EXPLAINABLE AI IN MEDICAL IMAGING:
INTERPRETING MULTI-MODALITY INFERENCE WITH NEUROIMAGING AND EHR

By

Cailey Irene Kerley

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
of the degree of

DOCTOR OF PHILOSOPHY
in
Electrical Engineering
December 17, 2022
Nashville, Tennessee

Approved:

Bennett A. Landman, Ph.D.
Laurie E. Cutting, Ph.D.
Yuankai Huo, Ph.D.
Thomas Lasko, Ph.D.
Seth Smith, Ph.D.

For Andrew

whose support, bad jokes, coffee, and hugs helped make this work a reality.

ACKNOWLEDGEMENTS

I have been immensely lucky to work with and learn from some of the most brilliant people I have ever met while on this adventure in graduate school. Since my very first day, I have constantly met faculty, staff, clinicians, and other students who are always eager to chat about science and discover new things. I will sincerely miss this environment of boundless curiosity and energy. I want to thank the various funding sources which supported my work, including the Vanderbilt Kennedy Center, for giving me the freedom to pursue many interesting rabbit holes over the last four years.

I want to express special thanks to my committee members for their support and guidance in shaping my dissertation research and growth as a scientist. I am very grateful to Drs. Cutting, Lasko, and Smith for investing time in answering my many questions, helping me think about the bigger picture, and finding the common threads flowing through my work. I am especially thankful to Dr. Huo, who has been a close mentor and teacher since my first day at Vanderbilt. And of course, I want to express my immense thanks to Dr. Landman for pushing me to grow, training me to be curious, and teaching me that collaboration is both the most important and most fulfilling part of science.

There are many other amazing people at Vanderbilt who were an important part of this time in my life. I would like to thank Dr. Rex for her support and enthusiasm when I was a new grad student. Thank you as well to Dr. Aboud for being both a great mentor and friend. There are too many awesome fellow students to mention, but you all have made this adventure worth it even when the work was stressful. I want to thank VISE and WoV for simultaneously fostering excellent science and social communities. In particular, thank you to Roza, Haley, Nhung, and Sarah for reminding me to have fun and always being willing to chat.

My time in grad school would not have been nearly as fun or productive without the amazing members of the MASI lab. Thank you especially to Cam, for never getting tired of my questions and making me play Pokémon Go; to Karthik for fixing my many technical problems and teaching me what real Biryani tastes like; to Lucas for showing me his plots and reminding me that life is a journey; and to Leon for answering all of my medical questions and being the best coauthor and desk/snack buddy.

Finally, none of this work would have been possible without the love and encouragement of my friends and family, of whom there are too many to name here. In particular, thank you to Aaron for the copious amounts of pasta made during my final months of research; to my sister Ashley for being my best friend and confidant; and to my parents Bill and Angie for always believing in me. Above all, thank you to my amazing partner Andrew (and our two mischievous cats Hamilton and Francie); their support and love got me through the most difficult times.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter I

Introduction

All models are wrong, but some are useful. – George Box, Ph.D.

1. Overview

Elucidating the intricate inner workings of the human body has long been a topic of interest for humankind. Over the last century, innovations in medical imaging methods have allowed us to examine internal structures *in vivo*, revolutionizing our understanding of our own biology. This revolution began with the discovery of X-ray imaging in 1895 by Wilhelm Röntgen [1]. Since then, the introductions of other imaging modalities, such as computed tomography, ultrasound, and magnetic resonance imaging (MRI), have increasingly provided researchers and clinicians alike with clearer, more detailed depictions of the complex systems and structures which give rise to human life. As this discipline of medical imaging grew, the field of medical image analysis emerged alongside. With increasing availability of routine high-resolution medical imaging, researchers saw an opportunity to extract objective, quantitative information from these images [1].

Originally, all medical image analysis was performed by trained experts; these clinicians would visually assess medical images to identify lesions and diagnose disease. Expert human labor, however, is both time-consuming and expensive, necessitating a shift toward automated medical image analysis systems. Today, those expert clinicians work alongside a multi-disciplinary team of biologists, computer scientists, engineers, and mathematicians to develop quantitative and efficient automated image processing methods. This collaboration generally begins with the clinician manually annotating a small set of medical images for a particular task, such as organ segmentation or disease classification. This dataset of raw images and annotations is then given to the non-clinical scientists, whose goal is to design a model that mimics the task of the clinician. Manual annotations, also called the *ground truth*, are considered the be the high-quality "gold standard," and all analysis techniques invented by the non-clinical team are compared to it [2].

In the early days of medical image analysis, classic machine learning models were the primary analysis tool. Adapted from the field of computer vision [3], [4], these methods consisted of two primary phases: feature extraction and analysis. In the first phase, anatomical features, such as the volume or average intensity of various structures, were carefully extracted by domain experts. Following this, the feature set may be used for several different analyses, including detecting differences between healthy and disease

populations and training predictive models. These early models were both accurate and interpretable due to their specially curated feature sets and small sample sizes. At the same time, however, generating the features for such a model was a laborious undertaking, and these small cohort sizes prohibited these models from generalizing to unseen populations [5].

Recent advances in computing power and data storage systems have enabled a rapid increase in the availability and accessibility of medical imaging and electronic health record (EHR) data, introducing the "big data" era of medical research [6]–[8]. This increase in cohort size enables medical image analysis researchers to employ deep learning techniques, a subclass of machine learning which is able to perform highly complex tasks from sufficiently large datasets [9]. Applications of deep learning for medical imaging tasks have boasted incredible performance statistics, rivaling those of human experts [10].

Yet, challenges remain. Despite the fact that we now have access to more diverse sources of data than ever before, most current methods in medical image analysis research focus on a single data source. In contrast, clinicians fuse information from many data sources during the process of diagnosis and treatment, including multi-modal imaging, patient demographics, medical history, and others. One major area of opportunity in medical image analysis involves the development of deep learning models that can combine diverse data sources in a similar way. Another crucial research direction is that of interpreting deep learning models. Although deep learning models have quickly proven to excel at many medical imaging tasks, they function as black-boxes: the complex mapping from input data to output prediction is uninterpretable even to the scientists that train them [11]. Regardless of the brilliance of reported accuracy metrics, reliable interpretation methods must be developed for these models before we can expect them to be deployed in the high-stakes environment of a medical clinic.

In this dissertation, I outline several methods aimed at addressing specific challenges in performing multi-modal medical image analysis via interpretable machine learning techniques. I demonstrate these methods on multi-modal MRI and EHR inference in the context of mild traumatic brain injury, developmental disorders, and mild cognitive impairment. The remainder of the chapter is dedicated to providing context for topics relevant to these experiments including: modern machine learning methods for inference in medical image analysis; approaches for multi-modal image analysis with EHR; interpretable artificial intelligence techniques; and clinical applications.

2. Machine Learning Inference for Medical Image Analysis

Machine learning is the art and science of extracting hidden information from sets of data. The field may broadly be split into two types of tasks: *supervised* and *unsupervised* learning. Supervised learning

(Figure I-1) is concerned with finding a mapping $f(x)$ that captures the relationship between an input x and its corresponding output y. This is accomplished by learning from sets of paired examples (x, y).

Examples of supervised learning in medical image analysis include skin lesion classification [12] and organ segmentation [13]. In contrast, unsupervised learning focuses simply on identifying patterns within a dataset. These methods first make assumptions about the characteristics (for example, low variance) and structure (for example, clusters) of the hidden patterns, and then extract features from the dataset that adhere to the assumed structure. Dimensionality reduction, clustering algorithms, and anomaly detection are all unsupervised machine learning techniques.

A second dimension along which the field of machine learning may be examined is model type: *classic machine learning* versus *deep learning* models. The remainder of this section is dedicated to describing the characteristics of and distinctions between these two types of models. For the sake of brevity, this discussion will be limited to classification models and will focus on model interpretability, as these concepts encompasses the main applications of this dissertation.

2.1. Classic machine learning approaches

There are many classical machine learning methods that may be employed for disease classification and characterization [14], [15]. The logistic regression (Figure I-1-A) is one of the simplest and most

