0.06	0.78 +/- 0.05	**0.81 +/- 0.05**	0.57 +/- 0.23
G2c	0.76 +/- 0.04	0.82 +/- 0.04	**0.84 +/- 0.04**	0.52 +/- 0.23
G12	0.72 +/- 0.07	0.79 +/- 0.05	**0.82 +/- 0.05**	0.51 +/- 0.24

Table VI.2 Transfer learning cross-validation performance of reproducibility DSC (rDSC) between and within Group 1 and Group 2. Bold represents the method with highest rDSC for each group. OS: Original SLANT, TL: Transfer Learning.

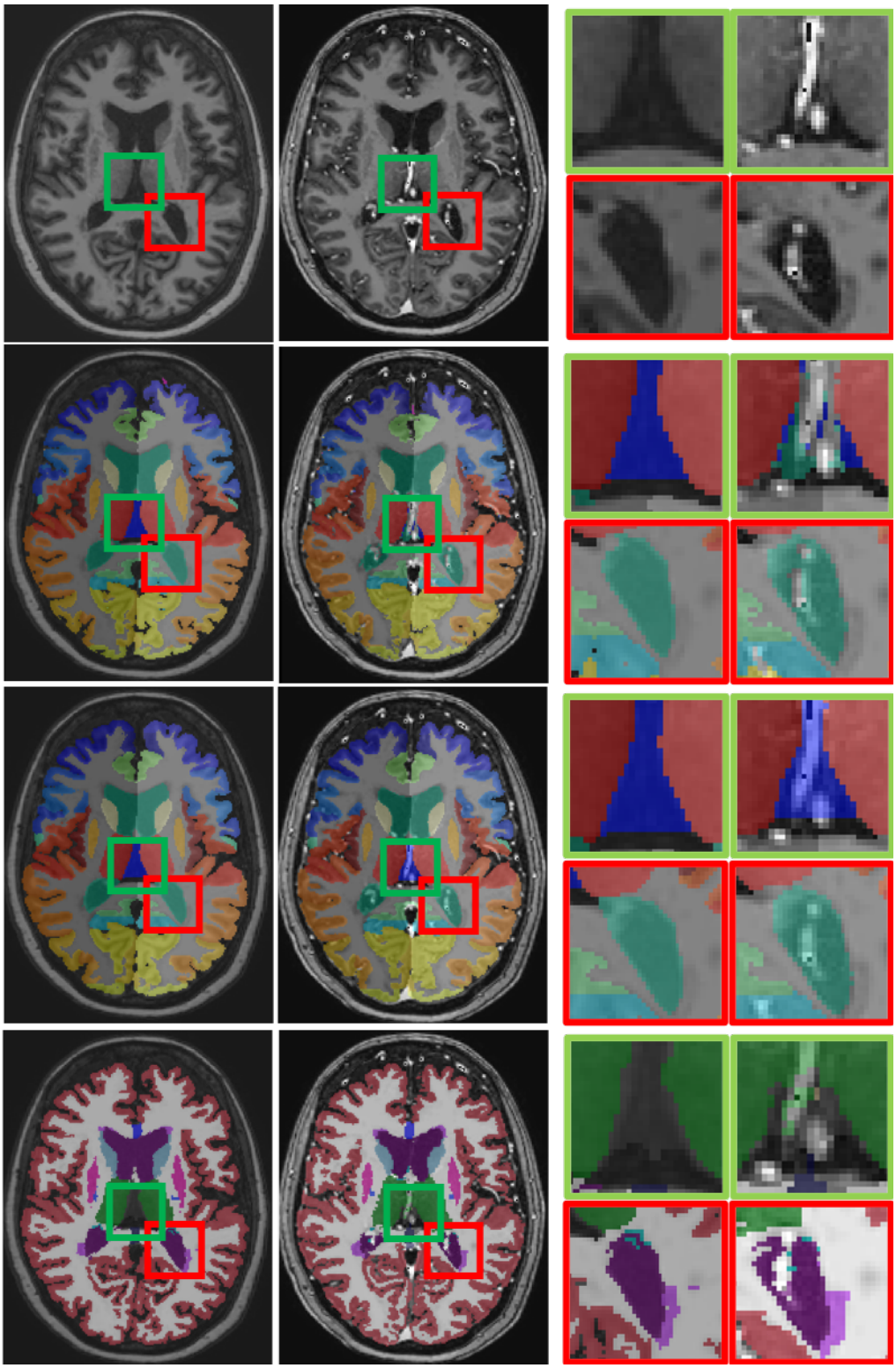

Figure VI.4 Segmentation results between pre- (left) and post-contrast (right) images. Top row shows the original image. Second row indicates the original SLANT followed by SLANT with transfer learning (TL). Fourth row shows FreeSurfer segmentation results. Green insets show segmentation at the third ventricle. Red inset shows segmentation at the lateral ventricle.

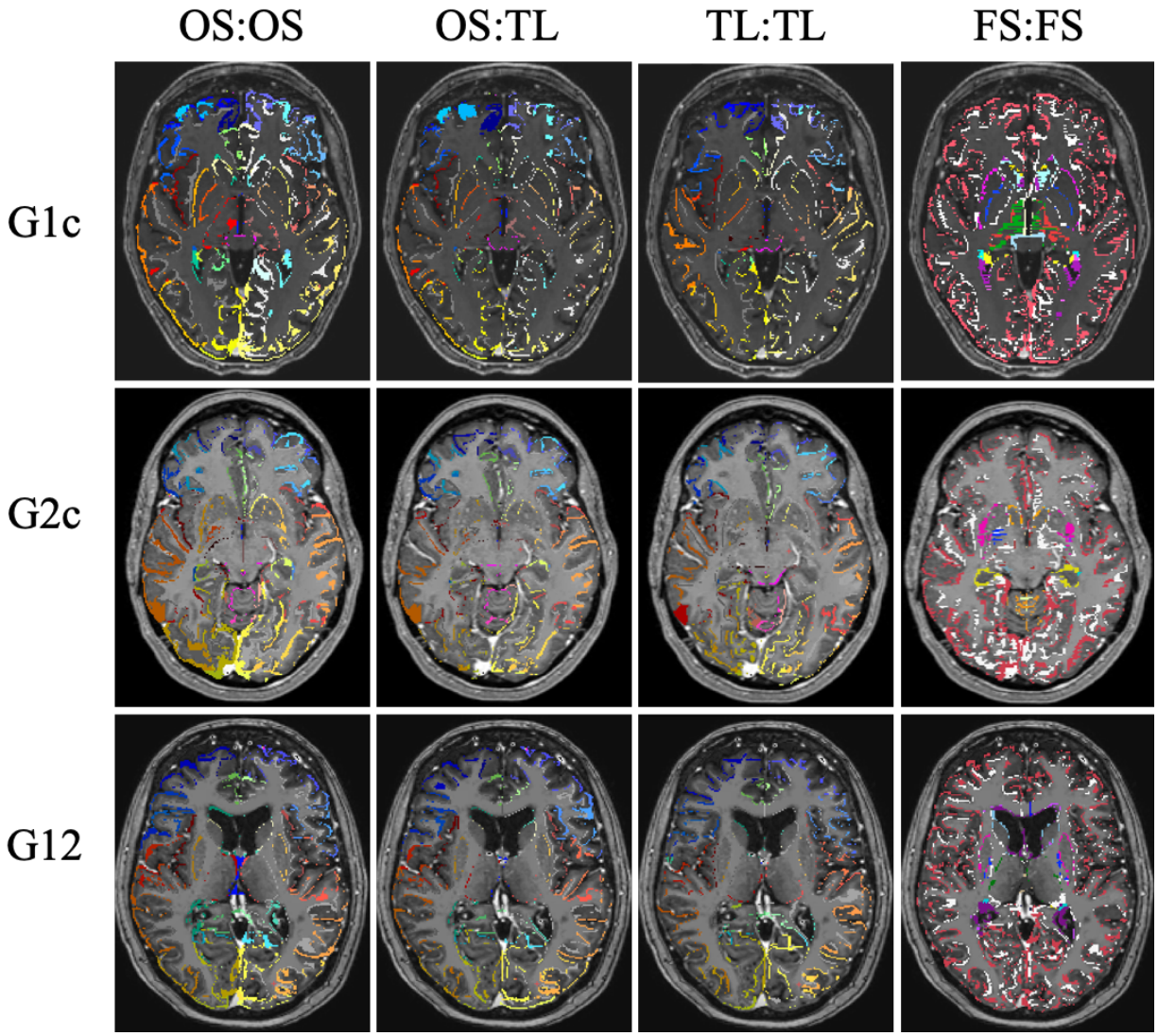

Figure VI.5 Areas of disagreement between labels during cross-validation of transfer learning on Groups 1 and Group 2. The disagreement labels are color-coded according to the original label. OS: Original SLANT, TL: Transfer Learning, FS: FreeSurfer.

3.3. Rescue of Original Manual Labels on cSLANT

After we perform TL on the scan-rescan images, we wish to evaluate whether the performance on the original SLANT dataset has changed. We note that the average DSC before TL against manual labels in this dataset was 0.70 +/- 0.02 (Table VI.3). There was a slight decrease in the average DSC after (TL 0.69 +/- 0.02, $p < 0.001$ Wilcoxon signed-rank test). Despite this being a statistically significant difference

across subjects, the slight drop in performance against manual labels is smaller than the gain in reproducibility in clinical imaging.

	Manual : OS	Manual : TL
OASIS	0.70 +/- 0.02	0.69 +/- 0.02

Table VI.3 Results on the original manual labels from the OASIS dataset before transfer learning and after transfer learning. There is a slight significant decrease between OS and TL ($p<0.001$, Wilcoxon signed-rank test).

3.4. Withheld sample of Group 3

We used the subjects from Group 3 as a withheld validation set and generated three cohorts: paired imaging with Groups 1 and 2, as well as pre- and post-contrast imaging of the same subject within Group 3. In total, we have 29 scans from 17 subjects. Table VI.4 shows the results on the withheld dataset are consistent with the cross-validation data, showing an increase in the learning task versus the original SLANT (OS:TL rDSC = 0.74 vs OS:OS rDSC = 0.66, $p < 0.001$ Wilcoxon signed-rank test) as well as a further increase when generating labels after TL on both scans from the same subject (TL:TL rDSC = 0.76, $p < 0.001$ Wilcoxon signed-rank test). Note that the results obtained from FreeSurfer are much lower than those from cSLANT and with a wider standard deviation (FreeSurfer rDSC = 0.34, $p < 0.001$ Wilcoxon signed rank test) . Figure VI.6 shows the areas of disagreement between predictions overlaid on subjects with median performance across experiments.

	OS:OS	OS:TL	TL:TL	FreeSurfer
G3c (N = 3)	0.76 +/- 0.05	0.84 +/- 0.02	**0.87 +/- 0.04**	0.31 +/- 0.33
G13 (N = 4)	0.72 +/- 0.08	0.78 +/- 0.09	**0.80 +/- 0.09**	0.52 +/- 0.14
G23 (N = 22)	0.64 +/- 0.13	0.72 +/- 0.11	**0.73 +/- 0.13**	0.30 +/- 0.28

Table VI.4 Transfer learning results on the withheld validation set. Bold represents the method with highest rDSC for each group. OS: Original SLANT, TL: Transfer Learning.

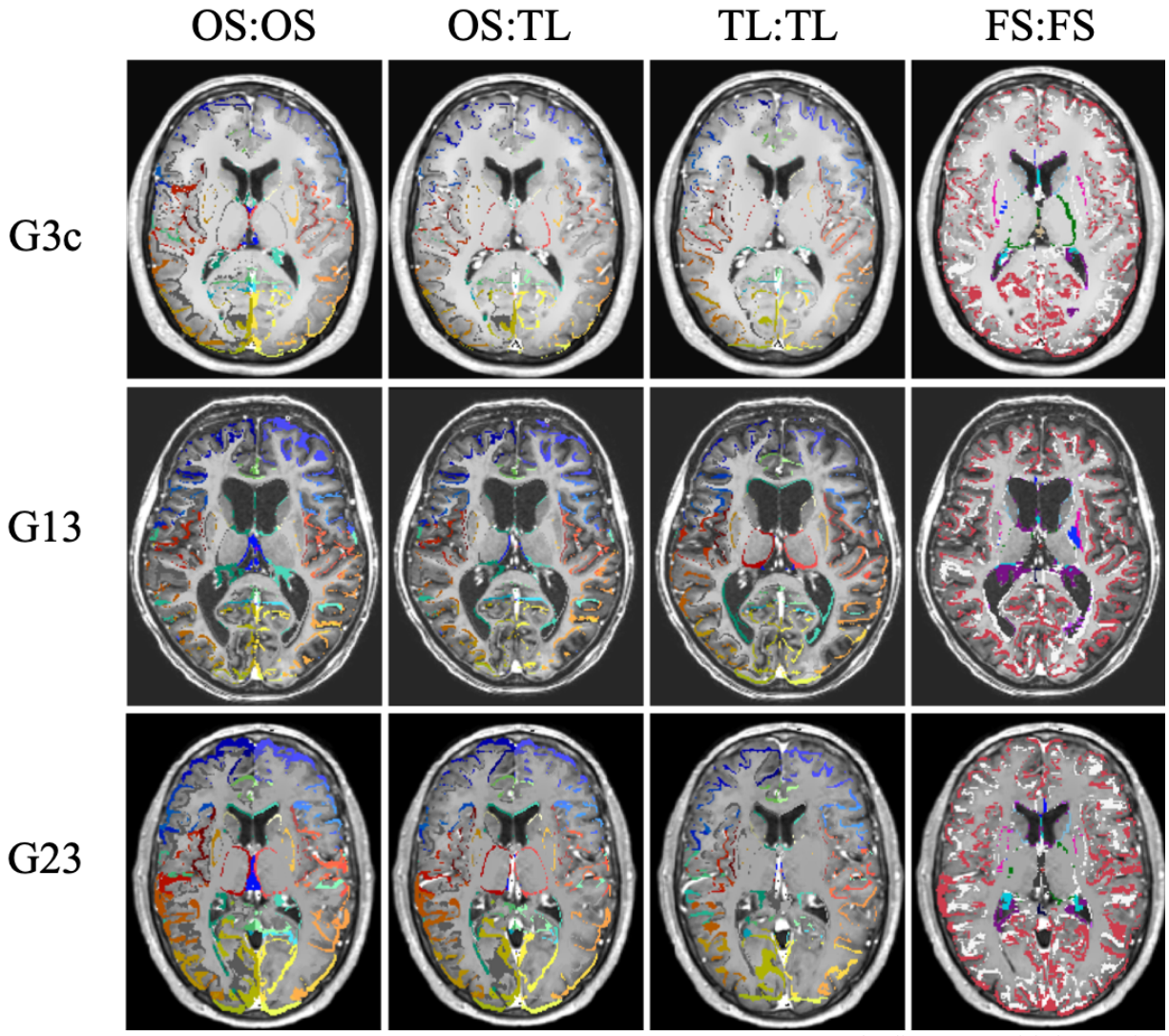

Figure VI.6 Areas of disagreement between labels of the withheld dataset in Group 3. The disagreement labels are color-coded according to the original label. OS: Original SLANT, TL: Transfer Learning, FS: FreeSurfer.

3.5. Volumetric Analysis of the Hippocampus

As an application of our method, we test the difference in volume of the hippocampus as estimated with three different methods: original SLANT, after transfer learning (cSLANT), and with FreeSurfer. First, we test it on the cross-validation dataset, which showed a decrease in volume RMSE between paired scans of 64% (Table VI.5). Similarly, we observe a large decrease in RMSE in the withheld dataset, except for subjects between Groups 1 and 3, which already had a low RMSE (Table VI.6). The RMSE obtained with

FreeSurfer was higher in both datasets than the original SLANT. Figures VI.7 and VI.9 show the change in hippocampal volume between pre- and post- imaging for all experiments. The difference in hippocampal segmentations overlaid on imaging is shown for the cross-validation dataset (Figure VI.8) and the withheld dataset (Figure VI.10).

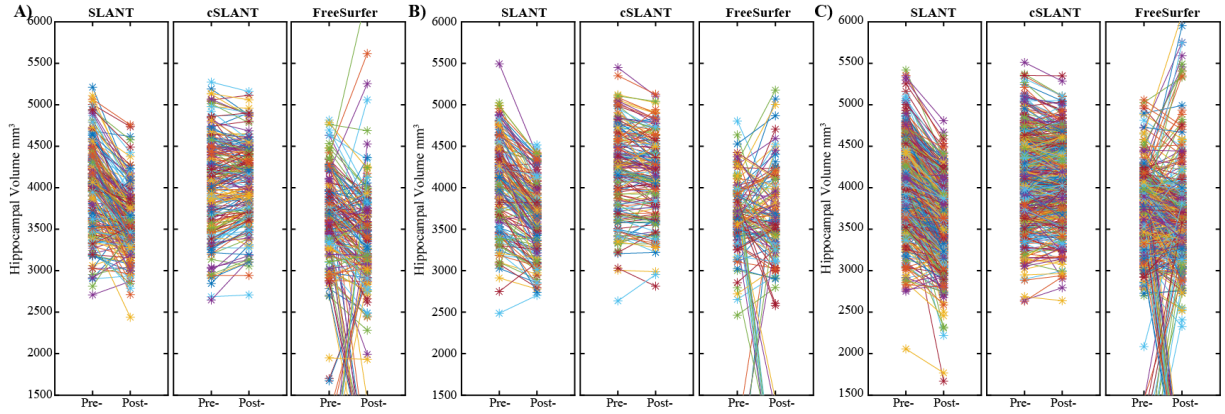

Figure VI.7 RMSE of hippocampus volume on the cross-validation data. Subplots represent performance on paired subgroups for A) Group 1c, B) Group 2c, and C) Group 12.

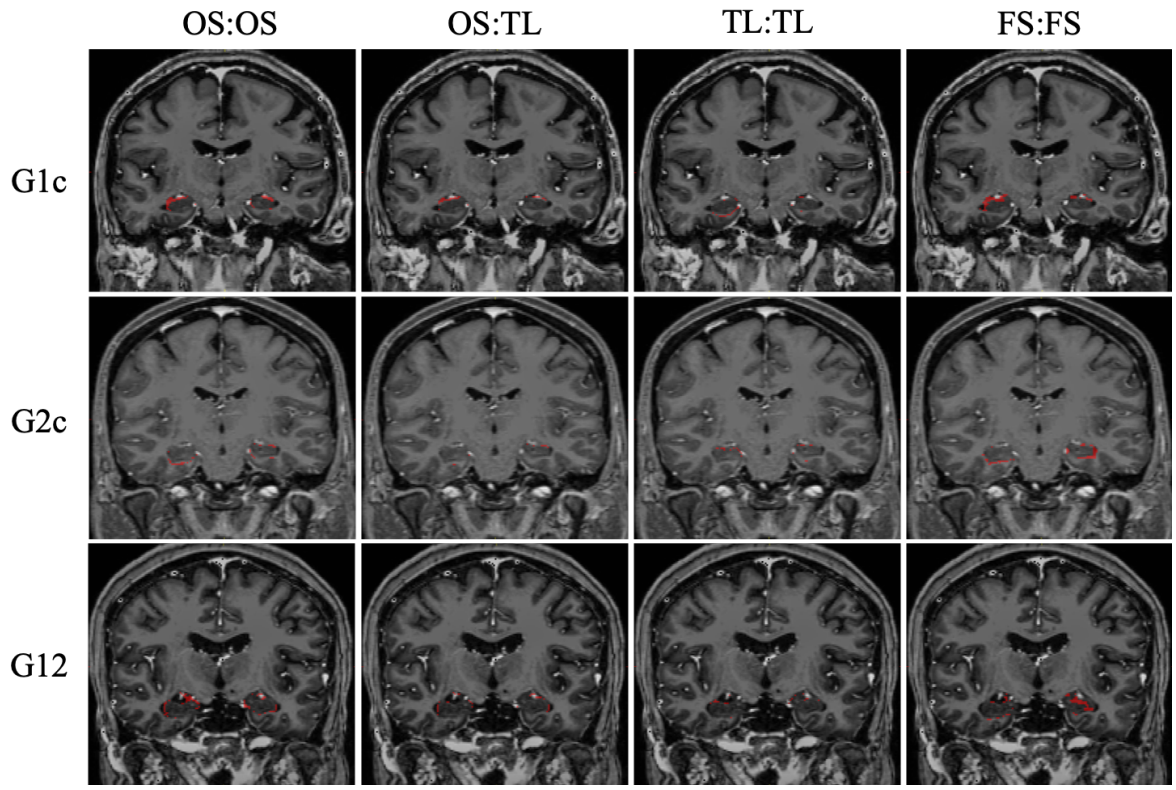

Figure VI.8 Areas of disagreement between pre- and post- segmentations of the hippocampus in the cross-validation dataset.