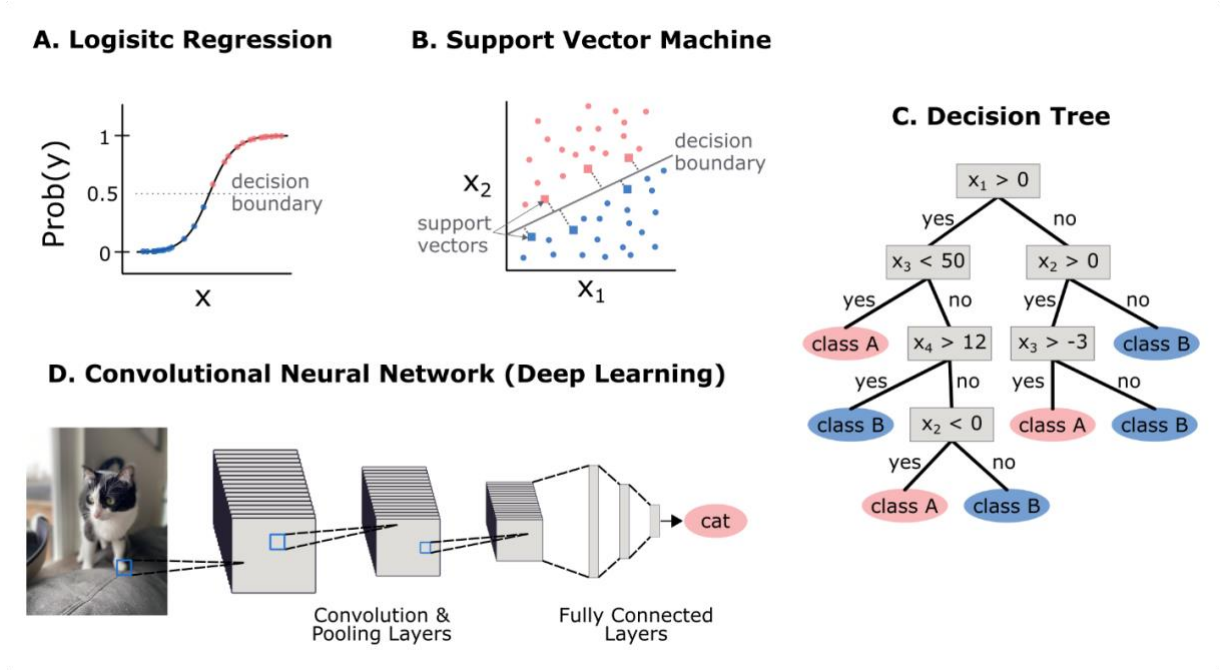


Figure I-1 Supervised learning models for classification

popular machine learning models for binary classification tasks; it models the probability of a sample being from the positive (disease) class as a logistic function of one or more variables. SVMs (Figure I-1-B) use a subset of samples from each class (support vectors) to find a boundary (hyperplane) which optimally separates the two classes. Decision trees (Figure I-1-C) learn to separate classes based on hierarchical decision making; often, multiple decision trees are combined to make an ensembled classifier called a Random Forest. These three methods have been widely used to successfully identify various conditions, including cognitive decline [16], diseases of the optic nerve [17], and migraine [18]. Additionally, predictions made by these models are fairly explainable. In the logistic regression model, each predictor variable has a corresponding constant, β, which describes the increase in the log-odds of the positive class for a unit increase in the predictor. SVM predictions may be explained in general by identifying which input features most influence the hyperplane and for specific inputs by measuring the distance between a sample and the hyperplane [19]. Similarly, predictions from decision trees may be explained by determining which input features are involved in the most influential decision nodes [20].

Despite this, these classical models suffer from several drawbacks. One of the primary challenges for classic machine learning approaches is their reliance on hand-crafted features [9]. Consider, for example, the task of predicting schizophrenia onset based on structural MRI measurements [21]. In this study, the raw data consists of an MRI volume and a label denoting whether or not each patient has schizophrenia. Before predictive modeling can occur, however, both the gray and white matter cortical surfaces of each patient must be parcellated into 68 regions, either by a domain expert or an automated segmentation algorithm. Following this, the cortical thickness must be measured for all 68 regions, along with the volume of 6 hand-selected interior brain regions, all before finally training a predictive model to determine the relationship between structural brain features and schizophrenia. The success of this pipeline is driven largely by both the quality and quantity of extracted features. Based on recent increases in computing power and innovations in automated segmentation, one possible solution to the manual feature extraction problem is to train a classification model using every possible automatically measured biomarker. This approach, however, leads to the curse of dimensionality: the idea that an algorithm becomes increasingly more difficult and costly to optimize as the number of input variables (or dimensions) increases [22]. It can be seen, then, that while classic approaches strike a decent balance between interpretability and accuracy, they are best suited for tasks that involve a small number of carefully constructed features. To overcome these challenges, in the next section we will delve into the topic of deep learning. The highly complex models in this research area can integrate both feature extraction and inference, allowing researchers to circumvent the laborious phase of manual feature extraction.

2.2. Deep learning approaches

To understand the challenges surrounding deep learning in medical image analysis, it is first necessary to have a high-level understand of how these models operate. Figure I-2 illustrates several essential concepts for the field of deep learning. The most basic building block of a deep learning model is the perceptron (Figure I-2-A). The perceptron may have any number of inputs, which are mapped to the output via a weighted summation followed by a non-linear activation function. Each input connection has a unique associated weight, which may be adjusted through supervised training to produce different patterns of output activation. In this way, perceptrons act as flexible feature detectors, with the output becoming active only when a particular combination of input features is detected. Though interesting in isolation, perceptrons are most powerful when strung together to create an *artificial neural network* (ANN) (Figure I-2-B). These networks consist of one or more *layers* of perceptrons; each layer in the network has a specific

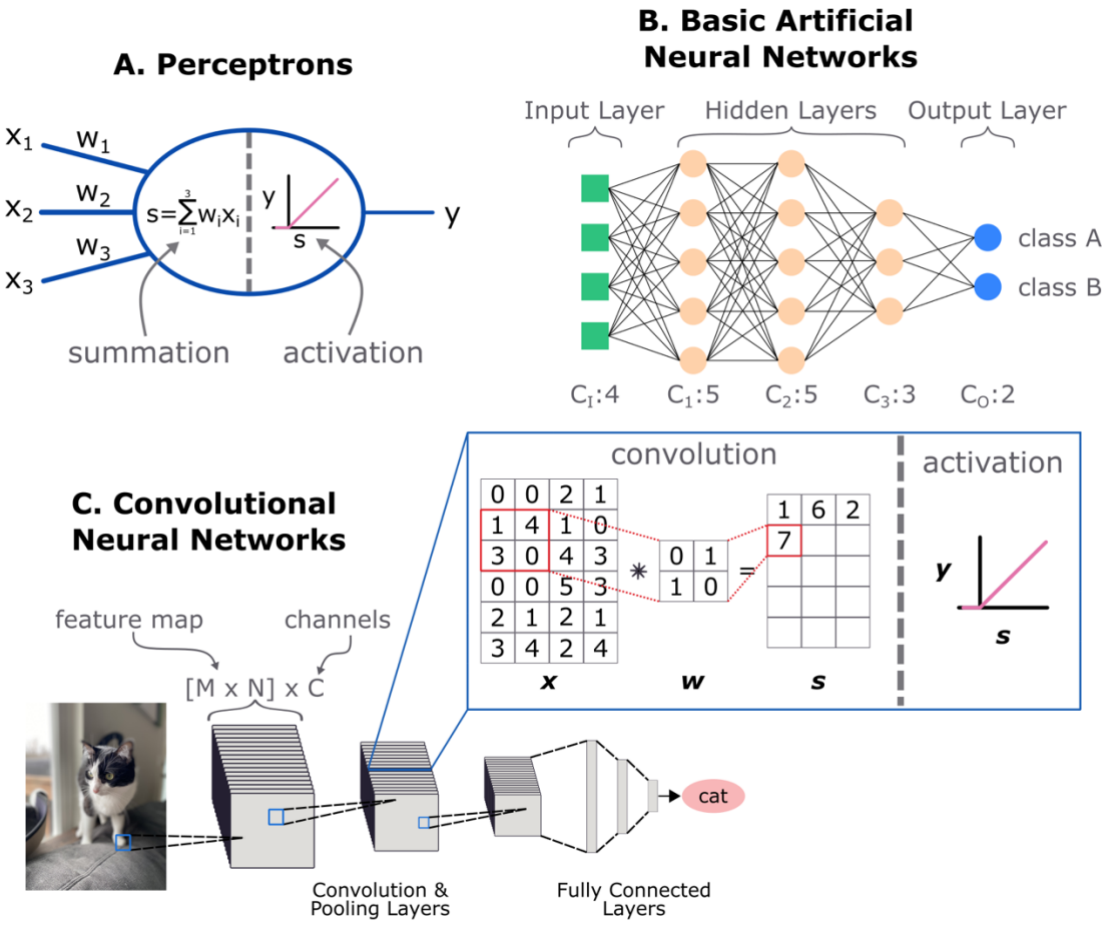


Figure I-2 Essential concepts in deep learning.

number of perceptrons (called channels), the outputs of which are connected to all channels in the next layer. These are called *fully connected* layers because all channels from one layer are connected to all channels of the next layer. As information flows from the input layer to the output layer, the hidden middle layers perform feature detection and pass their activations along to each successive layer deeper in the network. The final output layer is then composed of one channel per data class; to perform inference, a sample is fed through the network and the output channel with the highest activation determines the predicted class.

With this flexible architecture in place, ANNs may be taught to perform many different types of classification through example learning. The training process involves first performing inference: example data is fed into the network and the network's predicted class is recorded. Next, the *loss* (prediction error) is calculated by finding the difference between the predicted class and the true class. That loss is then backpropagated through the network, updating the perceptron weights at every layer with the goal of pushing the network's ultimate prediction toward that sample's true label. This process is then repeated many times for many different examples, with each example introducing a small amount of new information. In this way, ANNs learn complex non-linear mappings between input and output, allowing them to achieve incredible performances on computer vision and medical imaging tasks alike [9], [23].

Basic ANNs as described above are excellent for processing tabular information, but *convolutional* neural networks (CNNs) are used for analyzing two-dimensional (2D) or three-dimensional (3D) data [24], [25]. In CNNs, the perceptron's weighted-summation step is replaced by convolution (Figure I-2-C) to facilitate the capture of spatial information. Convolution involves sliding a small 2D or 3D filter across an image in incremental steps. At each step, pixels from the input image that overlap with the filter are extracted, multiplied by the filter's weights, and summed. In this way, convolution performs location-invariant feature extraction, where the type of extracted feature is determined by the arrangement of the filter's weights. Similar to the basic ANN's perceptron weights, the convolutional filter's weights are updated via backpropagation, so the network training phase iteratively adjusts what features are extracted at each layer of the network in an effort to find features relevant to the output task.

In the past decade, many methodological advancements have been made in an effort to make these models more accurate [26]. While such advancements in accuracy are good, performance is not the most pressing challenge associated with deep learning and medical image processing. One issue complicating the deployment of deep learning models is overfitting. Deep learning methods have so many trainable parameters compared to the relatively small size of training datasets, that they tend to overfit to the training data (for example, instead of learning universal discriminative features, they learn noise patterns in the training data). As a result, many models cannot generalize to unseen datasets. Methods such as dropout [27]

have been developed to reduce this overfitting, but the small size of medical imaging training datasets mean that this issue will persist for the near future.

Another major challenge associated with deep learning methods is that they are not interpretable. In the last section, we described several classic machine learning methods that provide easily-dissectible predictions. A trained CNN, however, contains thousands of learned parameters and feature maps in its hidden layers, reaching such a scale that is impossible for humans to comprehend. This lack of interpretability is particularly problematic in the high-stakes world of medicine, as the ultimate goal of medical image analysis research is to deploy our models into the clinic [28]. The field of explainable artificial intelligence (XAI) has recently emerged in an effort to address this interpretability problem, as we will later discuss in Section 4.

3. Multi-Modal Image Processing and Patient Context

Most medical image analysis is concerned with processing a single type of input data. Abdominal CT is used to segment organs [13]; the characteristics of autism spectrum disorder are examined via EHR [29]; or the neurological onset of Alzheimer's Disease is studied in longitudinal MRI [30]. In contrast, clinicians performing similar tasks synthesize varying data sources in order to reach a diagnosis, including a patient's demographic information, medical history, medications, lab testing, and different imaging modalities. Each of these data sources contains a unique piece of each patient's story. With this in mind, there has been a recent push within the medical image analysis community to design models which are able to capture more than one source of medical information [31]. In this section, we will first describe two active areas of research on this front (multi-modal image processing and context-aware image processing), after which we will discuss the challenges associated with combining techniques from these two research areas.

3.1. Multi-modal image processing

The various medical imaging modalities capture unique but complementary information; it is natural, then, to conceive of a model that attempts to leverage these complementary features for clinical applications. In general, multi-modal analysis approaches fall into three categories based on when and how the modalities are combined: pixel-level, feature-level, and classifier-level [31], [32]. Pixel-level fusion involves merging two different imaging modalities into a single volume; this volume may then be used for further image processing or predictive modeling [33]. There is a significant amount of literature dedicated to pixel-level medical image fusion, but much of this work is focused solely on generating the fused modality volume rather than utilizing it for predictive modeling [34]. One study that did focus on predictive

modeling utilized T1-weighted MRI and PET to investigate Alzheimer's disease [35]. In this analysis, researchers used a white matter segmentation map from T1-weighted MRI to remove the white-matter from a PET scan; this non-white matter PET scan was then used to train a discriminate model of Alzheimer's disease and mild cognitive impairment.

Rather than attempting to create a single fused representation of multiple modalities, feature-level fusion simultaneously extracts features from each modality; these multi-modality features may then be used to train a predictive model. An example of this method is illustrated in a study from Sun et al. that used a deep-learning based method to segment brain tumors by simultaneously processing T1-weighted, T2-weighted, and FLAIR MRI [36]. In another study, parotid gland tumors were segmented and classified using a multi-modal "stack" of 2D image slices from T1-weighted, T2-weighted, and diffusion-weighted MRI [37].

By far the most popular multi-modality fusion method in medical imaging, classifier-level fusion independently extracts features from each modality, then uses this set of single-modality features to train a multi-modal classifier. For example, one study used gray matter volume (MRI) and intensity (PET) from 93 regions to train an SVM to identify Alzheimer's disease subclasses [38]. Other studies have used classifier-level fusion approaches to investigate Alzheimer's disease [39], migraine [18], [40], osteoarthritis [41], and prostate cancer [42], [43].

3.2. Patient context: analyzing electronic health records

Patient context for medical image analysis is obtained via electronic health records. This multi-faceted longitudinal health information is stored by clinics after patient care; it includes demographic information (ages, sex, height, etc.), insurance billing codes, procedural codes, lab testing, medications, and any other information associated with a patient's diagnosis and treatment. Recent innovations in electronic record handling and big data analysis methods have enabled many successful studies to take advantage of this rich information source [6]. For example, Zihni et al. design a predictive model of stroke outcome based on demographic and health history EHR features [44]. Many EHR studies have focused on characterizing clinical phenotypes for different diseases. Luong et al. developed a probabilistic phenotyping method for identifying sub-types of chronic kidney disease based on heterogenous clinical data (age/height/weight/sex, lab results, and medications) [45]. In a similar line of research, Lee et al. introduced a clustering method for identifying temporal EHR phenotypes which capture disease progression in cystic fibrosis and Alzheimer's disease [46]. Other research directions, such the phenome-wide association study (PheWAS)

[47] and phenome-disease association study (PheDAS) [48], use hand-crafted EHR phenotypes to investigate phenotypic associations with various diseases.

Building from this foundational analysis, information from EHRs has increasingly been used to enrich medical image processing algorithms. Most models that blend imaging and clinical data perform classification-level information fusion. For example, Chaganti et al. combined International Classification of Disease (ICD) billing codes [49] with imaging biomarkers in a logistic regression model for the classification of diabetes and optic nerve diseases; this analysis found that the combination of EHR and imaging features resulted in a more powerful predictive model than using either data source in isolation [48], [50]. Another study performed a characterization of breast cancer via a joint analysis of CNN-derived histologic image features and genetic features [51]. Other interesting advances have been made towards processing raw diagnostic report text instead of tabular EHR features. Zhang et al. have used such raw text to guide the training of a neural network to predict bladder cancer based on histologic images [52], [53].

3.3. Challenges

Building on the work described in the previous two sections, a major focus of this dissertation deals with designing contextual multi-modal imaging models; in other words, models that process both EHR data and multi-modal imaging together. Despite the obvious potential of this holistic approach to medical image analysis, there are several challenges to grapple with when designing these models. The first hurdle that must be considered is the EHR itself. Data obtained from EHR is known to be inherently noisy; errors can creep in at many point in the recording process, from patient observation to the influence of billing procedures and avoidance of liability [54]. Additionally, EHR data are biased towards sicker populations since sick patients naturally require more medical care than relatively healthy patients [6]. A second related hurdle for contextual multi-modal model design is simply acquiring the data necessary for such a study. Medical image analysis in general is already afflicted by a lack of large datasets, so curating a dataset of sufficient size that contains multiple imaging modalities and complete longitudinal EHR data for all individuals is not a trivial feat.

Another set of challenges arise in the design and training phases of contextual multi-modal image processing. The first of these is simply how to combine the EHR and imaging data. By far, the easiest level of fusion is classification-level; working with a large number of extracted imaging and clinical features, however, may invoke the curse of dimensionality. Additionally, extracting features independently from each imaging and EHR source risks the loss of complex multi-source patterns in the dataset. Yet feature-level fusion can also be challenging due simply to hardware limitations; some graphics processing units

(GPUs) cannot accommodate a neural network, two or more 3D imaging volumes, and EHR data all at once. Here, we may draw inspiration from some innovative patch-based multi-modal imaging studies to overcome these hardware limitations [39].

4. Explainable AI in Medical Image Analysis

As was discussed briefly in Section 2, many machine learning methods produce interpretable inferences, while deep learning methods are "black-boxes": interpreting the reasoning behind a particular prediction is far from trivial. To remedy this, the field of explainable AI (XAI) has blossomed in recent years. While preparing this manuscript, the author identified 19 review papers on the topic of XAI published since just 2018; these reviews included general overviews [55]–[60], method comparisons [61]–[64], user and societal impacts [65]–[68], and specific considerations for medical imaging applications [11], [69]–[71]. Though each review included its own taxonomy for organizing the growing landscape of interpretability techniques, the two most transcendental categories were *post-hoc explanations* and *intrinsically explainable models*. In the remainder of this section, we describe each of these approaches in detail, following a brief note on the distinction between *explanation* and *interpretation*.

4.1. Explanation vs. Interpretation

The exact definitions of "interpretation" and "explanation" in the field of XAI are debated [55]. These words are so closely related and so often used interchangeably that precise definitions may be impossible to obtain (let alone enforce). Consequently, the author believes that a better solution is to simply present explicit definitions for how these intertwined concepts are used within this manuscript. The following definitions are the result of both careful literature consideration and the desire to emphasize each term's unique function.

Explanation: A metric, textual description, or visualization that provides insight into what features a model extracts and/or which of those features contribute the most to a particular prediction.

Explainable: The ability of a model (or an external method acting on a trained model) to generate an explanation for a given prediction.

Interpretation: The act of examining explanations in order to gain an understanding of the decision process made by a model.

Interpretable: A model for which the discriminatory process is understandable by humans.