	OS:OS	TL:TL	FreeSurfer
G1c	568.48	197.13	1,196.77
G2c	497.71	188.82	1,160.79
G12	582.21	203.05	1,018.78

Table VI.5 RMSE between hippocampal volume of paired images in the cross-validation set. First, with the original SLANT (OS), after transfer learning (TL:TL), and with FreeSurfer.

	OS:OS	TL:TL	FreeSurfer
G3c	790.79	169.40	513.82
G13	223.30	251.70	435.04
G23	962.23	480.96	1,586.03

Table VI.6 RMSE between paired images in the withheld dataset. First, with the original SLANT (OS), after transfer learning (TL:TL), and with FreeSurfer.

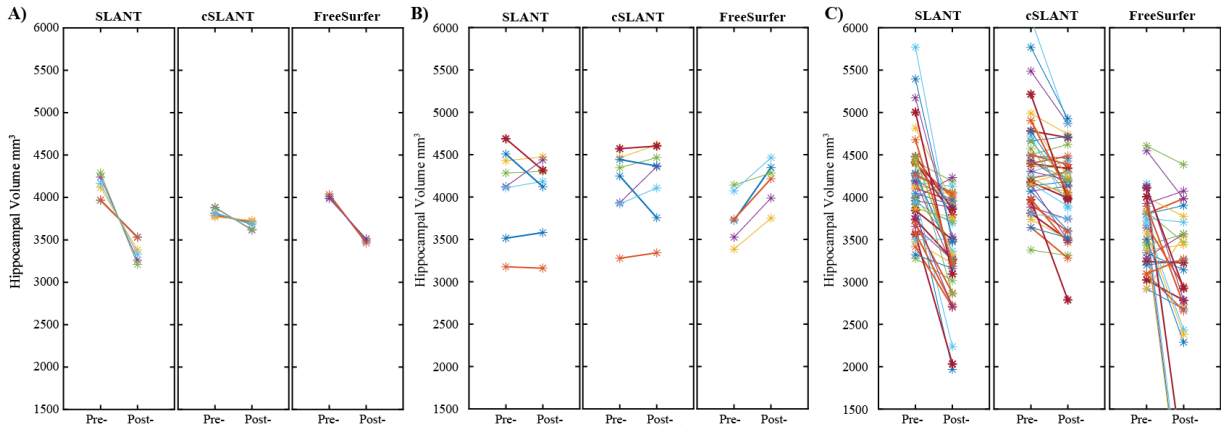

Figure VI.9 Hippocampus volume RMSE on withheld data. Subplots represent performance on paired subgroups for A) Group 3c, B) Group 1-3, and C) Group 2-3.

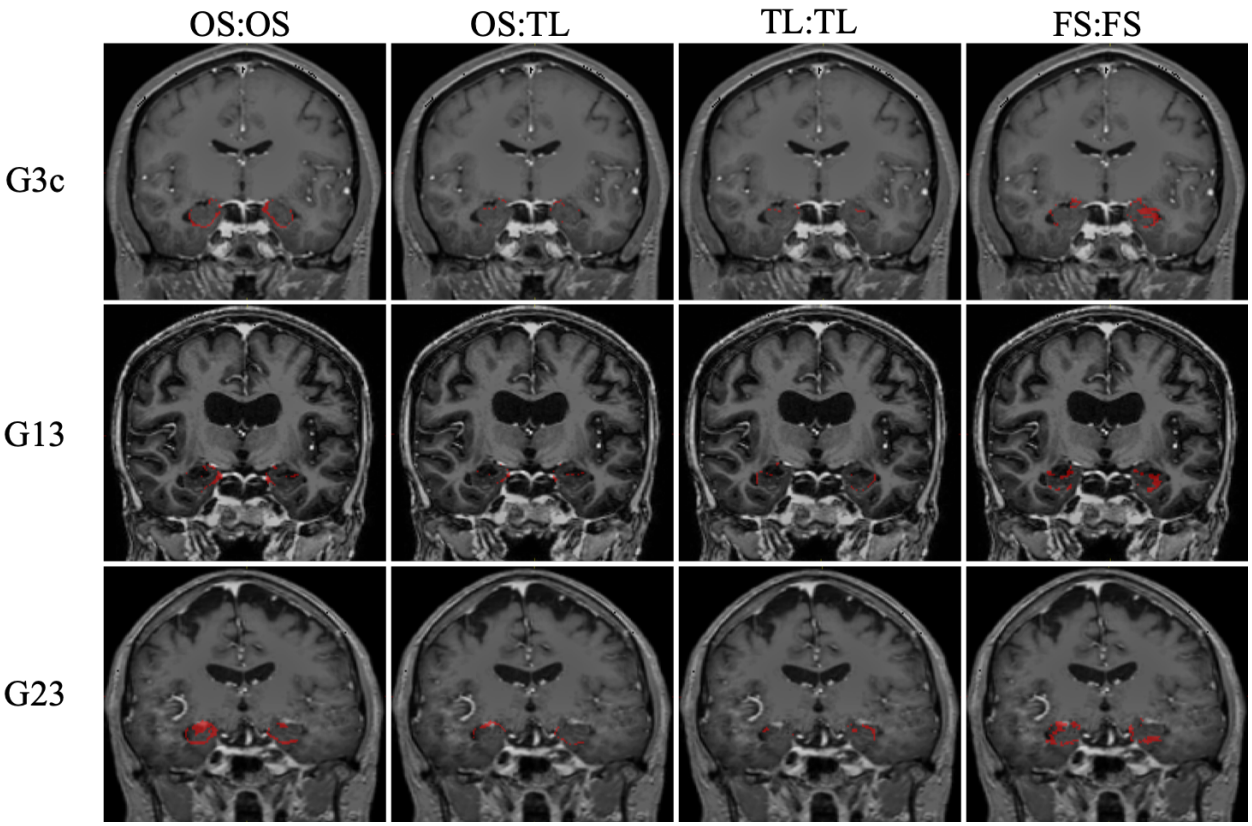

Figure VI.10 Areas of disagreement between pre- and post- segmentations of the hippocampus in the withheld dataset.

4. Discussion

In this work, we explore the effect of TL on the performance of deep neural networks. Our aim was to show the feasibility of taking a well-known segmentation tool such as SLANT and perform domain adaptation with TL by leveraging unlabeled clinical data instead of changing or generating the available code. We used TL to enforce similarity in manually corrected pediatric labels as well as automated segmentation between the pre- and post-contrast images and paired images with different acquisition parameters. We showed that the original SLANT segmentation algorithm has decreased performance in pediatric brains, likely due to smaller volume and altered gray/white matter proportions in the younger subjects compared to the initial training data used in SLANT. However, we can improve performance by 0.08 rDSC when we introduce pediatric subjects in the TL process in addition to the original adult subjects.

We showed that segmentation of difficult regions, such as the occipital poles, can be improved by providing mixed training examples. Importantly, we showed that the proposed method can reproduce the manual corrections done on the original multi-atlas labels using only 30 training examples. Furthermore, the manual corrections used as training examples were generated in a semi-automated method, which provides a time- and resource-efficient alternative to generate training data.

In the context of clinical imaging, we showed an improvement in performance of rDSC of about 0.1 between paired scans across all groups. This shows a significant increase in generalizability as well as robustness to heterogeneous clinical imaging. There is an improvement in the segmentation of the gray-white matter junction as well as areas with intravenous contrast particularly at the ventricles, as shown in Figures VI.4, VI.5 & VI.6. With an increasing demand for larger sample size in deep learning, the field will likely turn to clinically acquired medical data for examples rather than acquiring more research scans. Segmentation algorithms will likely need to be robust to inhomogeneities present in clinically acquired data, such as intravenous contrast, movement, noise, or artifacts.

Volumetric estimates, both in research and in clinical applications, will need to be consistent and reproducible within the same subject despite acquisition parameters or throughout a longitudinal timeline. We showed that cSLANT decreases the RMSE of volumetric estimation within the same subjects by 67%, which will result in more reliable and reproducible image processing on clinical data. The disagreement in segmentations of the hippocampus show that our method performs better at describing the boundary consistently across imaging sequences, with fewer confluent areas of disagreement (Figure VI.8 & VI.10).

We showed that we can achieve domain adaptation on the SLANT algorithm through TL on pediatric and clinically acquired imaging data. This work has great implications for medical and research domains, since it shows that pretrained algorithms can be translated to similar data in a different domain without resorting to separate segmentation pipelines for each domain [280, 281] or alterations to the original image, such as image synthesis [253, 257, 282]. Improving the generalizability of deep learning algorithms across large variability in imaging domains using TL can complement harmonization efforts which achieve a remapping of tissue intensities across small changes in acquisition parameters. Existing harmonization

frameworks may still achieve tissue mapping across the large variability found on clinical data, but such studies are still limited in the literature. Future efforts that explore the integration of intensity harmonization with TL may prove to further improve segmentation accuracy as well as the generalizability of many other medical image processing algorithms. It is possible that a different deep learning framework, such as the one proposed by Kamnitsas et al [283] may achieve higher agreement between imaging domains given further algorithm refinement and hardware resources. Here, we showed that SLANT can be refined for heterogenous clinical imaging through a data-driven process instead of new algorithm modifications or implementations.

A major limitation of this work is that we are enforcing similarity between unlabeled paired images, since we are not introducing any more information to the algorithm in the form of ground truth. We showed that there was a slight but significant decrease in DSC of the manual labels on the OASIS dataset after TL, despite substantial gains on rDSC on the clinical datasets. While this is not surprising, it is expected that as TL seeks to improve site-specific or scanner-specific performance, the overall generalizability to manual labels may decrease. Since the clinical data is unlabeled, the only metric to judge the algorithm accuracy is to compare against manually labelled ground truth of a different dataset. Implementation for site-specific applications may seek to perform additional transfer learning on manual labels generated on local data. Although implementation of this work requires paired data for TL, clinical protocols often acquire a pre-contrast image along with a post-contrast image, so collecting paired contrast imaging may be feasible depending on the imaging bioinformatics tools available at each institution. In this work, we showed that we were able to generalize to a withheld test set with 480 imaging pairs of 255 unique subjects. This will greatly limit the need to acquire new imaging and generate manual labels in order to share or refine medical imaging analysis tools.

Another limitation of this work is that we have evaluated a sampling of acquisition parameters for T1-weighted brain MRI. We showed that our method increased generalizability in the data present and even translated to an improvement in the withheld test set. However, there are still many different combinations of acquisition parameters that result in a T1-weighted brain MRI, which were not evaluated in this work.

Future work will also seek to use transfer learning to improve generalizability of whole-brain segmentation, or other deep learning tasks, on T2-weighted brain MRI and perhaps diffusion-weighted imaging. However, with this work we have shown that there is a valid framework constructed with stable, available algorithms, such as SLANT, to improve performance on clinical imaging. This kind of translation of image processing algorithm from MRI acquired in a research setting towards MRI acquired in the course of clinical practice will have great implication on the application of this tools at the bedside.

This work demonstrates the importance of generalizability in the era of deep learning. We used FreeSurfer as a standard comparison of image processing tools. It has been previously shown that manual correction is necessary as a post-processing step in FreeSurfer with heterogeneous anatomy in imaging data [266, 284, 285]. We observed in this work that the presence of intravenous contrast in particular decreases the generalizability performance between paired imaging. With the advances in performance achieved with deep learning, clinical translation of these tools has begun to emerge [149-151]. To achieve reliable and reproducible results on clinical data, the image processing community will have to demonstrate that the proposed algorithms can achieve a desired performance not only on well-controlled research data, but on heterogeneous clinical imaging. Here, we use hippocampus volumetry as an example of variable measurements across image processing platforms. The largest difference was observed in FreeSurfer, but we still observed a large RMSE in the original SLANT. As personalized imaging biomarkers continue to emerge in research and in clinical practice, it is encouraging that we can drastically improve volumetric estimates through transfer learning.

VII. Neuroimaging Signature of Heart Failure with Preserved Ejection Fraction in the Setting of Dementia

1. Introduction

Alzheimer's Disease and other dementias affect more than 5.3 million Americans and result in over $259 billion in medical expenditures annually[105]. Epidemiologic evidence suggests that heart failure with preserved ejection fraction (HFpEF) may be an independent risk factor for dementia disorders[107, 108, 286]. However, HFpEF and dementia also share many common risk factors such as advanced age[112], obesity[110], smoking[110], midlife hypertension[110], and diabetes[110]. However, it is unclear whether HFpEF confers dementia risk independently, in combination with, or through interactions with other comorbidities.

Several lines of evidence support an independent role for HFpEF in neurocognitive disease. For example, patients with HFpEF have specific changes in neuropsychological testing, particularly in executive function, attention, and memory, as well as and brain morphology compared to healthy adults[109, 113, 114]. Additionally, the proportion of vascular dementia to Alzheimer's disease is increased among patients with HFpEF [112, 115]. In parallel, advances in quantitative medical image processing and machine learning have made possible the discovery of imaging biomarkers, at scale, for disease prediction, detection, and surveillance[17, 39]. For example, these efforts have demonstrated associations between decline in cerebral blood flow and higher Framingham Cardiovascular Disease Risk Profile (FCRP) score[121] as well as left ventricular mass index and changes in white matter microstructure[124]. Collectively, these observations suggest that HFpEF has a primary effect on brain morphology, which may result in a distinct dementia phenotype. Moreover, an improved understanding of the imaging predictors of dementia in the setting of HFpEF may help characterize pathophysiologic mechanisms and lead to improved diagnostic and therapeutic strategies.

We hypothesized that dementia in the setting of HFpEF would be associated with distinct comorbidity and brain morphology imaging signatures compared to dementia in the absence of HFpEF. To assess this HFpEF-dementia phenotype, we performed a case-control retrospective analysis of clinical diagnoses and brain imaging from 331 ambulatory patients with dementia from Vanderbilt University Medical Center (VUMC). We used a bioinformatics framework to identify and validate an HFpEF-dementia cohort and compare its clinical phenotype to that of dementia alone. We also performed a volumetric analysis for changes in brain morphology associated with HFpEF-dementia identified on brain MRI. We also performed a PheWAS analysis to find the clinical comorbidities associated with HFpEF-dementia compared to dementia alone. The work presented in this chapter is currently under review at *Journal of the American Heart Association* with Camilo Bermudez as first author.

2. Methods

2.1. Cohort Selection

The study cohort was derived from all subjects in the VUMC EHR. We first identified all possible subjects using a computable dementia algorithm that included both ICD-9 and ICD-10 codes (Figure VII.1). We next limited the sample to patients with head MRI scans documented in the VUMC clinical imaging database ImageVU. Finally, we restricted the dataset to only patients with evidence of cognitive testing in their EHR. Specific strings used to identify cognitive testing in the chart are shown in the red box in Figure VII.1. HFpEF status was defined using a validated algorithm [287]. In order to validate the diagnosis of clinical dementia, we conducted text searches for cognitive testing to extract segments of the medical chart for manual review (Figure VII.1). We used the following cutoffs to validate the diagnosis of dementia: a) mini-mental status exam score $\leq$ 24, b) Montreal Cognitive Assessment $\leq$ 22, or c) formal neuropsychological assessment showing cognitive impairment.

Last, we limited the sample to those patients with a high-quality brain MRI imaging. Quality assessment was performed by visual inspection to ensure that the studies retrieved were in fact brain MRIs

without large artifacts that would limit structural analysis. The MRI scans were further sorted into T1-weighted brain MRI and T2-FLAIR brain MRI for subsequent studies. Only T1-weighted MRI with a resolution of 2.2 mm or finer were used.

We identified 5,913 subjects meeting criteria for an automated diagnosis of dementia, including 1,092 (18.5 %) with HFpEF (Figure VII.1). The diagnosis of dementia was validated in 1,654 patients, of which 393 (24%) had HFpEF. After selecting for patients with high-resolution imaging, the final cohort included 331 patients with dementia and high-resolution T1-weighted brain MRI, 30 (9%) of which met criteria for HFpEF.

Participant characteristics are provided in Table VII.1. All subjects with HFpEF and dementia were matched 10:1 to patients with dementia alone for age, sex, and image resolution. The mean age in the entire sample was 76.2 ± 8.48 years; 59% were women. The mean duration of dementia at time of imaging was 0.79 ± 1.75 years for cases and 1.15 ± 2.14 years for controls ($p = 0.65$; Wilcoxon rank-sum test) as indicated by the time between first incidence of dementia ICD code and imaging. After all inclusion criteria and matching, there were a total of 30 cases of HFpEF and dementia and 301 controls of dementia alone.

	Heart-Failure and Dementia (N =30)	Dementia without Heart Failure (N = 301)
Age (years)	76.9 ± 8.12	76.2 ± 8.52
Sex (% female)	60.7%	59.8%
Dementia Duration at MRI (years)	0.82 ± 1.83	1.20 ± 2.23
Heart Failure Duration at MRI (years)	3.29 ± 3.67	N/A
Hypertension (%)	100%	92.7%
Diabetes (%)	53.6%	19.6%

Table VII.1 Cohort demographics. Values represented as mean ± SD.