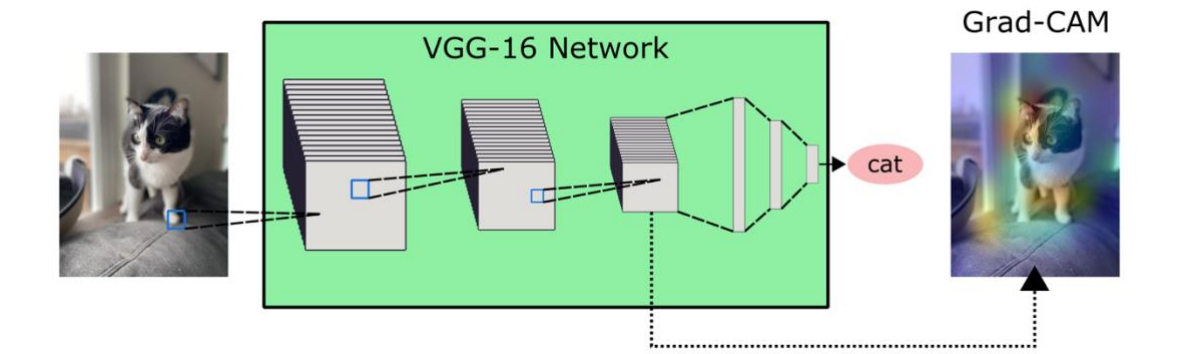


Figure I-3 Example of a post-hoc explanation in a deep learning model

4.2. Post-hoc explainability

Post-hoc explainability techniques aim to elucidate the decision-making processes of naturally unexplainable models. These techniques are advantageous as they can be applied to existing state-of-the-art deep learning architectures. Their explanations typically take the form of saliency maps – heatmaps overlaid on an input image for which the mapped intensity represents the relative importance of each pixel. There are three typical approaches to producing these explanatory saliency maps. The most popular method, activation visualization, involves backpropagating the predicted class activation to find related activations within hidden network layers. *Sensitivity maps* accomplish this by estimating the change in output activation with respect to the input values [72]. *Grad-CAMs* (Figure I-3), on the other hand, propose that the most relevant activations for image classification are found in the last convolutional layers of a CNN, since features extracted at this level are subsequently used to produce the output classification [73]. Yet another method, *Deep Taylor Decomposition*, departs from these sensitivity-focused methods; instead, this method considers the pixel-wise *relevance*, or connection strength, between particular input and output pairs [74], [75]. These techniques are relatively easy to apply, and therefore have been successfully employed in many studies, including those examining brain-computer interfacing [76], intracranial hemorrhage [77], pre-term fetus neurological structure [78], Alzheimer's Disease [79], and COVID-19 [80].

In contrast to activation visualization, perturbation explainability methods produce explanations by tracking changes in generated predictions as an input image is modified in some way. A popular perturbation explainability technique is occlusion sensitivity [81]; this method systematically "grays-out" patches of a sample image and measures how the model's prediction changes as different regions of the image are occluded. This method has recently been used to localize regions with ground glad opacities, vascular thickening, and other biomarkers in a study of COVID-19 chest X-rays [82].

A final type of post-hoc explainability analysis is termed model distillation. In this approach, rather than pull explanations from the opaque ANN itself, the behavior of a trained ANN is distilled into a simpler interpretable model, such as an SVM. Because this distilled model has access to the same information as the ANN and mimics its behavior, it is reasonable to use the distilled model as a proxy for identifying significant features and correlations in the input data. One example of this approach is the local interpretable model-agnostic explanations (LIME) method [83]. LIME first generates the distilled model, then performs an occlusion sensitivity analysis to estimate the importance of different input features. This method has demonstrated successful medical image interpretation capabilities in a study of Parkinson's Disease [84].

4.3. Intrinsic explainability

While post-hoc explainability methods aim to examine a model's decision-making after training, intrinsic explainability methods incorporate explanatory power into the model itself. This can be difficult to achieve with neural networks, as their complex non-linear architectures and legion of parameters are precisely what makes them so successful. Despite this, researchers have recently been exploring two particularly exciting approaches to intrinsically explainable deep learning models: attention mechanisms and prototypical networks. Attention mechanisms allow neural networks to build weighted contextual vectors which influence the relative importance of different inputs in downstream processing [59]. These context vectors may then be examined to visualize which parts of the input the neural network was attending to. This framework has been shown to boost neural network performance while simultaneously providing explanations for each prediction. In medical image processing applications, attention frameworks have also been successfully used for fusing information across imaging modalities and EHR [52]. In contrast, the prototypical network approach integrates interpretability by replicating human learning patterns [85], [86]. After using a CNN to extract imaging features, these networks learn explicit sets of "typical" examples for each class. New samples are then compared to class prototypes, and the predicted class is determined via maximizing prototype similarity.

4.4. Challenges

Despite the urgent need for interpretability in medical image analysis, the application of these described methods faces several challenges. Post-hoc explanations can be illuminating for individual examples, but there is not yet a consensus regarding which post-hoc methods produce the most relevant saliency maps [61], [63], [64], [87]. Intrinsically explainable models offer a potential solution to this issue, but these models tend to suffer from decreased classification performance due to the enforced explanatory constraints

[11], [71]. Furthermore, most explainability methods have been developed for computer vision applications and therefore, have not yet been thoroughly tested in medical imaging applications. This is particularly problematic for complex models that integrate more than one type of input data, including multi-modal imaging and EHR.

5. Test Bed Applications

5.1. Mild traumatic brain injury

Mild traumatic brain injury (mTBI) is a complex syndrome that affects up to 600 per 100,000 individuals in the United States [88], with a particular concentration among military personnel [89], [90]. Approximately half of all affected individuals experience chronic mTBI symptoms that persist long after the acute injury phase [91]. Despite this high prevalence, however, mTBI has proved to be a difficult condition to study; across many reports even the definition of this condition is disputed [92], and there is a large amount of heterogeneity in both symptoms and imaging findings across the mTBI population [93]. Due to the prevalence of mTBI and this long-term adverse symptomology, there is an urgent need for advanced imaging methods that can localize the effects of mTBI in the brain. Currently, the recommended clinical definition of mTBI is a Glasgow Coma Scale score of 13-15 paired with negative structural imaging findings [92]. Despite this definition, however, mTBI has previously been linked to abnormalities of the cortical surface [94] and disrupted white-matter pathways [95]. This makes mTBI a particularly good candidate for contextual multi-modal image analysis, as we expect pathological signals to be subtle patterns consisting of multiple neurological and sensory systems.

5.2. Developmental disabilities

It is estimated that between 2009 and 2017, the prevalence of developmental disabilities in the United States for children aged 3-17 was 16.93% [96]. Developmental disabilities encompass a group of conditions related to impairments in learning, behaviors, or physical growth. For example, Down syndrome is a genetic disorder characterized by the presence of an extra copy of chromosome 21 [97]. Patients with this condition are known to carry a heavy burden of comorbid conditions, including hypothyroidism, gastrointestinal disorders, congenital heart disease, and sleep apnea [98]. Across the range of developmental disabilities, patients tend to have a similarly wide array of comorbidities, requiring supportive care throughout their lifetimes [96]. In recent years, EHRs have been leveraged to study clinical patterns of autism spectrum disorder [29], [99]. Studying the EHRs and combined neuroanatomy of patients with this and other developmental disabilities could improve our understanding of the timing and intensity of these comorbid

patterns, enable the characterization of clusters of co-occurring disorders, and reveal currently unmet clinical needs of this community.

5.3. Mild cognitive impairment

A large proportion of medical image analysis literature is devoted to studying Alzheimer's disease and other dementias [16], [39], [108]–[117], [100], [118]–[122], [101]–[107]. This aggressive investigation is in large part due to the availability of three publicly available dementia-focused data repositories: the Alzheimer's Disease Neuroimaging Initiative (ADNI) [123] and the Open Access Series of Imaging Studies (OASIS) datasets OASIS-2 [124] and OASIS-3 [125]. Mild cognitive impairment (MCI) is an early phase of cognitive decline that precedes dementia in aging adults [126]. Most imaging studies that involve MCI only examine it as a secondary condition (with the primary focus of characterizing Alzheimer's disease) or use MCI as the baseline condition against which Alzheimer's Disease is compared. However, as more therapeutic interventions become available, the need for early detection of MCI in aging adults becomes more urgent; generally, earlier introduction of preventative treatments for MCI and dementia increase their effectiveness [127].

6. Contributed Work

The widespread use of deep learning has revolutionized the field of medical image analysis. Increasing amounts of available clinical and imaging data allow researchers to train more accurate deep learning models that encompass the whole person. Yet simultaneously, interpretability methods for these complex neural networks are lagging behind; we cannot hope to translate research models into clinical use until their predictions are accompanied by interpretable explanations grounded in anatomy. The work outlined in this dissertation lies at the intersection of these issues. We first focus on innovations in interpretability for traditional machine learning (**Contribution 1**), including developing multi-contrast MRI models, innovations in EHR analysis, and the introduction of explainability techniques into both. We next translate this work in interpretability to deep learning models of MRI and EHR (**Contribution 2**). Finally, these efforts come together to form a framework for interpretable multi-modal analysis of both MRI and EHR (**Contribution 3**). These efforts are essential for the field of medical image analysis, both in moving closer to individualized medicine and facilitating model trustworthiness for clinical translation.

6.1. Contribution 1: Interpretable Machine Learning with MRI and EHR

- We trained an exploratory multi-modal machine learning model of mild traumatic brain injury that demonstrates the utility of structural and diffusion-weighted MRI modalities in discriminating mild traumatic brain injury patients from controls.
- We created an interpretable joint model of T1-weighted MRI and diffusion tensor imaging able to identify white matter connectivity and cortical surface shape changes in mild traumatic brain injury patients relative to controls.
- We deployed pyPheWAS, an open-source python toolkit for conducting phenome-disease association studies.
- We designed pyPheWAS Explorer, an interactive visualization built on top of pyPheWAS that enables users to design, run, and visualize PheDAS models in real time. This tool was demonstrated in a case study of attention-deficit hyperactivity disorder.
- We investigated Down syndrome phenotype associations using pyPheWAS and explainable machine learning that revealed several significant phenotypic associations with Down syndrome generally and heart surgery in Down syndrome patients with co-morbid congenital heart disease.

6.2. Contribution 2: Interpretable Deep Learning with MRI and EHR

- We characterized the significant impact that batch size has on the interpretability of deep autoencoder embeddings of medical data.
- We extended hard case mining to unsupervised neural network training and demonstrated that this simple, computationally efficient technique may improve embedding interpretability and accelerate network convergence for MRI and EHR autoencoders.

6.3. Contribution 3: Interpretable Multi-Modal Modeling for MRI and EHR

- We developed a novel framework for interpretable joint analysis of longitudinal EHR subtypes and region-specific MRI-derived brain characteristics.

6.4. Outline of Dissertation

The remainder of this document proceeds as follows. In **Chapter II**, we develop a preliminary model of multi-modal MRI in the context of classifying mTBI. We establish T1-weighted MRI and diffusion tensor imaging as promising modalities for the study of this neurological disorder. Building on this effort, **Chapter III** presents an *interpretable* joint model of T1-weighted MRI and diffusion tensor imaging, again in the context of mTBI. From this study, we identify changes in white matter connectivity and cortical surface structure associated with mTBI and its chronic symptoms. Our focus then shifts from medical image analysis to EHR processing. **Chapter IV** describes pyPheWAS, a command line python-based toolkit developed for phenome-disease association studies (PheDAS). This toolkit allows users to identify phenotypic associations with disease based on diagnostic and procedural EHR codes. Following this, we present pyPheWAS Explorer (**Chapter V**), an interactive visualization of the pyPheWAS analysis pipeline that aims to make PheDAS models more transparent, interpretable, and accessible. **Chapter VI** then describes an investigation into the EHRs of Down syndrome patients via the pyPheWAS toolkit and interpretable machine learning techniques. This study identifies several significant associations between EHR phenotypes and Down syndrome generally, in addition to specific phenotypes that are associated with longitudinal surgery risk in Down syndrome patients with co-morbid congenital heart disease. In **Chapter VII**, we turn our attention to autoencoders, a deep learning architecture, and their use as an interpretable manifold embedding method for medical data. We investigate the substantial effect that batch size, a training hyperparameter, has on the interpretability of the trained autoencoder embedding space. Based on this work, **Chapter VIII** then presents unsupervised hard case mining, a cost-efficient optimization technique for training medical autoencoders; we demonstrate in models of MRI and EHR that this method may improve the interpretability of encoder embeddings and accelerate training convergence. This work all culminates in the development of a novel interpretable framework for joint multi-modal analysis of EHR and MRI in a study of autism spectrum disorder (**Chapter IX**). Finally, **Chapter X** draws together concluding thoughts and outlines future research directions in interpretable artificial intelligence models for multi-modal models of MRI and EHR.

Chapter II

MRI Correlates of Chronic Symptoms in Mild Traumatic Brain Injury

1. Overview

Some veterans with a history of mild traumatic brain injury (mTBI) have reported experiencing auditory and visual dysfunction that persist beyond the acute phase of the incident. The etiology behind these symptoms is difficult to characterize, since mTBI is defined by negative imaging findings on current clinical imaging. There are several competing hypotheses that could explain functional deficits; one example is shear injury, which may manifest in diffusion-weighted magnetic resonance (MR) imaging (DWI). Herein, we explore this alternative hypothesis in a pilot study of multi-parametric MR imaging. Briefly, we consider a cohort of 8 mTBI patients relative to 22 control subjects using structural T1-weighted imaging (T1w) and connectivity with DWI. 1,344 metrics were extracted per subject from whole brain regions and connectivity patterns in sensory networks. For each set of imaging-derived metrics, the control subject metrics were embedded in a low-dimensional manifold with principal component analysis, after which mTBI subject metrics were projected into the same space. These manifolds were employed to train support vector machines (SVM) to classify subjects as controls or mTBI. Two of the SVMs trained achieved near-perfect accuracy averaged across four-fold cross-validation. Additionally, we present correlations between manifold dimensions and 22 self-reported mTBI symptoms and find that five principal components from the manifolds (one component from the T1w manifold and four components from the DWI manifold) are significantly correlated with symptoms ($p<0.05$, uncorrected). The novelty of this chapter is that the DWI and T1w imaging metrics seem to contain information critical for distinguishing between mTBI and control subjects. This chapter presents an analysis of the pilot phase of data collection of the Quantitative Evaluation of Visual and Auditory Dysfunction and Multi-Sensory Integration in Complex TBI Patients study and defines specific hypotheses to be tested in the full sample.

2. Introduction

Mild TBI (mTBI) is a difficult condition to research; across many studies, even the definition of mTBI injury is disputed [92]. This is unsurprising, however, since across a population of mTBI subjects there is also often a large amount of heterogeneity in both symptoms and imaging findings [93]. Military veterans are particularly susceptible to mTBI due to their frequent proximity to blasts [90]. Many such veterans report experiencing chronic mTBI symptoms, but current clinical magnetic resonance (MR) imaging and computed tomography do not detect any TBI features. The current recommended clinical definition of

mTBI, in fact, is a Glasgow Coma Scale score of 13-15 paired with negative imaging findings [92]. Despite this, there are several possible explanations for the chronic dysfunction these veterans are experiencing. Shear injuries, another possible explanation, may manifest on diffusion-weighted MR images (DWI), where they may appear as changes in the geometry, trajectory, and volume of white matter pathways [95]. Based on this alternative hypothesis, in this chapter we perform a pilot study focused on distinguishing mTBI subjects from healthy controls via multi-parametric MR imaging.

To this end, a set of 1,344 imaging metrics are extracted from DWI and T1-weighted (T1w) MR imaging; these metrics are processed using Principal Component Analysis (PCA) to derive a lower-dimensional representation of the cohort. The PCA representation of each metric set is employed to train a nonlinear mTBI vs control classifier. Additionally, the connection between imaging metrics and patient symptoms is explored via computing correlations between the PCA representation and self-reported mTBI patient symptoms scores.

3. Methods

3.1. Data collection and preprocessing

T1w and DWI scans were acquired for 30 subjects, of whom 22 were controls (no history of TBI nor auditory and visual problems) and 8 were subjects with a history of mTBI (prior mTBI diagnosis confirmed via Electronic Medical Record with a Glasgow Coma Score in the range 13-15). The T1w scans were segmented into 132 brain regions as defined by the BrainCOLOR protocol via multi-atlas labeling as described in [128]. These labels were then registered to the DWI volume space for use in deriving region-based imaging metrics. Both 32 and 64 shell DWI scans were acquired, which were concatenated and corrected for eddy-current distortions and patient movement according to [129].

mTBI subjects filled out a questionnaire including the Neurobehavioral Symptom Inventory, which covers 22 TBI symptoms: dizziness, loss of balance, poor coordination, headaches, nausea, vision problems, light-sensitivity, hearing difficulty, noise sensitivity, numbness, taste or smell changes, appetite changes, poor concentration, forgetfulness, difficulty making decisions, slowed thinking, fatigue, difficulty sleeping, anxiety, feeling depressed, irritability, and frustration [130]. These self-reported symptoms were ranked on a scale of 1 to 5 with regard to its impact on their life since the injury, where 1 was unaffected and 5 was a significant impact on daily life.

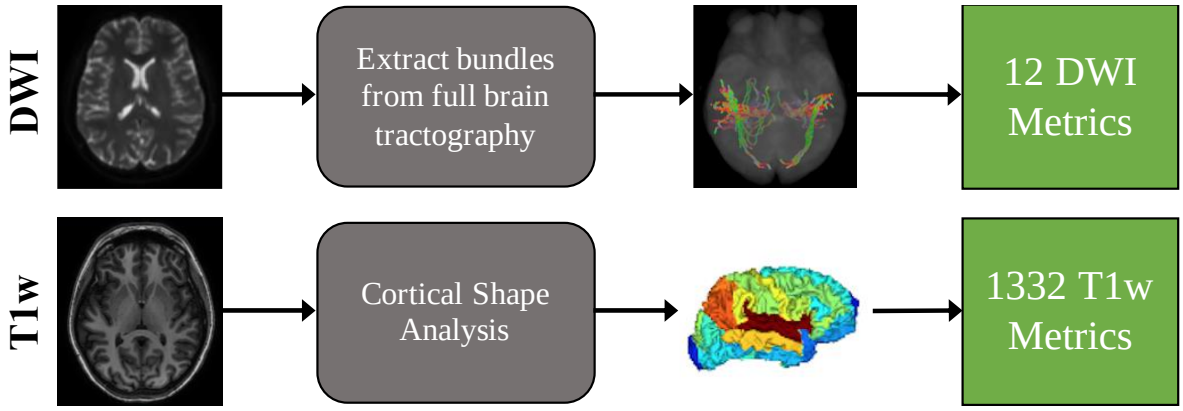


Figure II-1 An overview of imaging metric generation is presented. Full-brain tractography is performed on the preprocessed DWI volume, and four streamline bundles are extracted using the BrainCOLOR labels. The number of streamlines, bundle length, and bundle volume are calculated for each bundle, resulting in 12 connectivity metrics per subject. Cortical Shape Analysis is performed on the T1w volume; for each cortical surface region, curvature, shape index, sulcal depth, thickness, shape complexity index, and local gyrification index were calculated both along the region's sulci and averaged across the entire region, yielding 1,332 surface metrics.

3.2. Imaging metric extraction

An overview of the derivation of imaging metrics for each MR modality is shown in Figure II-1. The tractography pipeline was implemented using the MRTrix3 package [131]. The DWI volume was segmented into five tissue-type regions; anatomically-constrained full-brain tractography was then performed [132], and the resulting 10 million streamlines were sifted down to 1 million anatomically-probable streamlines [133]. The DWI-registered BrainCOLOR labels were used to extract four distinct streamline bundles that are associated with the auditory and visual sensory pathways (Figure II-2). In the right hemisphere, one bundle connects the thalamus to the superior temporal gyrus, and a second bundle connects the superior temporal gyrus to the calcarine cortex. The third and fourth bundles connect the same structures in the left hemisphere. For each bundle, three metrics were recorded: number of streamlines, average streamline length, and bundle volume, resulting in 12 total connectivity metrics per subject.

Structural metrics were acquired by first reconstructing the cortical surface and segmenting it into 111 regions via the MaCRUISE pipeline [134]. A cortical shape analysis was then performed on the cortical surface as described in [135]. For each surface region, the mean curvature, shape index, sulcal depth,

cortical thickness, shape complexity index, and local gyrification index were calculated both along the sulcal fundic region [135] and averaged across the region as a whole. This resulted in 1,332 structural imaging metrics.