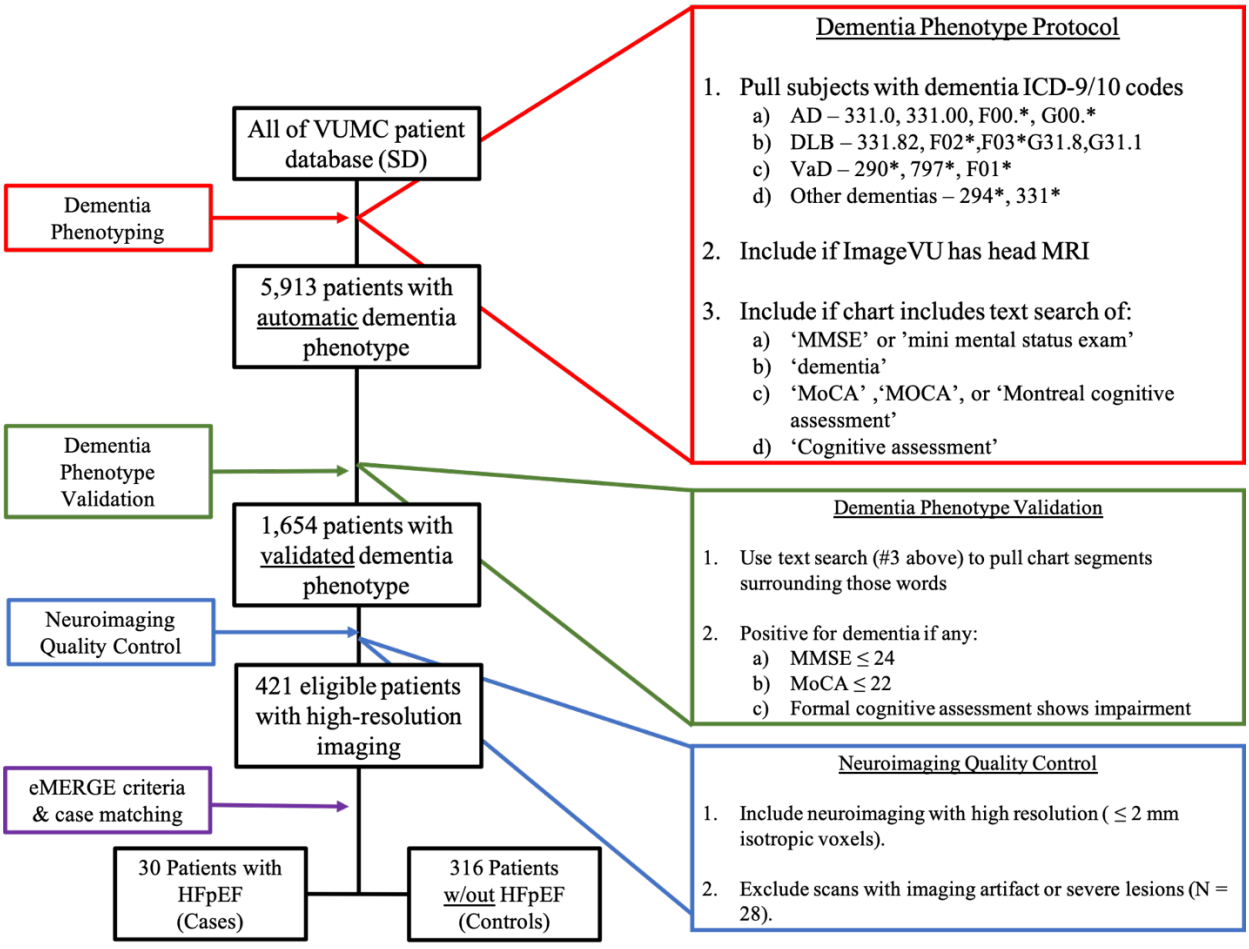

Figure VII.1 Inclusion and exclusion criteria protocol to identify patients with dementia and neuroimaging with and without heart failure.

2.2. Quantitative Image Analysis

We performed an automatic segmentation protocol on all T1-weighted brain MRI using the SLANT segmentation algorithm, which has been previously described and validated on clinical imaging[147, 243, 288]. SLANT performs a whole-brain segmentation into 132 cortical, subcortical, cerebellar, and white-matter regions of interest (ROI) following the BrainCOLOR labelling protocol[147]. The volume of each ROI in mm^3 was calculated by multiplying the number of labelled voxels by the voxel size. We used the automated library SIENAX from Fressurfer v6.0 to account for head size, which has been previously described and validated [137, 289].

2.3. Statistical Analysis

Two independent analyses were conducted: a PheWAS study to identify clinical variables predictive of HFpEF dementia and a volumetric MRI study to detect changes in regional brain volumes in HFpEF dementia. In order to characterize the clinical differences between cases and controls, we conducted a phenome-wide association study (PheWAS). The PheWAS method uses a validated, curated medical phenotypes (PheCodes) to rapidly identify phenome-wide association between exposures including clinical variation and disease phenotypes[116, 119, 120]. PheWAS codes are validated groupings of related ICD-9 (before October 2017) and ICD-10 (after October 2017) billing codes that capture the extended range of clinical diagnoses within an EHR data set.[23,24] We used a logistic regression model to identify the relationships between PheCodes and the presence of HFpEF in the setting of dementia versus no HFpEF in patients with dementia while matching for sex and age at imaging. Of the total 1,866 codes, we excluded phenotypes affecting a single gender or with ≤5% prevalence, which resulted in 397 clinical phenotypes included in the analyses. A Bonferroni correction was applied to account for multiple testing.

The second analysis used logistic regression to assess relationships between volumetric estimates of ROI from T1-weighted MRI and exposure to heart failure in the setting of dementia. To standardize volumetric measurements across participants, we included the SIENAX scaling factor for intracranial volume as a covariate to adjust for head size. Associations with a false discovery rate q value <0.05 were considered statistically significant.

3. Results

3.1. Anatomical Volumetry Differences Associated with Heart Failure

Brain volumes for six regions of interest were significantly different ($q<0.05$) between dementia patients with and without HFpEF (Table VII.2). Results for all regions are shown in the Appendix Table IX.2. All of the significant regions showed atrophy associated with exposure to HFpEF. The six regions with volume loss associated with HFpEF were the left accumbens, right amygdala, left posterior insula, left anterior orbital gyrus, right angular gyrus, and right cerebellar white matter. Five of these six significant regions are located in the temporo-parietal region of the brain, which is proximal to the posterior watershed territory between the middle cerebral artery and the posterior cerebral artery [290].

3.2. Clinical Comorbidities Associated with Heart Failure

PheWAS analysis identified a large number of clinical comorbidities associated with HFpEF in the setting of dementia (Figure VII.2). Not surprisingly, HFpEF with dementia was associated with a wide range of phenotypes dominated by cardiovascular disease and associated comorbidities, such as respiratory and metabolic abnormalities. All significant clinical phenotypes were positively associated (i.e., more common) with the presence of HFpEF. Importantly, dementia was not a significant comorbidity, since it is represented in both cases and controls. Details for model coefficients for the PheWAS analysis are shown in the Appendix Table IX.3.

Region of Interest	Log Odds Ratio	p-value
Left Accumbens Area	-4.4×10^{-3}	0.020
Right Amygdala	-2.5×10^{-3}	0.020
Left Posterior Insula	-1.6×10^{-3}	0.012
Left Anterior Orbital Gyrus	-1.4×10^{-3}	<0.001
Right Angular Gyrus	-2.0×10^{-4}	0.043
Right Cerebellar White Matter	-2.0×10^{-4}	0.04

Table VII.2 Brain regions with significant atrophy associated with the presence of HFpEF. P-values are corrected for multiple comparisons using the Bejamini-Hochberg method for false discovery rate.

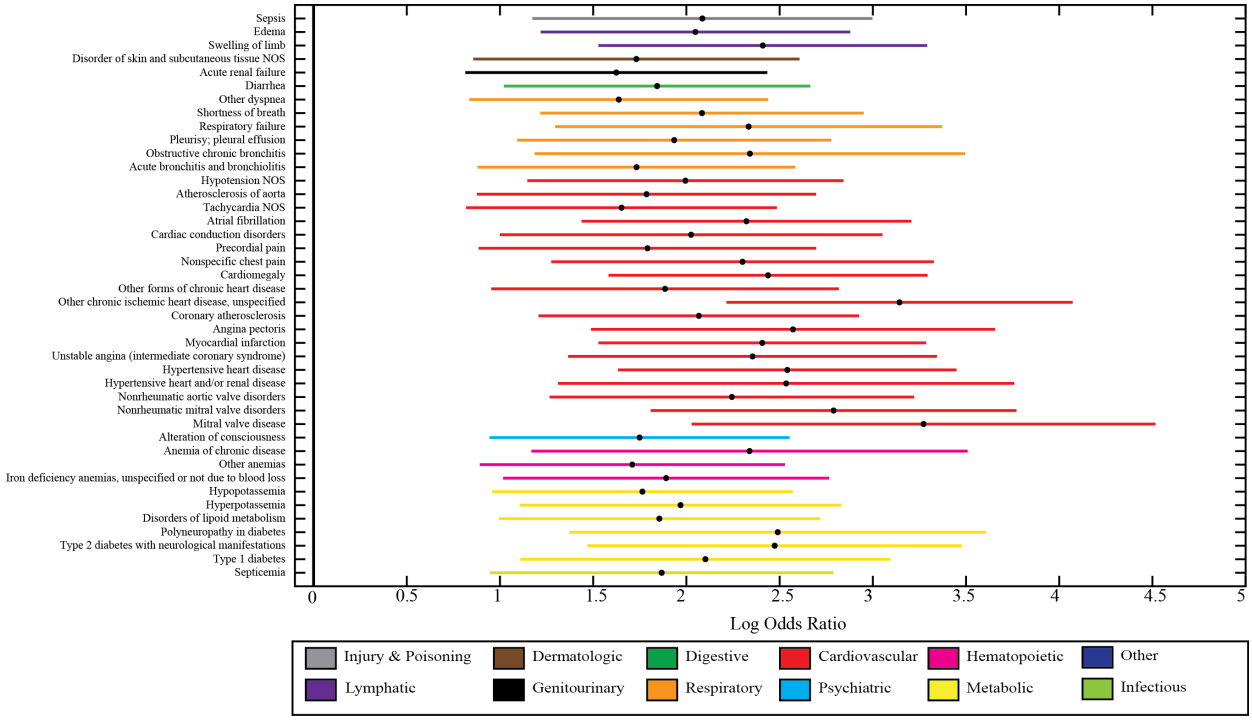

Table VII.3 PheWAS analysis done on HFpEF dementia cohort. Significant clinical phenotypes associated with heart failure with preserved ejection fraction (HFpEF) in patients with dementia. Clinical phenomes are organized by organ system. A full table of significant clinical variables is shown in Appendix Table IX.3.

4. Discussion

We extracted clinical data with neuroimaging to identify the clinical comorbidities and changes in brain imaging morphology of HFpEF-associated dementia. Collectively, the findings suggest that HFpEF has a unique clinical and anatomic signature from dementia in the absence of HFpEF. Moreover, this study demonstrates the feasibility of mining large clinical datasets to identify unique clinical and radiological signatures.

Previous work has shown associations between cortical atrophy and cerebral hypoperfusion in the temporo-parietal watershed territories in Alzheimer's disease[291, 292]. Of these regions, atrophy of the accumbens area and the amygdala has been associated with HF before, suggesting that HF may increase risk of Alzheimer's disease through early insult to subcortical brain structures[293]. Similarly, atrophy of the insula[294] and the orbital gyrus[295] has been previously reported in vascular dementia while

reduction of cerebral blood flow was shown in the angular gyrus in patients with Alzheimer's disease[296]. Previous work has also shown atrophy of the cerebellar white matter is associated with heart failure when compared to controls[297]. Additionally, we observed a lateralizing effect on these structures. Previous work has shown a lateralizing effect on brain atrophy in semantic dementia, particularly on temporal lobe structures like the entorhinal cortex, amygdala, and fusiform gyrus[298] as well as the hippocampus in the context of HF[299]. These associations may require further study to understand the mechanism of lateral differences in brain volume and how they are related to HFpEF's effect on brain perfusion and anatomy.

Importantly, the literature of brain structural changes associated with HF has focused on comparing patients with HF to healthy controls without taking into account the presence of clinical dementia[293, 299, 300]. In this work, we make similar observations to previous work, but focus on HF patients with dementia compared to controls with dementia. This suggests that the atrophy pattern identified may correspond to a HFpEF signature independent of the presence of dementia. Moreover, the neuroimaging signature of HFpEF in the setting of dementia seems to correspond to vascular watershed territories, suggesting that chronic hypoperfusion to these areas may result in the development of cognitive decline. This observation is also supported by the lack of significant atrophy in areas commonly associated with Alzheimer's disease, such as the hippocampus, since the effect of dementia may be accounted for in both groups.

To explore the possibility of atrophy due to an HFpEF-dementia interaction effect, we considered the effect size of non-significant regions (Appendix Figure VII.3). We note that only four regions have a larger effect size than the significant regions found in this study, and two of these correspond to the contralateral side of a significant structure. It is possible that a larger study would result in these structures being significant. Conversely, there are a large number of non-significant regions with atrophy where the lower bound of the confidence interval of the effect size is bounded by -1×10^{-3}. This suggests that regions such as the hippocampus, parahippocampal gyrus, frontal cortex, and the entorhinal area are possible candidates for HFpEF-dementia specific atrophy, but this effect is much smaller than that of HFpEF. Future studies may seek to study these regions directly to identify interaction effects between HFpEF and dementia. Lastly, a limitation of this work is the absence of a HFpEF group without dementia. However, it

is difficult to assume that clinical brain imaging acquired from patients with HFpEF would be cognitively normal. A prospective imaging study in patients with HFpEF and cognitive testing at the time of imaging would be necessary to account for brain changes due to HFpEF alone.

Some of the most significant clinical comorbities found in our PheWAS study were associated with the cardiovascular system. This is not surprising, as HFpEF occurs in context of other cardiovascular disease and its risk factors, which may play a role in the development of dementia[106, 113]. We also found previously reported HFpEF comorbidities in other organ systems, such as metabolic abnormalities, hematologic, and respiratory conditions[301]. A study to validate the presented systematic exploration in the context of PheWAS novelty[302] would be of interest. It is interesting that there were no clinical comorbidities inversely associated with HFpEF in the setting of dementia, suggesting that, in general, patients diagnosed with HFpEF are more medically ill than individuals with dementia without HFpEF, confirming a previously reported finding [303].

The protocol for phenotype selection shown in Figure VII.1 uses a semi-automated method of data extraction and validation to identify a clinical phenotype. These methods can be used to identify other clinical phenotypes with scarce datasets and examine associations with related diseases. One limitation of this approach is the role of important cardiovascular comorbidities on dementia, such as the presence of hypertension, diabetes, or smoking status. Future work may take these variables into account in both clinical phenome regressions as well as volumetric regressions to further understand the specific role of HFpEF in dementia. Furthermore, we included only patients who had undergone a brain MRI in order to perform volumetric segmentations. It is possible there are still a large number of patients with dementia and heart failure who have been diagnosed without acquiring neuroimaging. The clinical reasoning behind a lack of imaging, whether due to disease severity or other patient circumstances, may be a confounder in the description of a clinical phenotype as well as morphological brain changes.

In this work we demonstrate that a unique clinical and neuroimaging signature can be identified from clinical data on a narrow cohort. This approach can be applied to other large clinical datasets to examine clinical and imaging-based population differences in disorders and phenotypes which are difficult

to study prospectively. For our findings in dementia and HFpEF, further work is required to better characterize these populations and determine how management of vascular risk factors and HFpEF may ameliorate cognitive decline.

VIII. Conclusions & Future Work

1. Conclusion

The primary goal of this dissertation is to contribute to the understanding of how aging affects brain anatomy through the lens of neuroimaging. The motivation to study aging through neuroimaging stems from two main drives: a biological need and a technological opportunity. In the decades to come, the relative proportion of the elderly in the general population is estimated to increase. Aging is the result of a remarkably complex interplay of internal and external factors, resulting in highly variable outcomes in health and lifestyle for the individual and the use of resources for the healthcare system. However, a common resulting phenotype of aging is the development of cognitive decline and dementia. Neuroimaging offers the opportunity to visualize and measure how brain anatomy changes with age. Meanwhile, technology today offers new opportunities to store, share, and analyze large amounts of imaging data. At the time of this dissertation, advances in hardware made it possible to scale deep neural networks to tasks in computer vision and therefore apply them to medical imaging. Together, neuroimaging and deep learning offer the possibility of new types of analyses to better characterize brain morphology. By working on the engineering to generate data-driven, reliable, and reproducible metrics of brain anatomy, future work may rely on these to link brain morphology to cognitive performance with age as well as other diseases associated with aging.

The work done in this dissertation explored how deep neural networks can be applied to the field of medical image analysis in order to better study the aging brain. We developed a tool to accurately predict brain age in healthy individuals by introducing anatomical context in the form of traditional imaging features to both MRI and CT (Chapter II), and showed the importance of these anatomical features to protect against adversarial attacks (Chapter III). We then extended this work as an application by demonstrating that brain age gap may be a potential biomarker of depression in the elderly (Chapter IV). We then took

another look at segmentation of brain structures using deep learning, first focusing on the dentate nucleus of the cerebellum (Chapter V) then moving towards generalization of whole brain segmentation towards heterogenous data such as pediatric or clinically acquired imaging (Chapter VI). This work culminated in the application of these tools to study the neuroimaging signature of heart failure with preserved ejection fraction in patients with dementia by mining clinical imaging archives (Chapter VII).

2. Introducing Contextual Anatomical Features to Brain Age Estimation

2.1. Summary

Prediction of chronological age had been previously focused on using engineered imaging features or raw imaging from T1-weighted brain MRI alone. We proposed the first work to dynamically introduce engineered features, in the form of volumetric estimates of brain regions, during training of an imaging-based deep convolutional neural network (Chapter II). We also showed that these tools can be applied to develop age estimators for new imaging domains, such as orbital CT (Chapter II). We were also the first to show the susceptibility of brain age neural networks to adversarial noise attacks and the protective effects of introducing anatomical context against these attacks (Chapter III). We then provided clinical validation of the applicability of brain age on T1-weighted brain MRI by showing an increased brain age gap in elderly patients with depression when compared to healthy controls (Chapter IV). Overall, we generated an accurate and robust age estimator able to detect differences in disease states in both brain MRI and orbital CT.

2.2. Main Contributions

- We introduce the idea of enhancing deep learning with traditional engineered features in medical imaging to achieve higher accuracy in the task of age prediction from brain MRI and orbital CT.
- We show how some diseases of the orbit resulted in changes in the orbital CT age gap, suggesting this as a possible non-invasive biomarker of eye disease.

- We showed the susceptibility of medical imaging to adversarial noise and possible implications in the deployment of these tools to patient care. We provide insights into possible mechanisms for robustness against adversarial attacks using patient-specific anatomical context.
- We showed the clinical applicability of our MRI brain age estimator by finding differences in the Brain Age Gap between patients with depression and controls in an adult and a geriatric population. Furthermore, we showed associations between Brain Age Gap and cognitive performance in the geriatric cohort.

2.3. Discussion

The idea of introducing contextual features while training deep neural networks has evolved to focus on activation maps, which drive spatial attention towards regions with better performance. Although the role of engineered features seems to have been largely replaced by learned neural network features, there is still extensive domain knowledge that goes into crafting these features, which can be leveraged to improve deep learning accuracy, generalizability, or robustness. Here, we showed that by introducing volume estimates of brain regions, we could drive improvement in performance and protect against adversarial attacks. The goal was to enhance the network performance by presenting it with relevant imaging features. It will be interesting to see how non-imaging features, such as clinical features (e.g. labs, vitals, PheCodes) are integrated in image-based deep learning tasks to drive accuracy. This may play an important part as imaging datasets become larger and more heterogeneous. Non-imaging data may provide high-level separations in patient phenotypes, creating subgroups during training and therefore allowing the network or networks to refine features specific to each subgroup for higher accuracy and generalizability.

We also opened the idea of crafting tools to predict age in different locations. We showed that the same principles apply for age prediction in Orbital CT and were able to detect differenced in disease states. Aging is a multi-system process that affects every part of the human body. With enough data, future work may predict age in abdominal CT, chest X-rays, bone radiographs, or ultrasound, especially in modalities that already have expert-crafted features such as abdominal organ segmentation. Each location may provide

insight into the local effects of aging and even hint at mechanisms behind different diseases. As radiologic studies become more prevalent and feasible to analyze, tools like predicted age can provide a non-invasive biomarker not only to detect and tract disease but to tract aging as a form of "growth curve" for the elderly.

We showed that we can find differences in the gap between predicted age and true age using orbital CT in the case of eye diseases and brain MRI in depression. These retrospective, cross-sectional studies demonstrate the feasibility of validating non-invasive imaging tools to find differences between disease states. Future work may choose to use these tools prospectively to better track the progression of disease and better understand the clinical scenarios where disease progression diverges, as in the case mild cognitive impairment and Alzheimer's disease. One major factor limiting the implementation of current technology to clinical practice is the barrier of interpretability of neural networks. Although deep neural networks can achieve high accuracy, it is often difficult to probe what anatomical regions are informing the network. For instance, we were able to detect differences in brain age gap in patients with depression, but cannot at this time quantitatively describe the morphological changes in the brain that drive this response. In order to better assess the potential uses of these tools in patient care and generate mechanistic hypotheses of disease, much work needs to be done to better visualize and interpret what anatomical information is being integrated in convolutional networks to drive performance.

Lastly, we provided a characterization of the susceptibility of brain age prediction to adversarial noise at several intensity levels. Although anatomical features were able to reduce this effect, adversarial learning was still able to drastically change the prediction results. This susceptibility presents another limitation to implement medical image processing algorithms to clinical practice, since small, invisible perturbations may drastically alter the result and could have great implications in computer-assisted detection and diagnosis. Some of these effects can be attenuated by introducing noise during training, but it is still possible to learn new adversarial noise fields to alter prediction results. Future work may follow the steps we presented in our work, and use contextual features, image-based or otherwise, to protect against adversarial attacks. Even if these features do not drastically improve network performance, it may be less susceptible to noise changes in the original image. Together with information security measures, new tools

will need to be developed in radiologic systems to detect and protect against noise perturbations in clinical image acquisitions.

3. Segmentation of Brain Structures from Heterogeneous Datasets

3.1. Summary

In recent years, advanced in deep convolutional neural networks showed a dramatic improvement in automatic segmentation of desired structures from images. We began with the first deep learning segmentation of the dentate nucleus (DN) of the cerebellum using multimodal imaging (Chapter V). This work showed that deep learning can better reproduce manual labels than atlas-based methods (Chapter V). We then implemented domain adaptation to whole brain segmentation, using label augmentation to improve segmentation of imaging datasets that are not represented in the research data, such as pediatric or post-contrast brain MRI (Chapter VI). We showed that we can leverage paired, unlabeled clinical data to better generalize whole brain segmentation across many acquisition parameters (Chapter VI). These efforts contribute towards making research algorithms more generally available across different sites, scanners, and acquisitions while preserving performance across domains.

3.2. Main Contributions

- We showed that deep learning segmentation of the DN can better reproduce manual labels than atlas-based methods.
- Introducing co-registered structural MR images during segmentation of the DN did not dramatically improve segmentation accuracy over using only the image in which the labels were generated.
- We showed semi-automated label augmentation can be used to perform domain adaptation via transfer learning to improve the accuracy of whole brain segmentation in pediatric imaging. Additionally, this technique does not largely alter segmentation performance on the original dataset.

- We showed that automated label augmentation can be used on paired, unlabeled clinical data to perform domain adaptation via transfer learning to improve whole brain segmentation in brain MRI with intravenous contrast. We showed that this technique can generalize to image acquisitions excluded from training. Additionally, this technique does not largely alter segmentation performance on the original dataset.
- We showed the importance of domain adaptation for volumetric estimation of brain regions of interest. Specifically, we showed a decrease in bias and variance in volume estimation of the hippocampus between imaging pairs of the same subject compared to non-harmonized methods in deep learning and common neuroimaging pipelines.

3.3. Discussion

Segmentation of substructures in the cerebellum continues to be a challenging task in medical image processing. Here, we showed that we could better reproduce manual labels generated on FA maps using deep learning than multi-atlas or single-atlas methods. We found that the diffusion-based labels differed from templates generated from susceptibility-weighted imaging (SWI). Although we hypothesize that this is due to the susceptibility effect of iron in the DN, future work may indeed be required to explore the effect of iron content on DN contrast across imaging modalities. If true, these two segmentation methods would prove valuable as complimentary measures of DN structural integrity and could be used to study the role of iron deposition in deep cerebellar structures in aging and neurodegenerative diseases, particularly movement disorders.

Generating reliable segmentations on pediatric imaging is a challenge not only due to site or scanner effects, but also the evolving grey-white matter contrast during the early years of development. For this reason, it is desirable to have flexible algorithms that can adapt to local data. We learned that we can leverage pre-existing algorithms with some manual editing to gain a meaningful improvement in segmentation performance. Here, we applied this work to pediatric subjects ages 2-4 years old, but a more challenging task would be infant imaging (<1 year old), where the grey-white matter contrast is much more

difficult to discern. However, the approach used for paired, contrast imaging can be leveraged to improve pediatric segmentation in longitudinal studies. For instance, infant labels could be learned from pediatric labels in the same subject after appropriate coregistration. This would help overlay the segmentation in the infant imaging and assist in the detection of the grey-white boundary.

Similarly, whole brain segmentation on imaging with intravenous contrast is an active area of research, particularly when deep convolutional neural networks are susceptible to inhomogeneities in the image, such as contrast. The work here shows that even if an algorithm is trained on non-contrast data alone, it is possible to alter the parameters slightly to achieve high performance on a contrast dataset. Furthermore, we showed a decrease in bias and variance of volume estimations in the hippocampus on clinical data. This opens many areas of exploration, including measurement of key structures in longitudinal studies where neurodegeneration is suspected. Moreover, these tools also encourage the use of clinical data for research and potentially clinical studies. A more robust estimation of hippocampal volume may be able to better detect changes in neuroanatomy associated with aging and disease from a patient population. Prospective studies using these tools on clinical data now have more imaging data available for large-scale analyses. Lastly, new methods may seek to integrate deep learning domain adaptation with techniques in image intensity harmonization to achieve better correspondence between acquisition types. Together, this work suggests an avenue to share algorithms and improve generalizability in imaging without having to share data or create massive imaging datasets in a single site.

4. Neuroanatomical Signature of Heart Failure with Preserved Ejection Fraction

4.1. Summary

As the population ages, neurological and cardiovascular diseases comprise a large part of comorbidities affecting patient lifestyle and healthcare resources. Understanding the interplay of these two diseases can result in a better understanding of disease and optimization of treatment strategies. Here, we used clinical imaging data mined from the electronic health record to describe the neuroimaging signature

of heart failure with preserved ejection fraction (HFpEF) (Chapter VII). We demonstrated the feasibility of extracting and validating a desired patient phenotype derived from routine clinical visits. We then analyzed the imaging data using volume estimates robust to clinical imaging (Chapter VI) to find structural changes in the brain associated with a chronic heart condition associated with aging (Chapter VII).

4.2. Main Contributions

- We demonstrated the feasibility of crafting a patient cohort via data extraction from the VUMC EHR using ICD-9 and ICD-10 codes for dementia and the presence of heart failure. We further included only subjects who had at least one head MRI documented in the clinical imaging database ImageVU and evidence of some form of cognitive testing on their medical chart supporting the diagnosis of dementia. Together, these tools show the availability of large-scale imaging datasets to study aging and disease directly from patient records.
- We analyzed the changes in patient phenotype according to PheCodes with respect to the presence of high-resolution brain imaging within the described cohort. Our findings show that filtering for high-resolution results in a different patient phenotype, suggesting some degree of clinical inference that dictates ordering such studies.
- Using clinical imaging, we found regional atrophy in the parieto-temporal regions characteristic of HFpEF patients with dementia. We found that these effects reflect those found in previous literature comparing HFpEF against healthy controls, suggesting that these may be independent heart failure effects. The interaction effect of HFpEF and dementia on brain anatomy is likely smaller than could be detected.

4.3. Discussion

Deep learning tools have shown a dramatic increase in performance over the past few years, with perhaps yet more improvements coming as dataset sizes increase through collaborations and data retrieval. Instead of acquiring more research scans to drive performance, the field of medical imaging may choose to

focus on tools and techniques that leverage pre-existing data, such as that found in electronic health records. These data are not without problems, often coming with heterogenous acquisition parameters, noise, and poor annotations. Future work will have to focus on extensive quality assessment and harmonization efforts to detect the signal present in the data. However, this is a key step towards translating image processing tools to assist in patient care. Here, we showed that this process is indeed feasible and it opens the possibility of studying clinical phenotypes that are not possible through research subject recruitment, such as critically ill patients or rare diseases.

In our work, we observed a large drop off in eligible patients with high-resolution imaging. Instead, most of these patients had low-resolution acquisitions along each axis. Work dedicated at harmonizing the information present in each view could reconstruct an approximate rendering of a 3D acquisition. However, there are still many applications available for 2D imaging, such as classification, lesion detection, and biomarker regression, to name a few. More importantly, further work is required to understand the change in clinical phenotype when selecting patients with high-resolution imaging. One approach would be prospective, since there are some applications that are known to request high-resolution imaging, such as pediatric patients, searching for epileptogenic lesions, or presurgical plan on deep brain stimulation surgery. Another possibility may be that imaging protocols change over time and high-resolution imaging is acquired depending on the specialty, provider, or year. This may introduce bias in the selection of patients that undergo further analyses relative to the whole patient population. Overall, clinical imaging offers a new frontier of opportunities for phenotype exploration, but there are still many challenges to solve in order to harness the full potential of these datasets.

An important step forward will be to develop tools that can quantitatively describe the association between imaging data with clinical data, such as labs, medications, or PheCodes. Understanding how brain anatomy changes in response to clinical variables would provide some explainability to regional differences in deep convolutional neural networks. One such method would be to use large imaging datasets to describe a manifold between imaging and clinical variables, similar to the work done in natural language processing. This would enable exploration of this manifold to measure how changes in one domain affect the other.

However, this would require large amounts of well-curated data, but perhaps well-known associations, such as genetic mutations leading to Alzheimer's disease and hippocampal atrophy may lead the way to the development of these tools.

5. Concluding Remarks

With the development of deep learning, international learning communities, and powerful methods of data extraction and curation, there are yet many discoveries to be made in this field. In this dissertation, we addressed some of the challenges of using clinical data to describe aging and its associated diseases. We generated data-driven metrics to extract information from imaging that we can the associate with patient phenotypes to illuminate underlying biological changes. We showed the feasibility of linking research tools to clinical data and considered many of the challenges ahead to make better use of these data.

Our ultimate goal is to improve the lives of patients through our work. Although aging is a process that affects everyone of us, it does not affect us all equally. By bringing this technology to the bedside, we can leverage large amounts of clinical data available and provide patients with personalized, quantitative, actional measures of health and aging as well as disease. We can help patients and physicians better understand what is happening in their bodies and what we can change to help them. The amount of medical information is growing rapidly and, now more than ever, we are given the opportunity to help thousands of people by measuring it, analyzing it, and explaining it. Now is a remarkable time for medical image processing.

IX. Appendix

1. Chapter VI

Figure IX.1 Reproducibility DSC (rDSC) on each ROI. The table with corresponding ROI names is provided below (Table 6).

Figure VI.10 shows the results of cSLANT compared to baseline SLANT for each ROI individually. Each box indicates the rDSC between scan pairs for original slant (OS), and SLANT after transfer learning (TL) for each of the 132 ROIs. A lookup table is also included below to indicate the anatomical structure corresponding to each ROI.

ROI #	Anatomical Structure	ROI #	Anatomical Structure
4	3rd Ventricle	138	Right MCgG middle cingulate gyrus
11	4th Ventricle	139	Left MCgG middle cingulate gyrus
23	Right Accumbens Area	140	Right MFC medial frontal cortex
30	Left Accumbens Area	141	Left MFC medial frontal cortex
31	Right Amygdala	142	Right MFG middle frontal gyrus
32	Left Amygdala	143	Left MFG middle frontal gyrus
35	Brain Stem	144	Right MOG middle occipital gyrus
36	Right Caudate	145	Left MOG middle occipital gyrus
37	Left Caudate	146	Right MOrG medial orbital gyrus
38	Right Cerebellum Exterior	147	Left MOrG medial orbital gyrus
39	Left Cerebellum Exterior	148	Right MPoG postcentral gyrus medial segment
40	Right Cerebellum White Matter	149	Left MPoG postcentral gyrus medial segment
41	Left Cerebellum White Matter	150	Right MPrG precentral gyrus medial segment
44	Right Cerebral White Matter	151	Left MPrG precentral gyrus medial segment
45	Left Cerebral White Matter	152	Right MSFG superior frontal gyrus medial segment
47	Right Hippocampus	153	Left MSFG superior frontal gyrus medial segment
48	Left Hippocampus	154	Right MTG middle temporal gyrus
49	Right Inf Lat Vent	155	Left MTG middle temporal gyrus
50	Left Inf Lat Vent	156	Right OCP occipital pole
51	Right Lateral Ventricle	157	Left OCP occipital pole
52	Left Lateral Ventricle	160	Right OFuG occipital fusiform gyrus
55	Right Pallidum	161	Left OFuG occipital fusiform gyrus
56	Left Pallidum	162	Right OpIFG opercular part of the inferior frontal gyrus

57	Right Putamen	163	Left OpIFG opercular part of the inferior frontal gyrus
58	Left Putamen	164	Right OrIFG orbital part of the inferior frontal gyrus
59	Right Thalamus Proper	165	Left OrIFG orbital part of the inferior frontal gyrus
60	Left Thalamus Proper	166	Right PCgG posterior cingulate gyrus
61	Right Ventral DC	167	Left PCgG posterior cingulate gyrus
62	Left Ventral DC	168	Right PCu precuneus
71	Cerebellar Vermal Lobules I-V	169	Left PCu precuneus
72	Cerebellar Vermal Lobules VI-VII	170	Right PHG parahippocampal gyrus
73	Cerebellar Vermal Lobules VIII-X	171	Left PHG parahippocampal gyrus
75	Left Basal Forebrain	172	Right PIns posterior insula
76	Right Basal Forebrain	173	Left PIns posterior insula
100	Right ACgG anterior cingulate gyrus	174	Right PO parietal operculum
101	Left ACgG anterior cingulate gyrus	175	Left PO parietal operculum
102	Right AIns anterior insula	176	Right PoG postcentral gyrus
103	Left AIns anterior insula	177	Left PoG postcentral gyrus
104	Right AOrG anterior orbital gyrus	178	Right POrG posterior orbital gyrus
105	Left AOrG anterior orbital gyrus	179	Left POrG posterior orbital gyrus
106	Right AnG angular gyrus	180	Right PP planum polare
107	Left AnG angular gyrus	181	Left PP planum polare
108	Right Calc calcarine cortex	182	Right PrG precentral gyrus
109	Left Calc calcarine cortex	183	Left PrG precentral gyrus
112	Right CO central operculum	184	Right PT planum temporale
113	Left CO central operculum	185	Left PT planum temporale
114	Right Cun cuneus	186	Right SCA subcallosal area
115	Left Cun cuneus	187	Left SCA subcallosal area
116	Right Ent entorhinal area	190	Right SFG superior frontal gyrus
117	Left Ent entorhinal area	191	Left SFG superior frontal gyrus
118	Right FO frontal operculum	192	Right SMC supplementary motor cortex
119	Left FO frontal operculum	193	Left SMC supplementary motor cortex
120	Right FRP frontal pole	194	Right SMG supramarginal gyrus

121	Left FRP frontal pole	195	Left SMG supramarginal gyrus
122	Right FuG fusiform gyrus	196	Right SOG superior occipital gyrus
123	Left FuG fusiform gyrus	197	Left SOG superior occipital gyrus
124	Right GRe gyrus rectus	198	Right SPL superior parietal lobule
125	Left GRe gyrus rectus	199	Left SPL superior parietal lobule
128	Right IOG inferior occipital gyrus	200	Right STG superior temporal gyrus
129	Left IOG inferior occipital gyrus	201	Left STG superior temporal gyrus
132	Right ITG inferior temporal gyrus	202	Right TMP temporal pole
133	Left ITG inferior temporal gyrus	203	Left TMP temporal pole
134	Right LiG lingual gyrus	204	Right TrIFG triangular part of the inferior frontal gyrus
135	Left LiG lingual gyrus	205	Left TrIFG triangular part of the inferior frontal gyrus
136	Right LOrG lateral orbital gyrus	206	Right TTG transverse temporal gyrus
137	Left LOrG lateral orbital gyrus	207	Left TTG transverse temporal gyrus

Table IX.1 List of ROI names used in the whole-brain segmentation scheme.

2. Chapter VII

In this appendix we include a figure of all non-significant brain regions and the estimated effect size associated wit exposure to heart failure and corresponding confidence intervals. We include an associated table for the effect size and p-values for all brain regions. Lastly, we include a table with the effect size and p-value for all clinical comorbidities included in the PheWAS study.

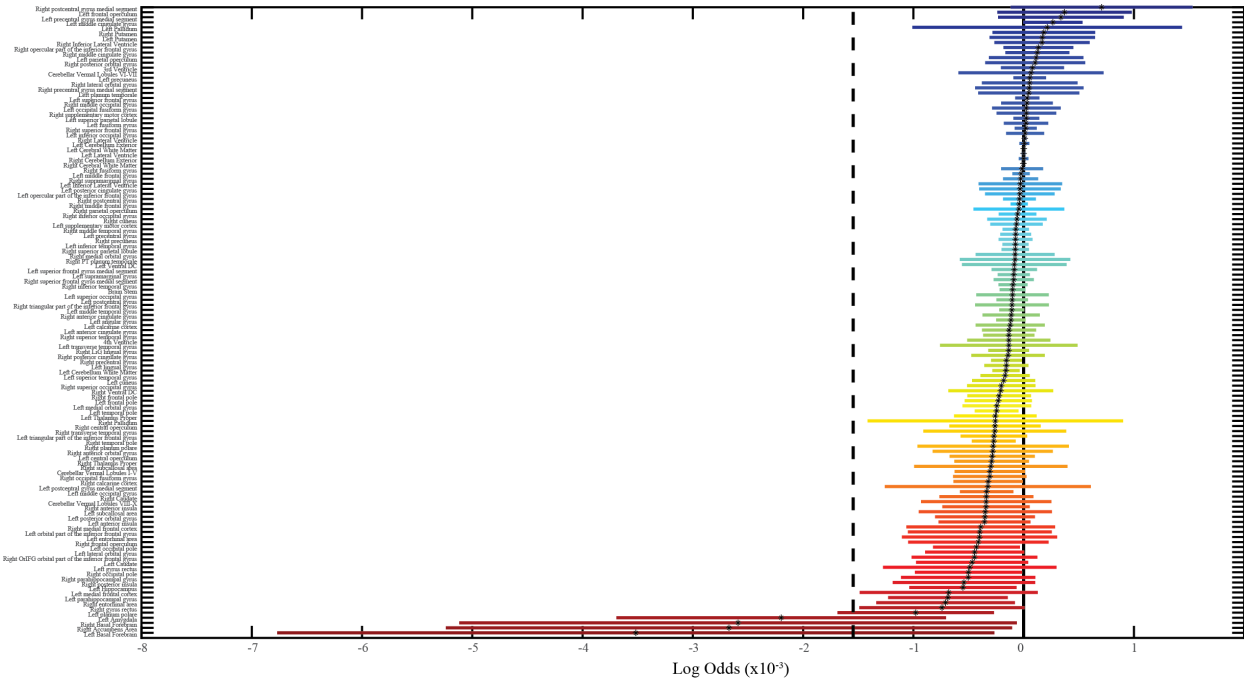

Figure IX.2 Volumetric analysis of non-significant regions. All non-significant regions are shown on the y-axis with the corresponding log odds and 95% confidence intervals. These regions are ranked according to predicted effect size, where negative effect size signifies volume loss associated with heart failure with preserved ejection fraction (HFpEF).