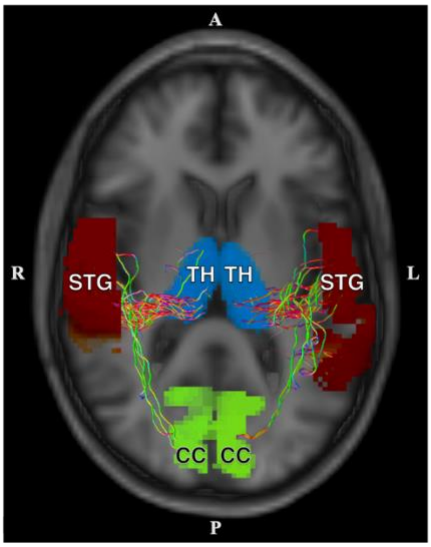


Figure II-2 An illustration of the four streamline bundles with the BrainCOLOR regions they connect. In both the right and the left hemispheres, bundles connect the thalamus (TH) to the superior temporal gyrus (STG) and the superior temporal gyrus to the calcarine cortex (CC).

3.3. Imaging metric analysis

Figure II-3 outlines the analysis of the 1,344 imaging-derived metrics described in section 2.2. The two types of imaging metrics were each analyzed individually to evaluate the effect that each had on the final subject-wise classification. Within each set, the metrics were normalized by calculating the z-score with respect to the mean and standard deviation of the control subjects. PCA was then applied to the z-scores of the healthy controls, producing two individual lower-dimensional PCA spaces (one each for DWI and T1w), which the z-scores of the mTBI subjects were projected into.

Next, each metric set's ability to distinguish between mTBI and control subjects was assessed individually. To this end, an SVM classifier [136] was trained on the PCA space of each metric set combined with subject age. For all SVMs trained, mTBI was defined as the positive class, and control was defined as the negative class. SVMs were trained and validated using four-fold cross validation with a radial basis function kernel. The box constraint and kernel scale hyperparameters were optimized on the training set at each fold using five-fold Bayesian optimization. The PCA spaces of each metric set were iteratively swept, so that a single principal component was added to the SVM training data at each iteration, starting with the first principal component. This sweeping procedure was used to determine how the addition of each component impacted the performance of the SVM classifier.

Once the entire PCA space was swept, the number of components, C_O, which produced the most optimal classifier was determined for each metric set. For the purposes of this analysis, "optimal" was defined as the classifier which maximized validation recall (averaged across the four cross-validation folds) based on the fewest number of principal components. Finally, the C_O components from the DWI and T1w metric sets' PCA spaces were combined, and the iterative SVM training procedure was repeated to analyze how the metric sets might work together to distinguish between mTBI and control subjects. It is important to note that in this combined sweeping, a principal component from both metric sets was added at each iteration, starting with the first principal components from each PCA space. Similar to the individual analysis, the optimal number of components for the combined SVM classifier was determined after the entire set of C_O components from the DWI and T1w metric sets were swept. Additionally, the Spearman

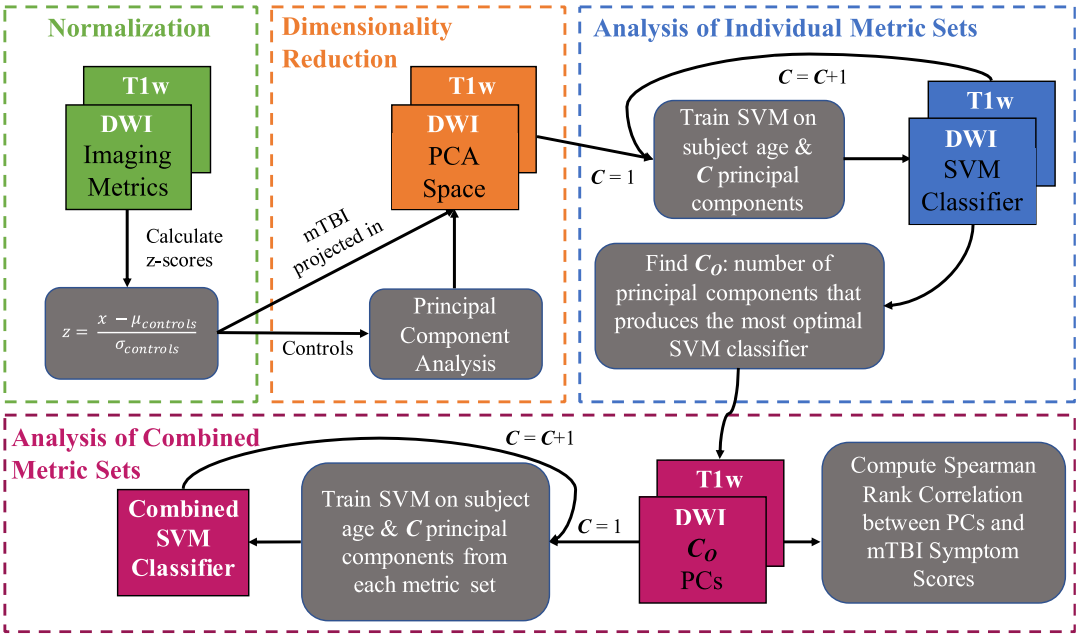


Figure II-3 A schematic overview of the imaging metric analysis. First, the imaging metrics are normalized by converting the raw imaging metrics to z-scores using the mean μcontrols and standard deviation σcontrols of the control subjects. PCA is performed using the z-scores of the control subjects, resulting in two lower-dimensional PCA spaces (one for each metric set), which the mTBI subjects' z-scores are projected into. Next, to analyze the metric sets individually, the PCA components of a single set and the subjects' ages are used to train a four-fold cross-validated SVM to classify subjects as controls or mTBI. Starting with the first principal component, the entire PCA space of each metric set is swept, adding a single component to the SVM at each iteration. After all components have been swept, CO, the number of principal components that produces the most optimal classifier, can be determined for each metric set based on the validation set performance (averaged across the four cross-validation folds). Finally, to analyze the metric sets together, the iterative SVM training process is repeated on the combined set of CO components from each metric set. In this step, the process starts with the first principal component from each metric set then adds an additional component from each metric set to the classifier at each iteration.

rank correlations were calculated between the 22 mTBI symptoms scores and these C_O components from the Combined classifier.

4. Results

4.1. SVM classifier performance

Figure II-4 shows SVM classifier performance as the PCA components are swept for both the individual metric sets and the combined set. This performance is represented by classification accuracy, recall, and specificity, all averaged across the four cross-validation folds. C_O is denoted for each metric set by a vertical bar; for DWI C_O = 11, for T1w C_O =13, and for the combined set C_O = 11. Note that for the combined set, C_O represents the number of components per metric set (i.e. C_O = 11 means that the optimal SVM for the combined set included 11 DWI principal components and 11 T1w principal components). SVM classifier performance at the operating point C_O for each metric set is shown in Table II-1.

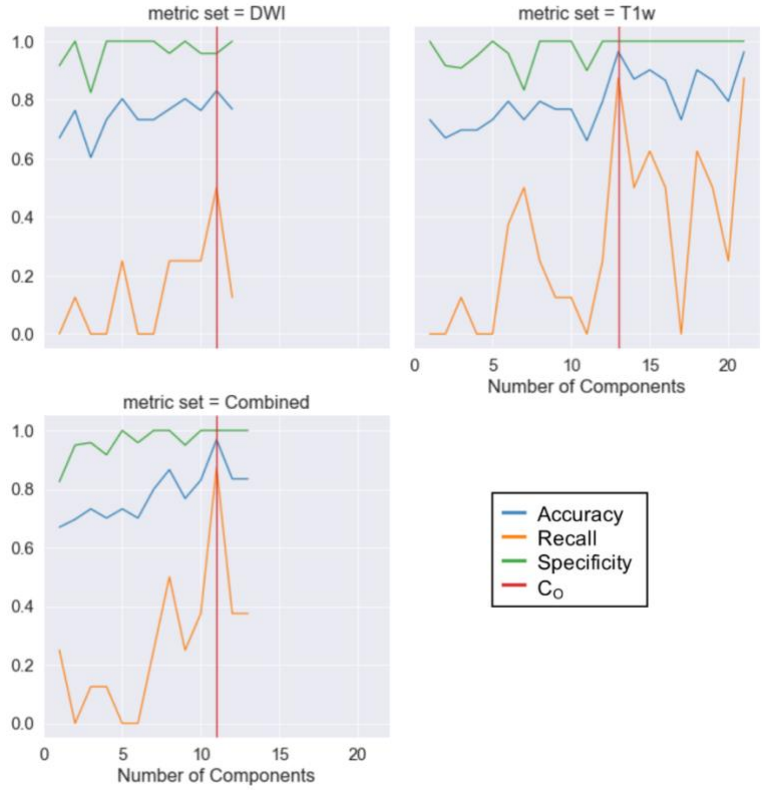


Figure II-4 Performance averaged across the 4 folds of individual metric set classifiers and combined metric set classifier as PCA components are added. The optimal operating point is displayed as a red vertical line. The top two plots show that the classifiers trained on the DWI and T1w metric sets individually are able to distinguish between the two classes. The plot in the bottom left shows that the SVM trained on DWI and T1w metric sets combined can also distinguish between the two classes, but not better than the SVM trained only on the T1w metric set.

For both the DWI and T1w metric sets the sweep of the PCA spaces shows that given enough components, the SVM is able to distinguish between mTBI and control subjects, with T1w producing a near perfect classifier. When DWI and T1w were combined, the SVM was again able to distinguish between mTBI and control subjects with performance similar to T1w individually.

4.2. Symptom score correlations

Out of the 22 symptom scores, 11 DWI principal components, and 11 T1w principal components, five statistically significant ($p < 0.05$, uncorrected) Spearman rank correlations were found. Table II-2 summarizes these correlations, showing that four of the five significant symptom correlations are related to the DWI metric set, and two of the five symptoms listed are appetite change. Due to limited sample size and the large number of correlations, these individual tests should be interpreted as exploratory; when false discovery rate correction was applied across all correlations, no tests surpassed the corrected 0.05 threshold.

Table II-1 Classifier performance for the optimal operating point of each metric set averaged across the four cross-validation folds (values in parentheses are standard deviations)

Metric Set	Co	Accuracy	Recall	Specificity
DWI	11	0.830 (0.067)	0.500 (0.354)	0.958 (0.072)
T1w	13	0.964 (0.062)	0.875 (0.217)	1.000 (0.000)
Combined	11	0.968 (0.054)	0.875 (0.217)	1.000 (0.000)

Table II-2 Significant ($p < 0.05$, uncorrected) correlations found between mTBI symptoms and the PCA components used to train the optimal Combined SVM classifier

Metric Set	Symptom	Principal Component	Correlation Coefficient	p-value
T1w	Appetite Change	5	0.7910	0.0196
DWI	Appetite Change	5	-0.8456	0.0178
	Poor Concentration	5	-0.8648	0.0357
	Feeling Depressed	5	-0.8225	0.0083
	Frustration	5	-0.8225	0.0167

5. Discussion

SVM classifiers trained on the DWI, T1w, and Combined PCA components were all able to distinguish between mTBI and control subjects. The individual T1w and Combined classifiers were both able to achieve near perfect accuracy in this task. It is interesting that the optimal Combined classifier achieved this near-perfect performance using only 11 T1w components, whereas the individual T1w classifier required 13 components to achieve its optimal performance; however, the performance of the two classifiers is too similar to tell whether the T1w or Combined classifier has any true advantage over the other. A second interesting observation is that the DWI metric set produced more significant symptom correlations than the T1w metric set. This suggests that despite its inferior performance in classification, the DWI metric set may still contain some information relevant to mTBI.

6. Conclusion

The key finding of this chapter was that the DWI and T1w imaging metrics seem to contain information critical for distinguishing between mTBI and control subjects. For all metric sets, the PCA dimensionality reduction step was performed using data only from controls, yet both the T1w and Combined classifiers achieved near-perfect four-fold cross-validation accuracy. The SVM classification performance indicates that most of the distinguishing information is in the T1w metrics, but the symptom correlations suggest that the DWI metrics may yet prove useful.

In summary, a novel combination of MRI modalities and imaging-derived metrics are presented in an effort to begin characterizing mTBI in MR imaging. Through PCA and SVM, these metrics were leveraged to produce two near-perfect classifiers for a condition that is currently identified by the absence of imaging findings. We conclude, therefore, that the methods described in this chapter show promise towards characterizing mTBI via MR imaging, but a deeper analysis and larger cohort are needed to clearly determine which individual imaging metrics are contributing the most to subject classification and symptom correlations. As more data is acquired for this study, we intend to improve the image extraction methods for DWI by including more streamline bundles and extracting more metrics from each bundle (i.e. fractional anisotropy along the bundle, connectivity profile, etc.). Additionally, we plan to deepen the classification and symptom correlation analyses by moving from analyzing whole metric sets to analyzing each metric individually to pinpoint precisely which metrics provide the distinguishing mTBI information.

Chapter III

Joint Analysis of Structural Connectivity and Cortical Surface Features: Correlated with Mild Traumatic Brain Injury

1. Overview

Mild traumatic brain injury (mTBI) is a complex syndrome that affects up to 600 per 100,000 individuals, with a particular concentration among military personnel. About half of all mTBI patients experience a diverse array of chronic symptoms which persist long after the acute injury. Hence, there is an urgent need for better understanding of the white matter and gray matter pathologies associated with mTBI to map which specific brain systems are impacted and identify courses of intervention. Previous works have linked mTBI to disruptions in white matter pathways and cortical surface abnormalities. Herein, we examine these hypothesized links in an exploratory study of joint structural connectivity and cortical surface changes associated with mTBI and its chronic symptoms. Briefly, we consider a cohort of 12 mTBI and 26 control subjects. A set of 588 cortical surface metrics and 4,753 structural connectivity metrics were extracted from cortical surface regions and diffusion weighted magnetic resonance imaging in each subject. Principal component analysis (PCA) was used to reduce the dimensionality of each metric set. We then applied independent component analysis (ICA) both to each PCA space individually and together in a joint ICA approach. We identified a stable independent component across the connectivity-only and joint ICAs which presented significant group differences in subject loadings ($p<0.05$, corrected). Additionally, we found that two mTBI symptoms, slowed thinking and forgetfulness, were significantly correlated ($p<0.05$, corrected) with mTBI subject loadings in a surface-only ICA. These surface-only loadings captured an increase in bilateral cortical thickness.

2. Introduction

Mild traumatic brain injury is a disruption of normal brain function caused by any injury to the head. It is estimated that in North America, mTBI has an incidence rate of more than 600 per 100,000 inhabitants [88]. Military personnel and veterans are particularly burdened by this disorder; 11%-23% of soldiers are expected to experience a traumatic brain injury while deployed, with the majority of cases classified as mild [89]. Additionally, approximately half of all mTBI patients are left with chronic symptoms that persist long after the injury's acute phase [91]. Due to the prevalence of mTBI and this long-term adverse symptomology, there is an urgent need for advanced imaging methods that can localize mTBI effects in the brain.

mTBI has previously been linked to abnormalities of the cortical surface [94] and disrupted white matter pathways [95]. One hypothesis is that mTBI may be classified as a "disconnection syndrome" due to the prevalence of these white matter disconnections [137]. In a previous work, we demonstrated the potential of combining cortical surface and sensory white matter pathway features in a discriminatory model of mTBI [138]. In this article, we build on that work by investigating a joint analysis of magnetic resonance imaging (MRI) derived cortical surface and structural connectivity features in mTBI and control subjects. This exploratory analysis aims to identify joint cortical surface and structural connectivity changes associated with mTBI and related symptoms, which may be tested in a larger sample.

Briefly, this work involved first extracting a) a set of 588 cortical shape metrics and b) a set 4,753 structural connectivity metrics from 98 cortical surface regions for each subject. The dimensionality of each metric set was reduced via principal component analysis. Independent component analysis was then applied to the PCA spaces of the cortical surface metrics, structural connectivity metrics, and a joint metric set (a concatenation of cortical surface and structural connectivity PCA spaces). Subjects' independent component (IC) loadings were compared across the mTBI and control groups to assess group differences. The correlation between mTBI subjects' IC loadings and self-reported symptom severity scores was also examined.

3. Methods

3.1. Imaging and symptom data

This exploratory study considered a cohort of 38 subjects, of whom 12 were mTBI subjects (previous mTBI diagnosis confirmed via Electronic Medical Record with a Glasgow Coma Score in the range 13-15) and 26 were controls with no history of mTBI or audiovisual problems. All subjects had both T1-weighted MRI (T1w) and diffusion weighted MRI (DWI) scans acquired during a single session at the Vanderbilt University Institute of Imaging Science. Three DWI volumes were acquired for each subject: a 32-direction b=1000 s/mm^2, a 64-direction b=2000 s/mm^2, and a corresponding b=0 s/mm^2 volume. These 32 and 64 shell DWI volumes were concatenated and corrected for eddy current distortions and patient movement following the protocol in [129].

For all mTBI subjects, chronic mTBI symptoms were assessed via the Neurobehavioral Symptom Inventory [130], a questionnaire that tracks 22 TBI symptoms: dizziness, loss of balance, poor coordination, headaches, nausea, vision problems, light sensitivity, hearing difficulty, noise sensitivity, numbness, taste or smell changes, appetite changes, poor concentration, forgetfulness, difficulty making decisions, slowed thinking, fatigue, difficulty sleeping, anxiety, feeling depressed, irritability, and frustration. The severity of

each symptom was ranked by the subject on a scale of 1 to 5, where 1 was unaffected and 5 was a significant impact on daily life.

3.2. Metric generation

The cortical surface and structural connectivity imaging metric generation processes are shown in Figure III-1. First, the T1w volume was segmented into 132 BrainColor regions via multi-atlas segmentation [128], and MaCRUISE [139] was used to reconstruct cortical surfaces from the volumetric segmentation. To define boundaries of regions of interest (ROI) on the cortical surfaces, the cortical surfaces were mapped to a unit sphere and then rigidly aligned to the adult template created by [140]. For each individual cortical surface, the cortical ROIs were determined via spherical convolutional networks [141]. The resulting labels are a subset of BrainColor, which contains 98 regions out of the original 132. A cortical surface analysis was then performed, yielding six shape metrics averaged over each cortical surface region, including mean curvature, shape index [142], sulcal depth [135], cortical thickness [139], shape complexity index [143], and local gyrification index [144], [145]. This resulted in a total of 588 surface metrics.