Region of Interest	**Log Odds**	**Corrected p-value**
Left Accumbens Area	-4.42E-03	0.024
Left Basal Forebrain	-3.58E-03	0.204
Right Basal Forebrain	-2.72E-03	0.220
Right Accumbens Area	-2.62E-03	0.273
Right Amygdala	-2.46E-03	0.024
Left Amygdala	-2.25E-03	0.065
Left posterior insula	-1.58E-03	0.012
Left anterior orbital gyrus	-1.37E-03	0.000

Left planum polare	-9.60E-04	0.108
Right entorhinal area	-7.06E-04	0.204
Right gyrus rectus	-6.86E-04	0.324
Left parahippocampal gyrus	-6.80E-04	0.154
Right Hippocampus	-6.41E-04	0.067
Left medial frontal cortex	-6.31E-04	0.387
Right occipital pole	-5.43E-04	0.204
Left Hippocampus	-5.42E-04	0.204
Right posterior insula	-5.35E-04	0.365
Right parahippocampal gyrus	-5.32E-04	0.332
Left gyrus rectus	-4.85E-04	0.472
Left occipital pole	-4.84E-04	0.173
Left Caudate	-4.71E-04	0.323
Left lateral orbital gyrus	-4.63E-04	0.258
Left postcentral gyrus medial segment	-4.60E-04	0.556
Right orbital part of the inferior frontal gyrus	-4.54E-04	0.387
Left entorhinal area	-4.42E-04	0.470
Left subcallosal area	-4.22E-04	0.430
Right medial frontal cortex	-4.08E-04	0.479
Right frontal operculum	-4.06E-04	0.465
Right occipital fusiform gyrus	-3.98E-04	0.181
Right subcallosal area	-3.92E-04	0.516
Left middle occipital gyrus	-3.55E-04	0.067
Right calcarine cortex	-3.54E-04	0.204
Left central operculum	-3.39E-04	0.332
Left anterior insula	-3.27E-04	0.387
Left orbital part of the inferior frontal gyrus	-3.25E-04	0.556
Right anterior insula	-3.10E-04	0.387
Cerebellar Vermal Lobules VIII-X	-3.05E-04	0.556
Right temporal pole	-3.02E-04	0.065
Right central operculum	-3.01E-04	0.422
Right Caudate	-3.01E-04	0.422
Left Thalamus Proper	-3.01E-04	0.387
Left posterior orbital gyrus	-3.00E-04	0.448
Right Thalamus Proper	-2.97E-04	0.332
Left temporal pole	-2.93E-04	0.067
Cerebellar Vermal Lobules I-V	-2.82E-04	0.332
Right anterior orbital gyrus	-2.72E-04	0.556
Left triangular part of the inferior frontal gyrus	-2.53E-04	0.360

Right superior occipital gyrus	-2.38E-04	0.390
Right transverse temporal gyrus	-2.22E-04	0.705
Right angular gyrus	-2.22E-04	0.043
Right planum polare	-2.21E-04	0.705
Left transverse temporal gyrus	-2.15E-04	0.705
Left medial orbital gyrus	-2.10E-04	0.446
Left cuneus	-2.08E-04	0.422
Right frontal pole	-2.07E-04	0.422
Left superior temporal gyrus	-2.05E-04	0.324
Left frontal pole	-2.03E-04	0.448
Right Cerebellum White Matter	-1.94E-04	0.044
Left lingual gyrus	-1.93E-04	0.322
Right Ventral DC	-1.85E-04	0.679
Right precentral gyrus	-1.66E-04	0.186
Left Cerebellum White Matter	-1.59E-04	0.136
Right lingual gyrus	-1.57E-04	0.360
Right Pallidum	-1.56E-04	0.918
Right superior temporal gyrus	-1.53E-04	0.448
Left anterior cingulate gyrus	-1.53E-04	0.472
Right anterior cingulate gyrus	-1.46E-04	0.526
Left superior occipital gyrus	-1.43E-04	0.636
Right triangular part of the inferior frontal gyrus	-1.36E-04	0.678
Right planum temporale	-1.34E-04	0.777
Right superior frontal gyrus medial segment	-1.24E-04	0.446
Left angular gyrus	-1.24E-04	0.323
Left calcarine cortex	-1.14E-04	0.705
Brain Stem	-1.13E-04	0.322
Left superior frontal gyrus medial segment	-1.10E-04	0.551
Left postcentral gyrus	-1.10E-04	0.390
Right posterior cingulate gyrus	-1.09E-04	0.705
Left middle temporal gyrus	-1.08E-04	0.322
Left supramarginal gyrus	-1.04E-04	0.422
Right inferior temporal gyrus	-1.01E-04	0.391
Right superior parietal lobule	-9.40E-05	0.397
Left Ventral DC	-9.33E-05	0.841
4th Ventricle	-9.16E-05	0.779
Left precentral gyrus	-8.74E-05	0.472
Right cuneus	-8.71E-05	0.705
Left opercular part of the inferior frontal gyrus	-8.04E-05	0.779

Right precuneus	-7.81E-05	0.556
Left supplementary motor cortex	-7.65E-05	0.705
Left inferior temporal gyrus	-7.63E-05	0.465
Right middle temporal gyrus	-7.15E-05	0.479
Right inferior occipital gyrus	-6.68E-05	0.679
Right middle frontal gyrus	-4.77E-05	0.477
Right medial orbital gyrus	-4.55E-05	0.920
Left middle frontal gyrus	-3.95E-05	0.556
Right supramarginal gyrus	-3.93E-05	0.779
Right postcentral gyrus	-3.83E-05	0.779
Left posterior cingulate gyrus	-3.82E-05	0.943
Left occipital fusiform gyrus	-2.73E-05	0.946
Right parietal operculum	-2.32E-05	0.964
Left inferior occipital gyrus	-3.78E-06	0.982
Right Cerebral White Matter	-3.29E-06	0.705
Right fusiform gyrus	-1.70E-06	0.994
Left Cerebral White Matter	-7.72E-07	0.950
Right precentral gyrus medial segment	-4.53E-07	0.999
Left Lateral Ventricle	-3.97E-07	0.982
Right Cerebellum Exterior	3.90E-06	0.946
Left superior parietal lobule	8.49E-06	0.957
Right Lateral Ventricle	9.03E-06	0.556
Left Cerebellum Exterior	1.10E-05	0.781
Left Inf Lat Vent	1.28E-05	0.978
Right middle occipital gyrus	1.29E-05	0.964
Right supplementary motor cortex	1.38E-05	0.964
Right superior frontal gyrus	1.62E-05	0.884
Left fusiform gyrus	2.13E-05	0.943
Left planum temporale	2.21E-05	0.964
Left superior frontal gyrus	3.16E-05	0.748
Right lateral orbital gyrus	4.30E-05	0.943
Left precuneus	5.51E-05	0.701
Left parietal operculum	8.12E-05	0.852
3rd Ventricle	8.49E-05	0.740
Cerebellar Vermal Lobules VI-VII	9.06E-05	0.918
Right opercular part of the inferior frontal gyrus	1.03E-04	0.705
Right middle cingulate gyrus	1.04E-04	0.705
Right posterior orbital gyrus	1.50E-04	0.705
Left Putamen	2.01E-04	0.647

Right Putamen	2.15E-04	0.604
Right Inf Lat Vent	2.24E-04	0.556
Left middle cingulate gyrus	2.42E-04	0.332
Left Pallidum	2.49E-04	0.838
Left frontal operculum	2.69E-04	0.628
Left precentral gyrus medial segment	3.31E-04	0.505
Right postcentral gyrus medial segment	8.24E-04	0.291

Table IX.2 Volumetric analysis of regions of interest in the brain. A negative log odds signifies volume atrophy associated with heart failure. P-values are corrected for multiple comparisons using the Benjamini-Hochberg false discovery rate.

PheCode	Effect Size	95% Confidence Interval		Corrected p-value
Septicemia	1.87	0.95	2.79	0.028
Type 1 diabetes	2.10	1.11	3.09	0.013
Type 2 diabetes with neurological manifestations	2.47	1.47	3.48	0.001
Polyneuropathy in diabetes	2.49	1.37	3.61	0.005
Disorders of lipoid metabolism	1.85	0.99	2.72	0.010
Hyperpotassemia	1.97	1.11	2.83	0.003
Hypopotassemia	1.76	0.96	2.57	0.007
Iron deficiency anemias, unspecified or not due to blood loss	1.89	1.02	2.77	0.009
Other anemias	1.71	0.89	2.53	0.017
Anemia of chronic disease	2.34	1.17	3.51	0.035
Alteration of consciousness	1.75	0.94	2.55	0.008
Mitral valve disease	3.27	2.03	4.52	<0.001
Nonrheumatic mitral valve disorders	2.79	1.81	3.77	<0.001
Nonrheumatic aortic valve disorders	2.24	1.27	3.22	0.003
Hypertensive heart and/or renal disease	2.54	1.31	3.76	0.019
Hypertensive heart disease	2.54	1.63	3.45	<0.001
Unstable angina (intermediate coronary syndrome)	2.35	1.37	3.34	0.001
Myocardial infarction	2.41	1.53	3.29	<0.001
Angina pectoris	2.57	1.49	3.66	0.001
Coronary atherosclerosis	2.07	1.21	2.93	0.001
Other chronic ischemic heart disease, unspecified	3.14	2.21	4.07	<0.001
Other forms of chronic heart disease	1.89	0.95	2.82	0.029
Cardiomegaly	2.44	1.58	3.29	<0.001
Nonspecific chest pain	2.30	1.28	3.33	0.004
Precordial pain	1.79	0.89	2.70	0.042
Cardiac conduction disorders	2.03	1.00	3.05	0.044
Atrial fibrillation	2.32	1.44	3.21	<0.001
Tachycardia NOS	1.65	0.82	2.49	0.041

Atherosclerosis of aorta	1.79	0.88	2.70	0.047
Hypotension NOS	1.99	1.15	2.84	0.002
Acute bronchitis and bronchiolitis	1.73	0.88	2.59	0.027
Obstructive chronic bronchitis	2.34	1.19	3.50	0.028
Pleurisy; pleural effusion	1.93	1.09	2.78	0.003
Respiratory failure	2.33	1.30	3.37	0.004
Shortness of breath	2.08	1.22	2.95	0.001
Other dyspnea	1.64	0.84	2.44	0.025
Diarrhea	1.84	1.02	2.66	0.004
Acute renal failure	1.62	0.81	2.43	0.034
Disorder of skin and subcutaneous tissue NOS	1.73	0.86	2.61	0.042
Swelling of limb	2.41	1.53	3.29	<0.001
Edema	2.05	1.22	2.88	0.001
Sepsis	2.09	1.17	3.00	0.003

Table IX.3 PheWAS results for significant clinical associations with HFpEF in the setting of dementia in patients with high resolution imaging. P-values are corrected for multiple comparisons using Bonferroni corrections.

References

1. Prince, J.L., Links, J.M.: Medical imaging signals and systems. Pearson Prentice Hall Upper Saddle River, NJ (2006)
2. Diedrichsen, J., Maderwald, S., Küper, M., Thürling, M., Rabe, K., Gizewski, E., Ladd, M.E., Timmann, D.: Imaging the deep cerebellar nuclei: a probabilistic atlas and normalization procedure. Neuroimage 54, 1786-1794 (2011)
3. Andrushko, V.A., Verano, J.W.: Prehistoric trepanation in the Cuzco region of Peru: a view into an ancient Andean practice. American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists 137, 4-13 (2008)
4. Wilkins, R.H., Classic-XVII, N.: The Edwin Smith surgical papyrus. Journal of Neurosurgery 5, 240-244 (1964)
5. Finger, S., Boller, F., Tyler, K.L.: History of neurology. Elsevier (2009)
6. Gross, C.G., Marshall, J.C.: Brain, Vision, Memory: Tales in the History of Neuroscience. Nature 394, 143-143 (1998)
7. Hughes, J.: Thomas Willis 1621–1675: his life and work. Singapore Med J 50, 1041 (2009)
8. Macmillan, M.: Restoring phineas gage: A 150th retrospective. Journal of the History of the Neurosciences 9, 46-66 (2000)
9. Moritz, D.J., Kasl, S.V., Berkman, L.F.: Cognitive functioning and the incidence of limitations in activities of daily living in an elderly community sample. American Journal of Epidemiology 141, 41-49 (1995)
10. Herzog, A.R., Wallace, R.B.: Measures of cognitive functioning in the AHEAD Study. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 52, 37-48 (1997)
11. Röntgen, W.C.: Eine neue art von strahlen. Stahel (1896)
12. Griscom, N.T.: History of pediatric radiology in the United States and Canada: images and trends. Radiographics 15, 1399-1422 (1995)
13. Labadie, R.F.: Image-guided surgery: fundamentals and clinical applications in otolaryngology. Plural Publishing (2016)
14. Radue, E.-W., Weigel, M., Wiest, R., Urbach, H.: Introduction to magnetic resonance imaging for neurologists. Continuum: Lifelong Learning in Neurology 22, 1379-1398 (2016)
15. Ichikawa, L.E., Barlow, W.E., Anderson, M.L., Taplin, S.H., Geller, B.M., Brenner, R.J.: Time trends in radiologists' interpretive performance at screening mammography from the community-based Breast Cancer Surveillance Consortium, 1996–2004. Radiology 256, 74-82 (2010)
16. Smith-Bindman, R., Miglioretti, D.L., Larson, E.B.: Rising use of diagnostic medical imaging in a large integrated health system. Health affairs 27, 1491-1502 (2008)
17. Greenspan, H., van Ginneken, B., Summers, R.M.: Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. IEEE Transactions on Medical Imaging 35, 1153-1159 (2016)
18. Van Horn, J.D., Toga, A.W.: Human neuroimaging as a "Big Data" science. Brain imaging and behavior 8, 323-331 (2014)
19. Zhang, S., Metaxas, D.: Large-Scale medical image analytics: Recent methodologies, applications and Future directions. Elsevier (2016)
20. Wells III, W.M.: Medical Image Analysis–past, present, and future. Elsevier (2016)
21. Duncan, J.S., Ayache, N.: Medical image analysis: Progress over two decades and the challenges ahead. IEEE transactions on pattern analysis and machine intelligence 22, 85-106 (2000)
22. Rueckert, D., Glocker, B., Kainz, B.: Learning clinically useful information from images: past, present and future. Elsevier (2016)
23. Frangi, A.F., Taylor, Z.A., Gooya, A.: Precision Imaging: more descriptive, predictive and integrative imaging. Elsevier (2016)

24. de Bruijne, M.: Machine learning approaches in medical image analysis: From detection to diagnosis. Elsevier (2016)
25. Criminisi, A.: Machine learning for medical images analysis. Elsevier (2016)
26. Bluemke, D.A.: Editor's Note: Publication of AI Research in Radiology. Radiological Society of North America (2018)
27. Park, S.H., Han, K.: Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology 286, 800-809 (2018)
28. Caserta, M.T., Bannon, Y., Fernandez, F., Giunta, B., Schoenberg, M.R., Tan, J.: Normal brain aging: clinical, immunological, neuropsychological, and neuroimaging features. International review of neurobiology 84, 1-19 (2009)
29. Lockhart, S., DeCarli, C., Fama, R.: Neuroimaging of the aging brain: Introduction to the special issue of neuropsychology review. Springer (2014)
30. Kochunov, P., Thompson, P.M., Coyle, T.R., Lancaster, J.L., Kochunov, V., Royall, D., Mangin, J.F., Riviere, D., Fox, P.T.: Relationship among neuroimaging indices of cerebral health during normal aging. Human brain mapping 29, 36-45 (2008)
31. Bonner, S., McGough, A.S., Kureshi, I., Brennan, J., Theodoropoulos, G., Moss, L., Corsar, D., Antoniou, G.: Data quality assessment and anomaly detection via map/reduce and linked data: a case study in the medical domain. In: 2015 IEEE International Conference on Big Data (Big Data), pp. 737-746. IEEE, (Year)
32. Elster, A.D., Burdette, J.H.: Questions and Answers in Magnetic Resonance Imaging, 2e. Mosby, Inc 1, 1-18 (2001)
33. Beaulieu, C.: The basis of anisotropic water diffusion in the nervous system–a technical review. NMR in Biomedicine 15, 435-455 (2002)
34. O'Donnell, L.J., Westin, C.-F.: An introduction to diffusion tensor image analysis. Neurosurgery Clinics 22, 185-196 (2011)
35. Alexander, A., Lee, J., Lazar, M., Field, A.: Diffusion tensor imaging of the brain. Neurotherapeutics: the journal of the American Society for Experimental NeuroTherapeutics. 2007; 4 (3): 316–29.
36. Vogenberg, F.R., Barash, C.I., Pursel, M.: Personalized medicine: part 1: evolution and development into theranostics. Pharmacy and Therapeutics 35, 560 (2010)
37. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis 42, 60-88 (2017)
38. Shen, D., Wu, G., Suk, H.-I.: Deep learning in medical image analysis. Annual review of biomedical engineering 19, 221-248 (2017)
39. Wernick, M.N., Yang, Y., Brankov, J.G., Yourganov, G., Strother, S.C.: Machine learning in medical imaging. IEEE signal processing magazine 27, 25-38 (2010)
40. Aboud, K.S., Huo, Y., Kang, H., Ealey, A., Resnick, S.M., Landman, B.A., Cutting, L.E.: Structural covariance across the lifespan: Brain development and aging through the lens of inter-network relationships. Human brain mapping 40, 125-136 (2019)
41. Huo, Y., Aboud, K., Kang, H., Cutting, L.E., Landman, B.A.: Mapping Lifetime Brain Volumetry with Covariate-Adjusted Restricted Cubic Spline Regression from Cross-Sectional Multi-site MRI. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 81-88. Springer, (Year)
42. Schmidhuber, J.: Deep learning in neural networks: An overview. Neural networks 61, 85-117 (2015)
43. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521, 436-444 (2015)
44. Gray, A., Gottbrath, C., Olson, R., Prasanna, S.: Deploying deep neural networks with nvidia tensorrt. Apr (2017)
45. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278-2324 (1998)

46. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp. 1097-1105. (Year)
47. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)
48. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, pp. 234-241. Springer, (Year)
49. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. (Year)
50. Nielsen, M.A.: Neural networks and deep learning. Determination press USA (2015)
51. Vellido, A., Martín-Guerrero, J.D., Lisboa, P.J.: Making machine learning models interpretable. In: ESANN, pp. 163-172. Citeseer, (Year)
52. Lipton, Z.C.: The mythos of model interpretability. arXiv preprint arXiv:1606.03490 (2016)
53. Montavon, G., Samek, W., Müller, K.-R.: Methods for interpreting and understanding deep neural networks. Digital Signal Processing 73, 1-15 (2018)
54. Lou, Y., Caruana, R., Gehrke, J.: Intelligible models for classification and regression. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 150-158. ACM, (Year)
55. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618-626. (Year)
56. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
57. Rasmussen, P.M., Madsen, K.H., Lund, T.E., Hansen, L.K.: Visualization of nonlinear kernel models in neuroimaging by sensitivity maps. NeuroImage 55, 1120-1131 (2011)
58. Ortman, J.M., Velkoff, V.A., Hogan, H.: An aging nation: the older population in the United States. United States Census Bureau, Economics and Statistics Administration, US … (2014)
59. Dall, T.M., Gallo, P.D., Chakrabarti, R., West, T., Semilla, A.P., Storm, M.V.: An aging population and growing disease burden will require alarge and specialized health care workforce by 2025. Health affairs 32, 2013-2020 (2013)
60. Ferrucci, L.: The Baltimore Longitudinal Study of Aging (BLSA): a 50-year-long journey and plans for the future. Oxford University Press (2008)
61. Resnick, S.M., Pham, D.L., Kraut, M.A., Zonderman, A.B., Davatzikos, C.: Longitudinal magnetic resonance imaging studies of older adults: a shrinking brain. Journal of Neuroscience 23, 3295-3301 (2003)
62. Jaffer, F.A., O'Donnell, C.J., Larson, M.G., Chan, S.K., Kissinger, K.V., Kupka, M.J., Salton, C., Botnar, R.M., Levy, D., Manning, W.J.: Age and sex distribution of subclinical aortic atherosclerosis: a magnetic resonance imaging examination of the Framingham Heart Study. Arteriosclerosis, thrombosis, and vascular biology 22, 849-854 (2002)
63. Salton, C.J., Chuang, M.L., O'Donnell, C.J., Kupka, M.J., Larson, M.G., Kissinger, K.V., Edelman, R.R., Levy, D., Manning, W.J.: Gender differences and normal left ventricular anatomy in an adult population free of hypertension: A cardiovascular magnetic resonance study of the Framingham Heart Study Offspring cohort. Journal of the American College of Cardiology 39, 1055-1060 (2002)
64. Scharf, E.L., Graff-Radford, J., Przybelski, S.A., Lesnick, T.G., Mielke, M.M., Knopman, D.S., Preboske, G.M., Schwarz, C.G., Senjem, M.L., Gunter, J.L.: Cardiometabolic health and longitudinal progression of white matter hyperintensity: the Mayo Clinic Study of Aging. Stroke 50, 3037-3044 (2019)
65. Pankratz, V.S., Roberts, R.O., Mielke, M.M., Knopman, D.S., Jack, C.R., Geda, Y.E., Rocca, W.A., Petersen, R.C.: Predicting the risk of mild cognitive impairment in the Mayo Clinic Study of Aging. Neurology 84, 1433-1442 (2015)
66. Skovronsky, D.M., Lee, V.M.-Y., Trojanowski, J.Q.: Neurodegenerative diseases: new concepts of pathogenesis and their therapeutic implications. Annu. Rev. Pathol. Mech. Dis. 1, 151-170 (2006)

67. Mueller, S., Schuff, N., Weiner, M.: Evaluation of treatment effects in Alzheimer's and other neurodegenerative diseases by MRI and MRS. NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In vivo 19, 655-668 (2006)
68. Cole, J.H., Marioni, R.E., Harris, S.E., Deary, I.J.: Brain age and other bodily 'ages': implications for neuropsychiatry. Molecular psychiatry 1 (2018)
69. Lemaitre, H., Goldman, A.L., Sambataro, F., Verchinski, B.A., Meyer-Lindenberg, A., Weinberger, D.R., Mattay, V.S.: Normal age-related brain morphometric changes: nonuniformity across cortical thickness, surface area and gray matter volume? Neurobiology of aging 33, 617. e611-617. e619 (2012)
70. Kaufmann, T., van der Meer, D., Doan, N.T., Schwarz, E., Lund, M.J., Agartz, I., Alnæs, D., Barch, D.M., Baur-Streubel, R., Bertolino, A.: Genetics of brain age suggest an overlap with common brain disorders. bioRxiv 303164 (2018)
71. Lewis, J.D., Evans, A.C., Tohka, J.: T1 white/gray contrast as a predictor of chronological age, and an index of cognitive performance. bioRxiv 171892 (2017)
72. Franke, K., Ziegler, G., Klöppel, S., Gaser, C., Initiative, A.s.D.N.: Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters. Neuroimage 50, 883-892 (2010)
73. Franke, K., Gaser, C.: Longitudinal Changes in Individual BrainAGE in Healthy Aging, Mild Cognitive Impairment, and Alzheimer's Disease 1Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni. loni. ucla. edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: adni. loni. ucla. edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List. pdf. GeroPsych (2012)
74. Löwe, L.C., Gaser, C., Franke, K., Initiative, A.s.D.N.: The effect of the APOE genotype on individual BrainAGE in normal aging, mild cognitive impairment, and Alzheimer's disease. PLoS One 11, e0157514 (2016)
75. Li, Y., Liu, Y., Wang, P., Wang, J., Xu, S., Qiu, M.: Dependency criterion based brain pathological age estimation of Alzheimer's disease patients with MR scans. Biomedical engineering online 16, 50 (2017)
76. Varikuti, D.P., Genon, S., Sotiras, A., Schwender, H., Hoffstaedter, F., Patil, K.R., Jockwitz, C., Caspers, S., Moebus, S., Amunts, K.: Evaluation of non-negative matrix factorization of grey matter in age prediction. NeuroImage 173, 394-410 (2018)
77. Koutsouleris, N., Davatzikos, C., Borgwardt, S., Gaser, C., Bottlender, R., Frodl, T., Falkai, P., Riecher-Rössler, A., Möller, H.-J., Reiser, M.: Accelerated brain aging in schizophrenia and beyond: a neuroanatomical marker of psychiatric disorders. Schizophrenia bulletin 40, 1140-1153 (2013)
78. Schnack, H.G., Van Haren, N.E., Nieuwenhuis, M., Hulshoff Pol, H.E., Cahn, W., Kahn, R.S.: Accelerated brain aging in schizophrenia: a longitudinal pattern recognition study. American Journal of Psychiatry 173, 607-616 (2016)
79. Nenadic, I., Dietzek, M., Langbein, K., Sauer, H., Gaser, C.: BrainAGE score indicates accelerated brain aging in schizophrenia, but not bipolar disorder. Psychiatry Res Neuroimaging 266, 86-89 (2017)
80. Cole, J.H., Leech, R., Sharp, D.J., Initiative, A.s.D.N.: Prediction of brain age suggests accelerated atrophy after traumatic brain injury. Ann Neurol 77, 571-581 (2015)
81. Wang, B., Pham, T.D.: MRI-based age prediction using hidden Markov models. J Neurosci Methods 199, 140-145 (2011)
82. Sabuncu, M.R., Van Leemput, K., Initiative, A.s.D.N.: The relevance voxel machine (RVoxM): a self-tuning Bayesian model for informative image-based prediction. IEEE Trans Med Imaging 31, 2290-2306 (2012)
83. Cole, J.H., Poudel, R.P., Tsagkrasoulis, D., Caan, M.W., Steves, C., Spector, T.D., Montana, G.: Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. NeuroImage 163, 115-124 (2017)

84. Wu, M., Rosano, C., Lopez-Garcia, P., Carter, C.S., Aizenstein, H.J.: Optimum template selection for atlas-based segmentation. NeuroImage 34, 1612-1618 (2007)
85. Guimond, A., Meunier, J., Thirion, J.-P.: Average brain models: A convergence study. Computer vision and image understanding 77, 192-210 (2000)
86. De Morsier, F., Tuia, D., Borgeaud, M., Gass, V., Thiran, J.-P.: Semi-supervised novelty detection using svm entire solution path. IEEE Transactions on Geoscience and Remote Sensing 51, 1939-1950 (2013)
87. Rohlfing, T., Russakoff, D.B., Brandt, R., Menzel, R., Maurer, C.J.: Performance-based multi-classifier decision fusion for atlas-based segmentation of biomedical images. In: 2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821), pp. 404-407. IEEE, (Year)
88. Heckemann, R.A., Hajnal, J.V., Aljabar, P., Rueckert, D., Hammers, A.: Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage 33, 115-126 (2006)
89. Asman, A.J., Huo, Y., Plassard, A.J., Landman, B.A.: Multi-atlas learner fusion: An efficient segmentation approach for large-scale data. Medical image analysis 26, 82-91 (2015)
90. Küper, M., Thürling, M., Maderwald, S., Ladd, M.E., Timmann, D.: Structural and functional magnetic resonance imaging of the human cerebellar nuclei. The Cerebellum 11, 314-324 (2012)
91. Deuschl, G., Wenzelburger, R., Löffler, K., Raethjen, J., Stolze, H.: Essential tremor and cerebellar dysfunction clinical and kinematic analysis of intention tremor. Brain 123, 1568-1580 (2000)
92. Paris-Robidas, S., Brochu, E., Sintes, M., Emond, V., Bousquet, M., Vandal, M., Pilote, M., Tremblay, C., Di Paolo, T., Rajput, A.H.: Defective dentate nucleus GABA receptors in essential tremor. Brain 135, 105-116 (2011)
93. Pinto, A.D., Lang, A.E., Chen, R.: The cerebellothalamocortical pathway in essential tremor. Neurology 60, 1985-1987 (2003)
94. Elble, R.J., Deuschl, G.: An update on essential tremor. Current neurology and neuroscience reports 9, 273-277 (2009)
95. Louis, E.D., Vonsattel, J.P.G., Honig, L.S., Lawton, A., Moskowitz, C., Ford, B., Frucht, S.: Essential tremor associated with pathologic changes in the cerebellum. Archives of neurology 63, 1189-1193 (2006)
96. Kwon, H.G., Hong, J.H., Hong, C.P., Lee, D.H., Ahn, S.H., Jang, S.H.: Dentatorubrothalamic tract in human brain: diffusion tensor tractography study. Neuroradiology 53, 787-791 (2011)
97. van Baarsen, K., Kleinnijenhuis, M., Konert, T., van Walsum, A.-M.v.C., Grotenhuis, A.: Tractography demonstrates dentate-rubro-thalamic tract disruption in an adult with cerebellar mutism. The Cerebellum 12, 617-622 (2013)
98. Petersen, K.J., Reid, J.A., Chakravorti, S., Juttukonda, M.R., Franco, G., Trujillo, P., Stark, A.J., Dawant, B.M., Donahue, M.J., Claassen, D.O.: Structural and functional connectivity of the nondecussating dentato-rubro-thalamic tract. NeuroImage 176, 364-371 (2018)
99. Dimitrova, A., Zeljko, D., Schwarze, F., Maschke, M., Gerwig, M., Frings, M., Beck, A., Aurich, V., Forsting, M., Timmann, D.: Probabilistic 3D MRI atlas of the human cerebellar dentate/interposed nuclei. Neuroimage 30, 12-25 (2006)
100. Diedrichsen, J.: A spatially unbiased atlas template of the human cerebellum. Neuroimage 33, 127-138 (2006)
101. Mirzaalian, H., Ning, L., Savadjiev, P., Pasternak, O., Bouix, S., Michailovich, O., Grant, G., Marx, C., Morey, R.A., Flashman, L.: Inter-site and inter-scanner diffusion MRI data harmonization. NeuroImage 135, 311-323 (2016)
102. Nath, V., al, e.: Inter-Scanner Harmonization of High Angular Resolution DW-MRI using Null Space Deep Learning. CD-MRI (2018)
103. Plassard, A.J., Harrigan, R.L., Newton, A.T., Rane, S., Pallavaram, S., D'Haese, P.F., Dawant, B.M., Claassen, D.O., Landman, B.A.: On the fallacy of quantitative segmentation for T1-weighted MRI. In: Medical Imaging 2016: Image Processing, pp. 978416. International Society for Optics and Photonics, (Year)

104. Mirzaalian, H., Ning, L., Savadjiev, P., Pasternak, O., Bouix, S., Michailovich, O., Karmacharya, S., Grant, G., Marx, C.E., Morey, R.A.: Multi-site harmonization of diffusion MRI data in a registration framework. Brain imaging and behavior 12, 284-295 (2018)
105. Association, A.s.: 2017 Alzheimer's disease facts and figures. Alzheimer's & Dementia 13, 325-373 (2017)
106. Gorelick, P.B., Scuteri, A., Black, S.E., DeCarli, C., Greenberg, S.M., Iadecola, C., Launer, L.J., Laurent, S., Lopez, O.L., Nyenhuis, D.: Vascular contributions to cognitive impairment and dementia. Stroke 42, 2672-2713 (2011)
107. Snyder, H.M., Corriveau, R.A., Craft, S., Faber, J.E., Greenberg, S.M., Knopman, D., Lamb, B.T., Montine, T.J., Nedergaard, M., Schaffer, C.B.: Vascular contributions to cognitive impairment and dementia including Alzheimer's disease. Alzheimer's & Dementia 11, 710-717 (2015)
108. Qiu, C., Winblad, B., Marengoni, A., Klarin, I., Fastbom, J., Fratiglioni, L.: Heart failure and risk of dementia and Alzheimer disease: a population-based cohort study. Archives of Internal Medicine 166, 1003-1008 (2006)
109. Heckman, G.A., Patterson, C.J., Demers, C., Onge, J.S., Turpie, I.D., McKelvie, R.S.: Heart failure and cognitive impairment: challenges and opportunities. Clinical interventions in aging 2, 209 (2007)
110. Gorelick, P.B., Scuteri, A., Black, S.E., DeCarli, C., Greenberg, S.M., Iadecola, C., Launer, L.J., Laurent, S., Lopez, O.L., Nyenhuis, D.: Vascular contributions to cognitive impairment and dementia: a statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke 42, 2672-2713 (2011)
111. Zuccalà, G., Marzetti, E., Cesari, M., Monaco, M.R.L., Antonica, L., Cocchi, A., Carbonin, P., Bernabei, R.: Correlates of cognitive impairment among patients with heart failure: results of a multicenter survey. The American journal of medicine 118, 496-502 (2005)
112. Cermakova, P., Lund, L.H., Fereshtehnejad, S.M., Johnell, K., Winblad, B., Dahlström, U., Eriksdotter, M., Religa, D.: Heart failure and dementia: survival in relation to types of heart failure and different dementia disorders. European journal of heart failure 17, 612-619 (2015)
113. Vogels, R.L., Scheltens, P., Schroeder-Tanka, J.M., Weinstein, H.C.: Cognitive impairment in heart failure: a systematic review of the literature. European journal of heart failure 9, 440-449 (2007)
114. Meguro, T., Meguro, Y., Kunieda, T.: Atrophy of the parahippocampal gyrus is prominent in heart failure patients without dementia. ESC heart failure 4, 632-640 (2017)
115. Festen, S., de Rooij, S.E.: Heart failure and brain failure: two of a kind? European journal of heart failure 17, 539-540 (2015)
116. Shameer, K., Denny, J.C., Ding, K., Jouni, H., Crosslin, D.R., De Andrade, M., Chute, C.G., Peissig, P., Pacheco, J.A., Li, R.: A genome-and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. Human genetics 133, 95-109 (2014)
117. Leader, J.B., Pendergrass, S.A., Verma, A., Carey, D.J., Hartzel, D.N., Ritchie, M.D., Kirchner, H.L.: Contrasting Association Results between Existing PheWAS Phenotype Definition Methods and Five Validated Electronic Phenotypes. In: AMIA Annual Symposium Proceedings, pp. 824. American Medical Informatics Association, (Year)
118. Denny, J.C., Bastarache, L., Ritchie, M.D., Carroll, R.J., Zink, R., Mosley, J.D., Field, J.R., Pulley, J.M., Ramirez, A.H., Bowton, E.: Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nature biotechnology 31, 1102-1111 (2013)
119. Hancock-Cerutti, W., Rader, D.J.: Opposing effects of ABCG5/8 function on myocardial infarction and gallstone disease. Journal of the American College of Cardiology (2014)
120. Ritchie, M.D., Denny, J.C., Zuvich, R.L., Crawford, D.C., Schildcrout, J.S., Bastarache, L., Ramirez, A.H., Mosely, J.D., Pulley, J.M., Basford, M.A.: Genome-and phenome-wide analysis of cardiac conduction identifies markers of arrhythmia risk. Circulation CIRCULATIONAHA. 112.000604 (2013)
121. Beason-Held, L.L., Thambisetty, M., Deib, G., Sojkova, J., Landman, B.A., Zonderman, A.B., Ferrucci, L., Kraut, M.A., Resnick, S.M.: Baseline cardiovascular risk predicts subsequent changes in resting brain function. Stroke 43, 1542-1547 (2012)

122. Romero, J.R., Beiser, A., Himali, J.J., Shoamanesh, A., DeCarli, C., Seshadri, S.: Cerebral microbleeds and risk of incident dementia: the Framingham Heart Study. Neurobiology of Aging 54, 94-99 (2017)
123. Romero, J.R., Preis, S.R., Beiser, A., Himali, J.J., Shoamanesh, A., Wolf, P.A., Kase, C.S., Vasan, R.S., DeCarli, C., Seshadri, S.: Cerebral Microbleeds as Predictors of Mortality. Stroke 48, 781-783 (2017)
124. Moore, E.E., Liu, D., Pechman, K.R., Terry, J.G., Nair, S., Cambronero, F.E., Bell, S.P., Gifford, K.A., Anderson, A.W., Hohman, T.J.: Increased Left Ventricular Mass Index Is Associated With Compromised White Matter Microstructure Among Older Adults. Journal of the American Heart Association 7, e009041 (2018)
125. Bermudez, C., Plassard, A.J., Chaganti, S., Huo, Y., Aboud, K.E., Cutting, L.E., Resnick, S., Landman, B.A.: Anatomic Context Improves Deep Learning on Brain Age Estimation Task. Neuroimage. Under review. (2018)
126. Li, Y., Zhang, H., Bermudez, C., Chen, Y., Landman, B.A., Vorobeychik, Y.: Anatomical Context Protects Deep Learning from Adversarial Perturbations in Medical Imaging. Neurocomputing (2019)
127. Noguera, C.B., Bao, S., Petersen, K.J., Lopez, A.M., Reid, J., Plassard, A.J., Zald, D.H., Claassen, D.O., Dawant, B.M., Landman, B.A.: Using deep learning for a diffusion-based segmentation of the dentate nucleus and its benefits over atlas-based methods. Journal of Medical Imaging 6, 044007 (2019)
128. Bermudez, C., Blaber, J., Remedios, S.W., Reynolds, J.E., Lebel, C., McHugo, M., Heckers, S., Huo, Y., Landman, B.A.: Generalizing deep whole brain segmentation for pediatric and post-contrast MRI with augmented transfer learning. In: Medical Imaging 2020: Image Processing, pp. 113130L. International Society for Optics and Photonics, (Year)
129. Cole, J.H., Franke, K.: Predicting Age Using Neuroimaging: Innovative Brain Ageing Biomarkers. Trends in Neurosciences (2017)
130. Cole, J.H., Ritchie, S.J., Bastin, M.E., Hernández, M.V., Maniega, S.M., Royle, N., Corley, J., Pattie, A., Harris, S.E., Zhang, Q.: Brain age predicts mortality. Molecular Psychiatry (2017)
131. Su, L., Wang, L., Hu, D.: Predicting the age of healthy adults from structural MRI by sparse representation. In: International Conference on Intelligent Science and Intelligent Data Engineering, pp. 271-279. Springer, (Year)
132. Brown, T.T., Kuperman, J.M., Chung, Y., Erhart, M., McCabe, C., Hagler, D.J., Venkatraman, V.K., Akshoomoff, N., Amaral, D.G., Bloss, C.S., Casey, B.J., Chang, L., Ernst, T.M., Frazier, J.A., Gruen, J.R., Kaufmann, W.E., Kenet, T., Kennedy, D.N., Murray, S.S., Sowell, E.R., Jernigan, T.L., Dale, A.M.: Neuroanatomical assessment of biological maturity. Curr Biol 22, 1693-1698 (2012)
133. Cherubini, A., Caligiuri, M.E., Peran, P., Sabatini, U., Cosentino, C., Amato, F.: Importance of Multimodal MRI in Characterizing Brain Tissue and its Potential Application for Individual Age Prediction. IEEE J Biomed Health Inform (2016)
134. Avants, B.B., Grossman, M., Gee, J.C.: Symmetric diffeomorphic image registration: Evaluating automated labeling of elderly and neurodegenerative cortex and frontal lobe. In: International Workshop on Biomedical Image Registration, pp. 50-57. Springer, (Year)
135. Asman, A.J., Dagley, A.S., Landman, B.A.: Statistical label fusion with hierarchical performance models. In: SPIE Medical Imaging, pp. 90341E-90341E-90348. International Society for Optics and Photonics, (Year)
136. Klein, A., Dal Canton, T., Ghosh, S.S., Landman, B., Lee, J., Worth, A.: Open labels: online feedback for a public resource of manually labeled brain images. In: 16th Annual Meeting for the Organization of Human Brain Mapping. (Year)
137. Smith, S.M., Zhang, Y., Jenkinson, M., Chen, J., Matthews, P., Federico, A., De Stefano, N.: Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. Neuroimage 17, 479-489 (2002)
138. Harrigan, R.L., Panda, S., Asman, A.J., Nelson, K.M., Chaganti, S., DeLisi, M.P., Yvernault, B.C., Smith, S.A., Galloway, R.L., Mawn, L.A.: Robust optic nerve segmentation on clinically acquired computed tomography. Journal of Medical Imaging 1, 034006-034006 (2014)
139. Rac, K.: keras-vis. GitHub. (2017)

140. Chaganti, S., Robinson, J.R., Bermudez, C., Lasko, T., Mawn, L.A., Landman, B.A.: EMR-Radiological Phenotypes in Diseases of the Optic Nerve and Their Association with Visual Function. Lecture Notes in Computer Science 10553, 373-381 (2017)
141. Moosavi-Dezfooli, S.-M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2574-2582. (Year)
142. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
143. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
144. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 427-436. (Year)
145. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
146. Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M.K.: Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 1528-1540. ACM, (Year)
147. Huo, Y., Xu, Z., Xiong, Y., Aboud, K., Parvathaneni, P., Bao, S., Bermudez, C., Resnick, S.M., Cutting, L.E., Landman, B.A.: 3D whole brain segmentation using spatially localized atlas network tiles. NeuroImage (2019)
148. Huo, Y., Terry, J.G., Wang, J., Nath, V., Bermudez, C., Bao, S., Parvathaneni, P., Carr, J.J., Landman, B.A.: Coronary Calcium Detection using 3D Attention Identical Dual Deep Network Based on Weakly Supervised Learning. arXiv preprint arXiv:1811.04289 (2018)
149. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115 (2017)
150. Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Jama 316, 2402-2410 (2016)
151. Bejnordi, B.E., Veta, M., van Diest, P.J., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama 318, 2199-2210 (2017)
152. Food & Drug Administration, U.S.A.: Approval Letter: Quantitative Insights. (2017)
153. Food & Drug Administration, U.S.A.: Approval Letter: Butterfly Network. (2017)
154. Food & Drug Administration, U.S.A.: Approval Letter: Arterys Software 2.0. (2016)
155. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135-1144. ACM, (Year)
156. Vorobeychik, Y., Kantarcioglu, M.: Adversarial machine learning. Synthesis Lectures on Artificial Intelligence and Machine Learning 12, 1-169 (2018)
157. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533 (2016)
158. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 39-57. IEEE, (Year)
159. Finlayson, S.G., Kohane, I.S., Beam, A.L.: Adversarial attacks against medical deep learning systems. arXiv preprint arXiv:1804.05296 (2018)
160. Kong, B., Wang, X., Li, Z., Song, Q., Zhang, S.: Cancer metastasis detection via spatially structured deep network. In: International Conference on Information Processing in Medical Imaging, pp. 236-248. Springer, (Year)

161. Yan, Z., Zhan, Y., Peng, Z., Liao, S., Shinagawa, Y., Zhang, S., Metaxas, D.N., Zhou, X.S.: Multi-instance deep learning: Discover discriminative local anatomies for bodypart recognition. IEEE transactions on medical imaging 35, 1332-1343 (2016)
162. Asman, A.J., Landman, B.A.: Hierarchical performance estimation in the statistical label fusion framework. Medical image analysis 18, 1070-1081 (2014)
163. McEvoy, F.J., Svalastoga, E.: Security of patient and study data associated with DICOM images when transferred using compact disc media. Journal of digital imaging 22, 65-70 (2009)
164. Li, B., Vorobeychik, Y.: Evasion-robust classification on binary domains. ACM Transactions on Knowledge Discovery from Data (TKDD) 12, 50 (2018)
165. Khan, S.S., Singer, B.D., Vaughan, D.E.: Molecular and physiological manifestations and measurement of aging in humans. Aging Cell 16, 624-633 (2017)
166. Sweet, J.A., Walter, B.L., Gunalan, K., Chaturvedi, A., McIntyre, C.C., Miller, J.P.: Fiber tractography of the axonal pathways linking the basal ganglia and cerebellum in Parkinson disease: implications for targeting in deep brain stimulation. Journal of neurosurgery 120, 988-996 (2014)
167. Ridout, K.K., Ridout, S.J., Price, L.H., Sen, S., Tyrka, A.R.: Depression and telomere length: A meta-analysis. J Affect Disord 191, 237-247 (2016)
168. Levine, A.J., Quach, A., Moore, D.J., Achim, C.L., Soontornniyomkij, V., Masliah, E., Singer, E.J., Gelman, B., Nemanim, N., Horvath, S.: Accelerated epigenetic aging in brain is associated with pre-mortem HIV-associated neurocognitive disorders. J Neurovirol 22, 366-375 (2016)
169. Darrow, S.M., Verhoeven, J.E., Revesz, D., Lindqvist, D., Penninx, B.W., Delucchi, K.L., Wolkowitz, O.M., Mathews, C.A.: The Association Between Psychiatric Disorders and Telomere Length: A Meta-Analysis Involving 14,827 Persons. Psychosom Med 78, 776-787 (2016)
170. Brown, P.J., Wall, M.M., Chen, C., Levine, M.E., Yaffe, K., Roose, S.P., Rutherford, B.R.: Biological Age, Not Chronological Age, Is Associated with Late-Life Depression. J Gerontol A Biol Sci Med Sci 73, 1370-1376 (2018)
171. Isaev, N.K., Genrikhs, E.E., Oborina, M.V., Stelmashook, E.V.: Accelerated aging and aging process in the brain. Rev Neurosci 29, 233-240 (2018)
172. Hajek, T., Franke, K., Kolenic, M., Capkova, J., Matejka, M., Propper, L., Uher, R., Stopkova, P., Novak, T., Paus, T., Kopecek, M., Spaniel, F., Alda, M.: Brain Age in Early Stages of Bipolar Disorders or Schizophrenia. Schizophr Bull 45, 190-198 (2019)
173. Kolenic, M., Franke, K., Hlinka, J., Matejka, M., Capkova, J., Pausova, Z., Uher, R., Alda, M., Spaniel, F., Hajek, T.: Obesity, dyslipidemia and brain age in first-episode psychosis. J Psychiatr Res 99, 151-158 (2018)
174. Koutsouleris, N., Davatzikos, C., Borgwardt, S., Gaser, C., Bottlender, R., Frodl, T., Falkai, P., Riecher-Rossler, A., Moller, H.J., Reiser, M., Pantelis, C., Meisenzahl, E.: Accelerated brain aging in schizophrenia and beyond: a neuroanatomical marker of psychiatric disorders. Schizophr Bull 40, 1140-1153 (2014)
175. Besteher, B., Gaser, C., Nenadic, I.: Machine-learning based brain age estimation in major depression showing no evidence of accelerated aging. Psychiatry Res Neuroimaging 290, 1-4 (2019)
176. Zannas, A.S., McQuoid, D.R., Steffens, D.C., Chrousos, G.P., Taylor, W.D.: Stressful life events, perceived stress, and 12-month course of geriatric depression: direct effects and moderation by the 5-HTTLPR and COMT Val158Met polymorphisms. Stress 15, 425-434 (2012)
177. Andreescu, C., Ajilore, O., Aizenstein, H.J., Albert, K., Butters, M.A., Landman, B.A., Karim, H.T., Krafty, R., Taylor, W.D.: Disruption of Neural Homeostasis as a Model of Relapse and Recurrence in Late-Life Depression. Am J Geriatr Psychiatry (2019)
178. Sheline, Y.I., Wang, P.W., Gado, M.H., Csernansky, J.G., Vannier, M.W.: Hippocampal atrophy in recurrent major depression. Proceedings of the National Academy of Sciences of the United States of America 93, 3908-3913 (1996)
179. Sheline, Y.I., Sanghavi, M., Mintun, M.A., Gado, M.H.: Depression duration but not age predicts hippocampal volume loss in medically healthy women with recurrent major depression. The Journal of Neuroscience 19, 5034-5043 (1999)

180. Taylor, W.D., McQuoid, D.R., Payne, M.E., Zannas, A.S., MacFall, J.R., Steffens, D.C.: Hippocampus Atrophy and the Longitudinal Course of Late-life Depression. Am J Geriatr Psychiatry 22, 1504-1512 (2014)
181. Diniz, B.S., Butters, M.A., Albert, S.M., Dew, M.A., Reynolds, C.F., 3rd: Late-life depression and risk of vascular dementia and Alzheimer's disease: systematic review and meta-analysis of community-based cohort studies. Br J Psychiatry 202, 329-335 (2013)
182. Taylor, W.D., McQuoid, D.R., Krishnan, K.R.: Medical comorbidity in late-life depression. International Journal of Geriatric Psychiatry 19, 935-943 (2004)
183. Gandelman, J.A., Albert, K., Boyd, B.D., Park, J.W., Riddle, M., Woodward, N.D., Kang, H., Landman, B.A., Taylor, W.D.: Intrinsic Functional Network Connectivity Is Associated With Clinical Symptoms and Cognition in Late-Life Depression. Biol Psychiatry Cogn Neurosci Neuroimaging 4, 160-170 (2019)
184. Franke, K., Gaser, C., Manor, B., Novak, V.: Advanced BrainAGE in older adults with type 2 diabetes mellitus. Front Aging Neurosci 5, 90 (2013)
185. Franke, K., Gaser, C.: Longitudinal changes in individual BrainAGE in healthy aging, mild cognitive impairment, and Alzheimer's disease. GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry 25, 235-245 (2012)
186. Gaser, C., Franke, K., Kloppel, S., Koutsouleris, N., Sauer, H., Alzheimer's Disease Neuroimaging, I.: BrainAGE in Mild Cognitive Impaired Patients: Predicting the Conversion to Alzheimer's Disease. PLoS One 8, e67346 (2013)
187. Montgomery, S.A., Asberg, M.: A new depression scale designed to be sensitive to change. British Journal of Psychiatry 134, 382-389 (1979)
188. Ekselius, L., Lindstrom, E., von Knorring, L., Bodlund, O., Kullgren, G.: SCID II interviews and the SCID Screen questionnaire as diagnostic tools for personality disorders in DSM-III-R. Acta Psychiatrica Scandinavica 90, 120-123 (1994)
189. Nasreddine, Z.S., Phillips, N.A., Bedirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J.L., Chertkow, H.: The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. Journal of the American Geriatrics Society 53, 695-699 (2005)
190. Folstein, M.F., Folstein, S.E., McHugh, P.R.: "Mini-mental state" a practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research 12, 189-198 (1975)
191. Albert, K.M., Potter, G.G., McQuoid, D.R., Taylor, W.D.: Cognitive performance in antidepressant-free recurrent major depressive disorder. Depress Anxiety 35, 694-699 (2018)
192. Saleh, A., Potter, G.G., McQuoid, D.R., Boyd, B., Turner, R., MacFall, J.R., Taylor, W.D.: Effects of early life stress on depression, cognitive performance and brain morphology. Psychol Med 47, 171-181 (2017)
193. Sheehan, D.V., Lecrubier, Y., Sheehan, K.H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., Dunbar, G.C.: The Mini-International Neuropsychiatric Interview (MINI): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. The Journal of clinical psychiatry (1998)
194. WorldHealthOrganization: Measuring health and disability : manual for WHO Disability Assessment Schedule WHODAS 2.0. WHO Press, Geneva (2010)
195. Taylor, W.D., Boyd, B., Turner, R., McQuoid, D.R., Ashley-Koch, A., MacFall, J.R., Saleh, A., Potter, G.G.: APOE epsilon4 associated with preserved executive function performance and maintenance of temporal and cingulate brain volumes in younger adults. Brain Imaging Behav 11, 194-204 (2017)
196. Evans, A.C., Collins, D.L., Mills, S., Brown, E., Kelly, R., Peters, T.M.: 3D statistical neuroanatomical models from 305 MRI volumes. In: 1993 IEEE conference record nuclear science symposium and medical imaging conference, pp. 1813-1817. IEEE, (Year)
197. Ourselin, S., Roche, A., Subsol, G., Pennec, X., Ayache, N.: Reconstructing a 3D structure from serial histological sections. Image and vision computing 19, 25-31 (2001)
198. Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C.: N4ITK: improved N3 bias correction. IEEE transactions on medical imaging 29, 1310 (2010)

199. Huo, Y., Blaber, J., Damon, S.M., Boyd, B.D., Bao, S., Parvathaneni, P., Noguera, C.B., Chaganti, S., Nath, V., Greer, J.M.: Towards portable large-scale image processing with high-performance computing. Journal of digital imaging 31, 304-314 (2018)
200. Schmidt, P., Gaser, C., Arsic, M., Buck, D., Forschler, A., Berthele, A., Hoshi, M., Ilg, R., Schmid, V.J., Zimmer, C., Hemmer, B., Muhlau, M.: An automated tool for detection of FLAIR-hyperintense white-matter lesions in Multiple Sclerosis. Neuroimage 59, 3774-3783 (2012)
201. Abi Zeid Daou, M., Boyd, B.D., Donahue, M.J., Albert, K., Taylor, W.D.: Frontocingulate cerebral blood flow and cerebrovascular reactivity associated with antidepressant response in late-life depression. J Affect Disord 215, 103-110 (2017)
202. Taylor, W.D., Aizenstein, H.J., Alexopoulos, G.S.: The vascular depression hypothesis: mechanisms linking vascular disease with depression. Mol Psychiatry 18, 963-974 (2013)
203. Park, J.H., Lee, S.B., Lee, J.J., Yoon, J.C., Han, J.W., Kim, T.H., Jeong, H.G., Newhouse, P.A., Taylor, W.D., Kim, J.H., Woo, J.I., Kim, K.W.: Epidemiology of MRI-defined vascular depression: A longitudinal, community-based study in Korean elders. J Affect Disord 180, 200-206 (2015)
204. Tromp, D., Dufour, A., Lithfous, S., Pebayle, T., Despres, O.: Episodic memory in normal aging and Alzheimer disease: Insights from imaging and behavioral studies. Ageing Res Rev 24, 232-262 (2015)
205. Czepielewski, L.S., Massuda, R., Panizzutti, B., Grun, L.K., Barbe-Tuana, F.M., Teixeira, A.L., Barch, D.M., Gama, C.S.: Telomere Length and CCL11 Levels are Associated With Gray Matter Volume and Episodic Memory Performance in Schizophrenia: Evidence of Pathological Accelerated Aging. Schizophr Bull 44, 158-167 (2018)
206. Degerman, S., Josefsson, M., Nordin Adolfsson, A., Wennstedt, S., Landfors, M., Haider, Z., Pudas, S., Hultdin, M., Nyberg, L., Adolfsson, R.: Maintained memory in aging is associated with young epigenetic age. Neurobiol Aging 55, 167-171 (2017)
207. Sheline, Y.I., Barch, D.M., Garcia, K., Gersing, K., Pieper, C., Welsh-Bohmer, K., Steffens, D.C., Doraiswamy, P.M.: Cognitive function in late life depression: relationships to depression severity, cerebrovascular risk factors and processing speed. Biol Psychiatry 60, 58-65 (2006)
208. Nebes, R.D., Butters, M.A., Mulsant, B.H., Pollock, B.G., Zmuda, M.D., Houck, P.R., Reynolds, C.F., 3rd: Decreased working memory and processing speed mediate cognitive impairment in geriatric depression. Psychol Med 30, 679-691 (2000)
209. Taylor, W.D., Wagner, H.R., Steffens, D.C.: Greater depression severity associated with less improvement in depression-associated cognitive deficits in older subjects. Am J Geriatr Psychiatry 10, 632-635 (2002)
210. Jokinen, H., Koikkalainen, J., Laakso, H.M., Melkas, S., Nieminen, T., Brander, A., Korvenoja, A., Rueckert, D., Barkhof, F., Scheltens, P., Schmidt, R., Fazekas, F., Madureira, S., Verdelho, A., Wallin, A., Wahlund, L.O., Waldemar, G., Chabriat, H., Hennerici, M., O'Brien, J., Inzitari, D., Lotjonen, J., Pantoni, L., Erkinjuntti, T.: Global Burden of Small Vessel Disease-Related Brain Changes on MRI Predicts Cognitive and Functional Decline. Stroke 51, 170-178 (2020)
211. Verlinden, V.J., van der Geest, J.N., de Groot, M., Hofman, A., Niessen, W.J., van der Lugt, A., Vernooij, M.W., Ikram, M.A.: Structural and microstructural brain changes predict impairment in daily functioning. Am J Med 127, 1089-1096 e1082 (2014)
212. Inzitari, D., Pracucci, G., Poggesi, A., Carlucci, G., Barkhof, F., Chabriat, H., Erkinjuntti, T., Fazekas, F., Ferro, J.M., Hennerici, M., Langhorne, P., O'Brien, J., Scheltens, P., Visser, M.C., Wahlund, L.O., Waldemar, G., Wallin, A., Pantoni, L.: Changes in white matter as determinant of global functional decline in older independent outpatients: three year follow-up of LADIS (leukoaraiosis and disability) study cohort. BMJ 339, b2477 (2009)
213. Teixeira, M.J., Cury, R.G., Galhardoni, R., Barboza, V.R., Brunoni, A.R., Alho, E., Lepski, G., de Andrade, D.C.: Deep brain stimulation of the dentate nucleus improves cerebellar ataxia after cerebellar stroke. Neurology 85, 2075-2076 (2015)
214. Wathen, C.A., Frizon, L.A., Maiti, T.K., Baker, K.B., Machado, A.G.: Deep brain stimulation of the cerebellum for poststroke motor rehabilitation: from laboratory to clinical trial. Neurosurgical focus 45, E13 (2018)

215. Coenen, V.A., Allert, N., Mädler, B.: A role of diffusion tensor imaging fiber tracking in deep brain stimulation surgery: DBS of the dentato-rubro-thalamic tract (drt) for the treatment of therapy-refractory tremor. Acta neurochirurgica 153, 1579-1585 (2011)
216. Fekete, R., Jankovic, J.: Revisiting the relationship between essential tremor and Parkinson's disease. Movement Disorders 26, 391-398 (2011)
217. Nicoletti, G., Manners, D., Novellino, F., Condino, F., Malucelli, E., Barbiroli, B., Tonon, C., Arabia, G., Salsone, M., Giofre, L.: Diffusion tensor MRI changes in cerebellar structures of patients with familial essential tremor. Neurology 74, 988-994 (2010)
218. Küper, M., Dimitrova, A., Thürling, M., Maderwald, S., Roths, J., Elles, H., Gizewski, E., Ladd, M.E., Diedrichsen, J., Timmann, D.: Evidence for a motor and a non-motor domain in the human dentate nucleus—an fMRI study. Neuroimage 54, 2612-2622 (2011)
219. Bernard, J.A., Peltier, S.J., Benson, B.L., Wiggins, J.L., Jaeggi, S.M., Buschkuehl, M., Jonides, J., Monk, C.S., Seidler, R.D.: Dissociable functional networks of the human dentate nucleus. Cerebral Cortex 24, 2151-2159 (2013)
220. Allen, G., McColl, R., Barnard, H., Ringe, W.K., Fleckenstein, J., Cullum, C.M.: Magnetic resonance imaging of cerebellar–prefrontal and cerebellar–parietal functional connectivity. Neuroimage 28, 39-48 (2005)
221. Solbach, K., Kraff, O., Minnerop, M., Beck, A., Schöls, L., Gizewski, E., Ladd, M., Timmann, D.: Cerebellar pathology in Friedreich's ataxia: atrophied dentate nuclei with normal iron content. NeuroImage: Clinical 6, 93-99 (2014)
222. Cocosco, C.A., Zijdenbos, A.P., Evans, A.C.: A fully automatic and robust brain MRI tissue classification method. Medical image analysis 7, 513-527 (2003)
223. Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P.: Automated model-based tissue classification of MR images of the brain. IEEE transactions on medical imaging 18, 897-908 (1999)
224. Wells, W.M., Grimson, W.E.L., Kikinis, R., Jolesz, F.A.: Adaptive segmentation of MRI data. IEEE transactions on medical imaging 15, 429-442 (1996)
225. He, N., Langley, J., Huddleston, D.E., Ling, H., Xu, H., Liu, C., Yan, F., Hu, X.P.: Improved Neuroimaging Atlas of the Dentate Nucleus. The Cerebellum 16, 951-956 (2017)
226. Diedrichsen, J., Balsters, J.H., Flavell, J., Cussans, E., Ramnani, N.: A probabilistic MR atlas of the human cerebellum. Neuroimage 46, 39-46 (2009)
227. Cabezas, M., Oliver, A., Lladó, X., Freixenet, J., Cuadra, M.B.: A review of atlas-based segmentation for magnetic resonance brain images. Computer methods and programs in biomedicine 104, e158-e177 (2011)
228. Doan, N.T., de Xivry, J.O., Macq, B.: Effect of inter-subject variation on the accuracy of atlas-based segmentation applied to human brain structures. In: Medical Imaging 2010: Image Processing, pp. 76231S. International Society for Optics and Photonics, (Year)
229. Dang, L.C., Castrellon, J.J., Perkins, S.F., Le, N.T., Cowan, R.L., Zald, D.H., Samanez-Larkin, G.R.: Reduced effects of age on dopamine D2 receptor levels in physically active adults. Neuroimage 148, 123-129 (2017)
230. Behrens, T.E., Berg, H.J., Jbabdi, S., Rushworth, M.F., Woolrich, M.W.: Probabilistic diffusion tractography with multiple fibre orientations: What can we gain? Neuroimage 34, 144-155 (2007)
231. Hagemeier, J., Geurts, J.J., Zivadinov, R.: Brain iron accumulation in aging and neurodegenerative disorders. Expert review of neurotherapeutics 12, 1467-1480 (2012)
232. Deoni, S.C., Catani, M.: Visualization of the deep cerebellar nuclei using quantitative T1 and ρ magnetic resonance imaging at 3 Tesla. Neuroimage 37, 1260-1266 (2007)
233. Acosta-Cabronero, J., Betts, M.J., Cardenas-Blanco, A., Yang, S., Nestor, P.J.: In vivo MRI mapping of brain iron deposition across the adult lifespan. Journal of Neuroscience 36, 364-374 (2016)
234. He, N., Huang, P., Ling, H., Langley, J., Liu, C., Ding, B., Huang, J., Xu, H., Zhang, Y., Zhang, Z.: Dentate nucleus iron deposition is a potential biomarker for tremor-dominant Parkinson's disease. NMR in Biomedicine 30, e3554 (2017)

235. Iglesias, J.E., Sabuncu, M.R.: Multi-atlas segmentation of biomedical images: a survey. Medical image analysis 24, 205-219 (2015)
236. Huo, Y., Xu, Z., Aboud, K., Parvathaneni, P., Bao, S., Bermudez, C., Resnick, S.M., Cutting, L.E., Landman, B.A.: Spatially Localized Atlas Network Tiles Enables 3D Whole Brain Segmentation from Limited Data. arXiv preprint arXiv:1806.00546 (2018)
237. de Brebisson, A., Montana, G.: Deep neural networks for anatomical brain segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 20-28. (Year)
238. Kaku, A., Hegde, C.V., Huang, J., Chung, S., Wang, X., Young, M., Lui, Y.W., Razavian, N.: DARTS: DenseUnet-based Automatic Rapid Tool for brain Segmentation. arXiv preprint arXiv:1911.05567 (2019)
239. Dai, C., Mo, Y., Angelini, E., Guo, Y., Bai, W.: Transfer Learning from Partial Annotations for Whole Brain Segmentation. Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data, pp. 199-206. Springer (2019)
240. Coupé, P., Mansencal, B., Clément, M., Giraud, R., de Senneville, B.D., Ta, V.-T., Lepetit, V., Manjon, J.V.: AssemblyNet: A large ensemble of CNNs for 3D Whole Brain MRI Segmentation. arXiv preprint arXiv:1911.09098 (2019)
241. Xiao, B., Cheng, X., Li, Q., Wang, Q., Zhang, L., Wei, D., Zhan, Y., Zhou, X.S., Xue, Z., Lu, G.: Weakly Supervised Confidence Learning for Brain MR Image Dense Parcellation. In: International Workshop on Machine Learning in Medical Imaging, pp. 409-416. Springer, (Year)
242. Zhao, Y.-X., Zhang, Y.-M., Song, M., Liu, C.-L.: Multi-view Semi-supervised 3D Whole Brain Segmentation with a Self-ensemble Network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 256-265. Springer, (Year)
243. Xiong, Y., Huo, Y., Wang, J., Davis, L.T., McHugo, M., Landman, B.A.: Reproducibility evaluation of SLANT whole brain segmentation across clinical magnetic resonance imaging protocols. In: Medical Imaging 2019: Image Processing, pp. 109492V. International Society for Optics and Photonics, (Year)
244. Mostapha, M., Styner, M.: Role of deep learning in infant brain MRI analysis. Magnetic resonance imaging (2019)
245. Wang, L., Nie, D., Li, G., Puybareau, É., Dolz, J., Zhang, Q., Wang, F., Xia, J., Wu, Z., Chen, J.-W.: Benchmark on automatic six-month-old infant brain segmentation algorithms: the iSeg-2017 challenge. IEEE transactions on medical imaging 38, 2219-2230 (2019)
246. Zhao, T., Liao, X., Fonov, V.S., Wang, Q., Men, W., Wang, Y., Qin, S., Tan, S., Gao, J.-H., Evans, A.: Unbiased age-specific structural brain atlases for Chinese pediatric population. Neuroimage 189, 55-70 (2019)
247. Cheplygina, V., de Bruijne, M., Pluim, J.P.: Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical image analysis 54, 280-296 (2019)
248. van Opbroek, A., Vernooij, M.W., Ikram, M.A., de Bruijne, M.: Weighting training images by maximizing distribution similarity for supervised segmentation across scanners. Medical image analysis 24, 245-254 (2015)
249. Kouw, W.M., Loog, M., Bartels, L.W., Mendrik, A.M.: Mr acquisition-invariant representation learning. arXiv preprint arXiv:1709.07944 (2017)
250. Guerrero, R., Ledig, C., Rueckert, D.: Manifold alignment and transfer learning for classification of Alzheimer's disease. In: International Workshop on Machine Learning in Medical Imaging, pp. 77-84. Springer, (Year)
251. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in neural information processing systems, pp. 3320-3328. (Year)
252. Torrey, L., Shavlik, J.: Transfer learning. Handbook of research on machine learning applications and trends: algorithms, methods, and techniques, pp. 242-264. IGI Global (2010)

253. Dar, S.U., Yurt, M., Karacan, L., Erdem, A., Erdem, E., Çukur, T.: Image synthesis in multi-contrast MRI with conditional generative adversarial networks. IEEE transactions on medical imaging (2019)
254. Narayana, P.A., Coronado, I., Sujit, S.J., Sun, X., Wolinsky, J.S., Gabr, R.E.: Are multi-contrast magnetic resonance images necessary for segmenting multiple sclerosis brains? A large cohort study based on deep learning. Magnetic resonance imaging 65, 8-14 (2020)
255. Gong, E., Pauly, J.M., Wintermark, M., Zaharchuk, G.: Deep learning enables reduced gadolinium dose for contrast-enhanced brain MRI. Journal of Magnetic Resonance Imaging 48, 330-340 (2018)
256. Chen, K.T., Gong, E., de Carvalho Macruz, F.B., Xu, J., Boumis, A., Khalighi, M., Poston, K.L., Sha, S.J., Greicius, M.D., Mormino, E.: Ultra–Low-Dose 18F-Florbetaben Amyloid PET Imaging Using Deep Learning with Multi-Contrast MRI Inputs. Radiology 290, 649-656 (2018)
257. Welander, P., Karlsson, S., Eklund, A.: Generative adversarial networks for image-to-image translation on multi-contrast MR images-A comparison of CycleGAN and UNIT. arXiv preprint arXiv:1806.07777 (2018)
258. Fortin, J.-P., Parker, D., Tunç, B., Watanabe, T., Elliott, M.A., Ruparel, K., Roalf, D.R., Satterthwaite, T.D., Gur, R.C., Gur, R.E.: Harmonization of multi-site diffusion tensor imaging data. Neuroimage 161, 149-170 (2017)
259. Shinohara, R.T., Oh, J., Nair, G., Calabresi, P.A., Davatzikos, C., Doshi, J., Henry, R.G., Kim, G., Linn, K.A., Papinutto, N.: Volumetric analysis from a harmonized multisite brain MRI study of a single subject with multiple sclerosis. American Journal of Neuroradiology 38, 1501-1509 (2017)
260. Prince, J.L., Carass, A., Zhao, C., Dewey, B.E., Roy, S., Pham, D.L.: Image synthesis and superresolution in medical imaging. Handbook of Medical Image Computing and Computer Assisted Intervention, pp. 1-24. Elsevier (2020)
261. Roy, S., Carass, A., Shiee, N., Pham, D.L., Prince, J.L.: MR contrast synthesis for lesion segmentation. In: Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on, pp. 932-935. IEEE, (Year)
262. Roy, S., Carass, A., Jog, A., Prince, J.L., Lee, J.: MR to CT registration of brains using image synthesis. In: Proceedings of SPIE. NIH Public Access, (Year)
263. Plassard, A.J., McHugo, M., Heckers, S., Landman, B.A.: Multi-scale hippocampal parcellation improves atlas-based segmentation accuracy. In: Medical Imaging 2017: Image Processing, pp. 101332D. International Society for Optics and Photonics, (Year)
264. Iglesias, J.E., Van Leemput, K., Augustinack, J., Insausti, R., Fischl, B., Reuter, M., Initiative, A.s.D.N.: Bayesian longitudinal segmentation of hippocampal substructures in brain MRI using subject-specific atlases. Neuroimage 141, 542-555 (2016)
265. Iglesias, J.E., Augustinack, J.C., Nguyen, K., Player, C.M., Player, A., Wright, M., Roy, N., Frosch, M.P., McKee, A.C., Wald, L.L.: A computational atlas of the hippocampal formation using ex vivo, ultra-high resolution MRI: application to adaptive segmentation of in vivo MRI. Neuroimage 115, 117-137 (2015)
266. Nogovitsyn, N., Souza, R., Muller, M., Srajer, A., Hassel, S., Arnott, S.R., Davis, A.D., Hall, G.B., Harris, J.K., Zamyadi, M.: Testing a deep convolutional neural network for automated hippocampus segmentation in a longitudinal sample of healthy participants. NeuroImage 197, 589-597 (2019)
267. Wu, Z., Gao, Y., Shi, F., Ma, G., Jewells, V., Shen, D.: Segmenting hippocampal subfields from 3T MRI with multi-modality images. Medical image analysis 43, 10-22 (2018)
268. Yushkevich, P.A., Amaral, R.S., Augustinack, J.C., Bender, A.R., Bernstein, J.D., Boccardi, M., Bocchetta, M., Burggren, A.C., Carr, V.A., Chakravarty, M.M.: Quantitative comparison of 21 protocols for labeling hippocampal subfields and parahippocampal subregions in in vivo MRI: towards a harmonized segmentation protocol. Neuroimage 111, 526-541 (2015)
269. Bannerman, D., Rawlins, J., McHugh, S., Deacon, R., Yee, B., Bast, T., Zhang, W.-N., Pothuizen, H., Feldon, J.: Regional dissociations within the hippocampus—memory and anxiety. Neuroscience & biobehavioral reviews 28, 273-283 (2004)

270. Eichenbaum, H., Dudchenko, P., Wood, E., Shapiro, M., Tanila, H.: The hippocampus, memory, and place cells: is it spatial memory or a memory space? Neuron 23, 209-226 (1999)
271. Hyman, B.T., Van Hoesen, G.W., Damasio, A.R., Barnes, C.L.: Alzheimer's disease: cell-specific pathology isolates the hippocampal formation. Science 225, 1168-1170 (1984)
272. Velakoulis, D., Pantelis, C., McGorry, P.D., Dudgeon, P., Brewer, W., Cook, M., Desmond, P., Bridle, N., Tierney, P., Murrie, V.: Hippocampal volume in first-episode psychoses and chronic schizophrenia: a high-resolution magnetic resonance imaging study. Archives of general psychiatry 56, 133-141 (1999)
273. Yonekawa, W.D., Kapetanovic, I.M., Kupferberg, H.J.: The effects of anticonvulsant agents on 4-aminopyridine induced epileptiform activity in rat hippocampus in vitro. Epilepsy research 20, 137-150 (1995)
274. Hao, Y., Wang, T., Zhang, X., Duan, Y., Yu, C., Jiang, T., Fan, Y., Initiative, A.s.D.N.: Local label learning (LLL) for subcortical structure segmentation: application to hippocampus segmentation. Human brain mapping 35, 2674-2697 (2014)
275. Reuter, M., Schmansky, N.J., Rosas, H.D., Fischl, B.: Within-subject template estimation for unbiased longitudinal image analysis. Neuroimage 61, 1402-1418 (2012)
276. Huo, Y., Asman, A.J., Plassard, A.J., Landman, B.A.: Simultaneous total intracranial volume and posterior fossa volume estimation using multi-atlas label fusion. Human brain mapping 38, 599-616 (2017)
277. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. (2017)
278. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
279. Fischl, B.: FreeSurfer. Neuroimage 62, 774-781 (2012)
280. Remedios, S., Pham, D.L., Butman, J.A., Roy, S.: Classifying magnetic resonance image modalities with convolutional neural networks. In: Medical Imaging 2018: Computer-Aided Diagnosis, pp. 105752I. International Society for Optics and Photonics, (Year)
281. Pizarro, R., Assemlal, H.-E., De Nigris, D., Elliott, C., Antel, S., Arnold, D., Shmuel, A.: Using deep learning algorithms to automatically identify the brain MRI contrast: implications for managing large databases. Neuroinformatics 17, 115-130 (2019)
282. Tang, Y., Lee, H.H., Xu, Y., Tang, O., Chen, Y., Gao, D., Han, S., Gao, R., Bermudez, C., Savona, M.R.: Contrast Phase Classification with a Generative Adversarial Network. arXiv preprint arXiv:1911.06395 (2019)
283. Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Nori, A., Criminisi, A., Rueckert, D.: Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: International conference on information processing in medical imaging, pp. 597-609. Springer, (Year)
284. Worker, A., Dima, D., Combes, A., Crum, W.R., Streffer, J., Einstein, S., Mehta, M.A., Barker, G.J., CR Williams, S., O'daly, O.: Test–retest reliability and longitudinal analysis of automated hippocampal subregion volumes in healthy ageing and A lzheimer's disease populations. Human brain mapping 39, 1743-1754 (2018)
285. Mueller, S.G., Yushkevich, P.A., Das, S., Wang, L., Van Leemput, K., Iglesias, J.E., Alpert, K., Mezher, A., Ng, P., Paz, K.: Systematic comparison of different techniques to measure hippocampal subfield volumes in ADNI2. NeuroImage: Clinical 17, 1006-1018 (2018)
286. Adelborg, K., Horváth-Puhó, E., Ording, A., Pedersen, L., Sørensen, H.T., Henderson, V.W.: Heart failure and risk of dementia: a Danish nationwide population-based cohort study. European journal of heart failure 19, 253-260 (2017)
287. Bielinski, S.J., Pathak, J., Carrell, D.S., Takahashi, P.Y., Olson, J.E., Larson, N.B., Liu, H., Sohn, S., Wells, Q.S., Denny, J.C.: A robust e-epidemiology tool in phenotyping heart failure with differentiation for preserved and reduced ejection fraction: the electronic medical records and genomics (eMERGE) network. Journal of cardiovascular translational research 8, 475-483 (2015)

288. Bermudez, C., Blaber, J., Remedios, S.W., Reynolds, J.E., Lebel, C., McHugo, M., Heckers, S., Huo, Y., Landman, B.A.: Generalizing Deep Whole Brain Segmentation for Pediatric and Post-Contrast MRI with Augmented Transfer Learning. arXiv preprint arXiv:1908.04702 (2019)
289. Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E.: Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23, S208-S219 (2004)
290. Momjian-Mayor, I., Baron, J.-C.: The pathophysiology of watershed infarction in internal carotid artery disease: review of cerebral perfusion studies. Stroke 36, 567-577 (2005)
291. Suter, O.-C., Sunthorn, T., Kraftsik, R., Straubel, J., Darekar, P., Khalili, K., Miklossy, J.: Cerebral hypoperfusion generates cortical watershed microinfarcts in Alzheimer disease. Stroke 33, 1986-1992 (2002)
292. Huang, C.-W., Hsu, S.-W., Chang, Y.-T., Huang, S.-H., Huang, Y.-C., Lee, C.-C., Chang, W.-N., Lui, C.-C., Chen, N.-C., Chang, C.-C.: Cerebral perfusion insufficiency and relationships with cognitive deficits in Alzheimer's disease: A multiparametric neuroimaging study. Scientific reports 8, 1-14 (2018)
293. Alosco, M.L., Hayes, S.M.: Structural brain alterations in heart failure: a review of the literature and implications for risk of Alzheimer's disease. Heart failure reviews 20, 561-571 (2015)
294. Seo, S.W., Ahn, J., Yoon, U., Im, K., Lee, J.M., Tae Kim, S., Ahn, H.J., Chin, J., Jeong, Y., Na, D.L.: Cortical thinning in vascular mild cognitive impairment and vascular dementia of subcortical type. Journal of Neuroimaging 20, 37-45 (2010)
295. Liu, C., Li, C., Gui, L., Zhao, L., Evans, A.C., Xie, B., Zhang, J., Wei, L., Zhou, D., Wang, J.: The pattern of brain gray matter impairments in patients with subcortical vascular dementia. Journal of the neurological sciences 341, 110-118 (2014)
296. Hirao, K., Ohnishi, T., Hirata, Y., Yamashita, F., Mori, T., Moriguchi, Y., Matsuda, H., Nemoto, K., Imabayashi, E., Yamada, M.: The prediction of rapid conversion to Alzheimer's disease in mild cognitive impairment using regional cerebral blood flow SPECT. Neuroimage 28, 1014-1021 (2005)
297. Kumar, R., Woo, M.A., Macey, P.M., Fonarow, G.C., Hamilton, M.A., Harper, R.M.: Brain axonal and myelin evaluation in heart failure. Journal of the neurological sciences 307, 106-113 (2011)
298. Chan, D., Fox, N.C., Scahill, R.I., Crum, W.R., Whitwell, J.L., Leschziner, G., Rossor, A.M., Stevens, J.M., Cipolotti, L., Rossor, M.N.: Patterns of temporal lobe atrophy in semantic dementia and Alzheimer's disease. Annals of neurology 49, 433-442 (2001)
299. Pan, A., Kumar, R., Macey, P.M., Fonarow, G.C., Harper, R.M., Woo, M.A.: Visual assessment of brain magnetic resonance imaging detects injury to cognitive regulatory sites in patients with heart failure. Journal of cardiac failure 19, 94-100 (2013)
300. Kumar, R., Woo, M.A., Birrer, B.V., Macey, P.M., Fonarow, G.C., Hamilton, M.A., Harper, R.M.: Mammillary bodies and fornix fibers are injured in heart failure. Neurobiology of disease 33, 236-242 (2009)
301. Lang, C.C., Mancini, D.M.: Non-cardiac comorbidities in chronic heart failure. Heart 93, 665-671 (2007)
302. Chaganti, S., Welty, V.F., Taylor, W., Albert, K., Failla, M.D., Cascio, C., Smith, S., Mawn, L., Resnick, S.M., Beason-Held, L.L.: Discovering novel disease comorbidities using electronic medical records. PloS one 14, (2019)
303. Braunstein, J.B., Anderson, G.F., Gerstenblith, G., Weller, W., Niefeld, M., Herbert, R., Wu, A.W.: Noncardiac comorbidity increases preventable hospitalizations and mortality among Medicare beneficiaries with chronic heart failure. Journal of the American College of Cardiology 42, 1226-1233 (2003)