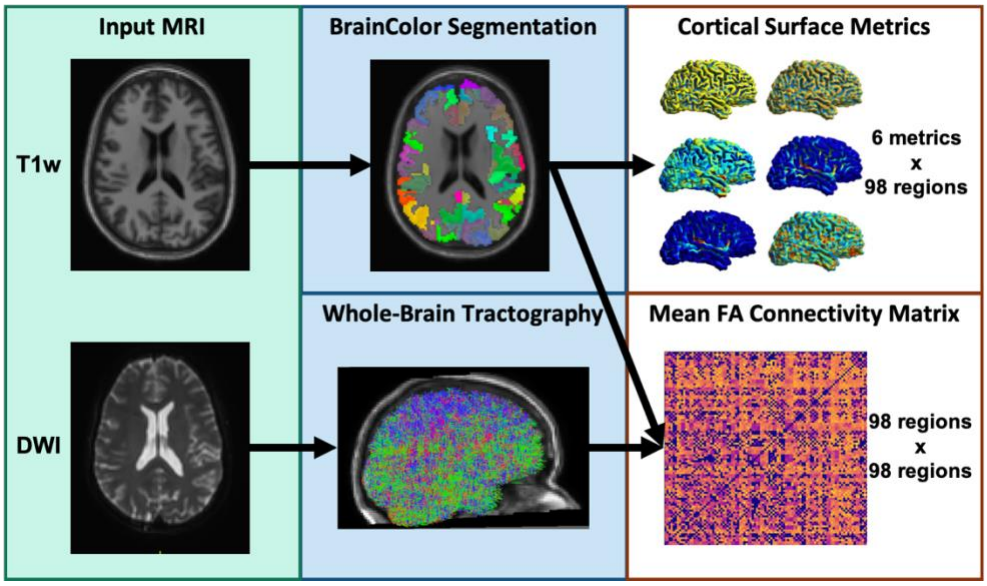


Figure III-1 Surface and connectivity metric generation. The T1w volume is segmented into 132 BrainColor regions, out of which 98 cortical surface regions are kept. A cortical shape analysis is performed on the 98 regions, yielding 6 shape metrics per region: mean curvature, shape index, sulcal depth, cortical thickness, shape complexity index, and local gyrification index. Whole brain tractography is performed on the DWI volume. This tractogram is used to construct a connectivity matrix for the 98 surface regions, where connection strength is equivalent to mean fractional anisotropy (FA) along streamlines connecting each pair of regions.

To derive the structural connectivity metrics, each subject's 98-region BrainColor segmentation was registered to their DWI volume, so that the same cortical surface regions were used for both the surface and connectivity metrics. Next, a whole-brain tractogram was generated for each subject. The response function for tracking was estimated via the iterative Tournier algorithm [146], after which the fiber orientation distribution was calculated using spherical deconvolution. The iFOD2 algorithm [147] was then used to perform probabilistic tractography and construct a one million streamline whole-brain tractogram seeded at random within a whole-brain mask. Finally, a connectivity matrix was generated for each subject, where the nodes were the 98 cortical surface regions, and the edges were defined as mean fractional anisotropy (FA) along the streamlines connecting each pair of regions. This mean FA linking pairs of regions was used as our structural connectivity metric. Self-connections were excluded, and streamline direction was ignored, yielding 4,753 metrics. All tractography and connectivity matrix operations were performed via the MRTrix3 package [148].

3.3. Metric analysis

An overview of the imaging metric analysis is shown in Figure III-2. Briefly, this analysis pipeline includes metric normalization and dimensionality reduction via principal component analysis (PCA) as preprocessing steps before applying independent component analysis (ICA). ICA can be sensitive to high variance, so PCA was applied prior to the ICA step in order to reduce variance in the dataset. Additionally, reconstruction ICA was used; this variant on traditional ICA optimizes on a soft reconstruction constraint [149]. Reconstruction ICA is a desirable method for this application, as it was designed to extract stable features which are not as sensitive to variance in the underlying data as traditional ICA. The following description includes more specific details of our method design.

The 588 surface and 4,753 connectivity metrics were normalized via conversion to z-score representation; z-scores are calculated for each individual metric by subtracting the group's mean and dividing by the group's standard deviation. Next, the dimensionality of each metric set was reduced via PCA. This reduction revealed that principal components (PCs) beyond the top 8 in both the surface and connectivity PCA spaces represented negligible explained variance in the data. Based on this and the need to balance the metric sets in the joint analysis, the PCs were narrowed down to the top 8 for both the surface and connectivity PCA spaces. A joint PCA space was then created by concatenating the surface and connectivity PCs, yielding a 16-component joint space. Finally, the reconstruction ICA was applied to the three PCA spaces. To fully explore all possible ICA features, X, the number of independent components (ICs), was varied from 2 to 7 for all three PCA spaces. To differentiate between these, we will refer to each individual ICA in the form of "ICA type – X", where possible types are *S* (surface-only ICA), *C*

(connectivity-only ICA), and *J* (joint ICA); for example, S-6 refers to surface-only ICA with 6 ICs. Note that 1 and 8 ICs were excluded because singular reduction (1 IC) and no reduction (8 ICs) were considered uninteresting edge cases.

Each ICA produces two important outputs: ICs and subject loadings. ICs are the set of independent features learned from the PCA space; they may be back-projected into the original data space to investigate relationships between brain regions. These ICs are related to study subjects according to each subject's set of IC loadings. These loadings are a set of weights (one per IC) that correspond to how strongly each IC is represented in an individual subject's input signal. So, while ICs describe overall patterns in the data, subject loadings describe variations of those patterns in individuals. The following sections explain how these two outputs are investigated.

3.4. Visualizing independent components

The many sets of ICs derived in section 2.3 would be largely useless without a means for interpretation. To facilitate visual interpretation, all ICs were back-projected into the original cortical surface and/or structural connectivity spaces and scaled to the range [0,1]. All ICs were visualized using the 98-region

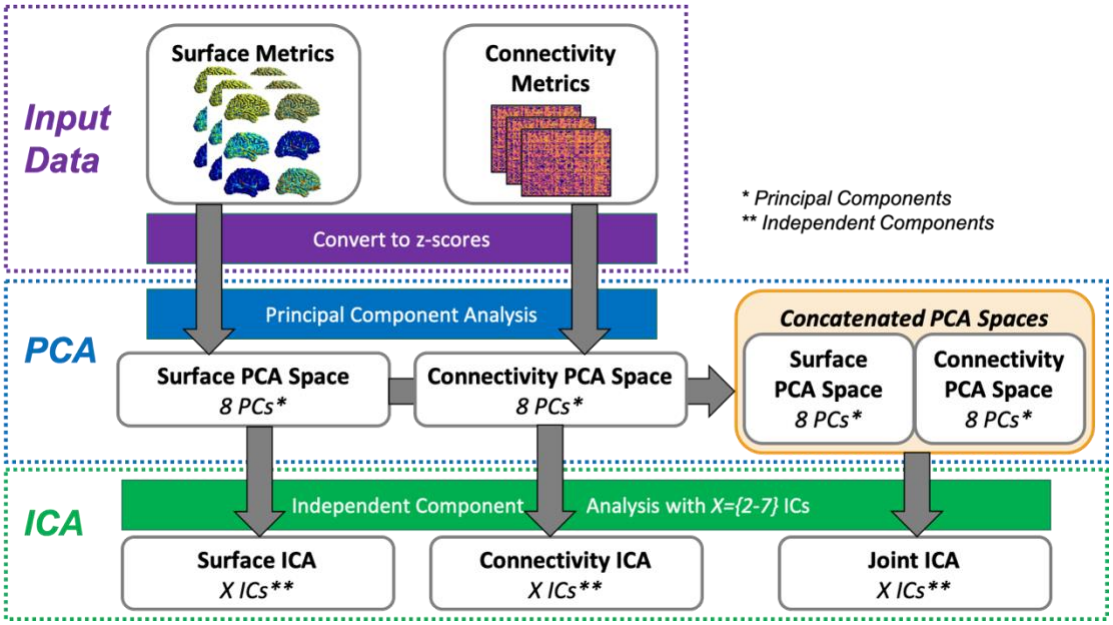


Figure III-2 Metric analysis pipeline. All metrics are first normalized by converting the raw data to z-scores. PCA is then performed separately on the surface and connectivity metrics; the dimensionality of both metric sets is reduced to the 8 most principal components from their respective PCA spaces. These two PCA spaces are further reduced via ICA to X ICs. The value of X is swept across all possible values, from 2 to 7 ICs. To perform a joint analysis of the surface and connectivity metrics, their 8-component PCA spaces are concatenated, and ICA is again performed on this joint data set for X={2-7} ICs.

cortical surface BrainColor segmentation of a single representative T1w volume. Since the surface data space contained only six metrics per region, surface ICs were visualized on six individual three-dimensional surface renderings, one for each metric. In these renderings, region-wise IC coefficients were mapped to each region's color.

Due to the high dimensionality of the structural connectivity data space, the connectivity ICs required a more specialized visualization approach. In this visualization, the entire connectome was presented on a single T1w surface, with region color and transparency denoting region-wise and overall connectivity, respectively. To illustrate, consider a single IC back-projected to a connectivity matrix *C*. First, all self-connections are set to 1 (fully connected). *C* is then converted to a dissimilarity matrix *D* by subtracting all elements from 1. Nonmetric multidimensional scaling with Kruskal's normalized stress1 criterion [150] is applied to *D,* so that each region is mapped to a two-dimensional space *D'* in which the Euclidian distance between any two points approximates a monotonic transformation of their corresponding dissimilarity in *D*. The two dimensions of *D'* are then used as the a* (green-red) and b* (blue-yellow) dimensions in the CIELAB colorimetric system [151]. The last dimension, *L (lightness), is calculated by summing a region's edges in *C* as a percentage of the summed edges of the most connected region. Thus, *C* is converted to a region-wise CIELAB color map. A three-dimensional T1w cortical surface is then generated in which each region is colored according to this color map, and each region's transparency is scaled according to its L* dimension. In this way, the IC is visualized on a single surface, such that similarly colored regions are more connected to each other, and more opaque regions are more connected overall.

4. Results

4.1. Group differences

A primary aim of this analysis was to identify cortical surface and structural connectivity changes associated with mTBI. To this end, we investigated group differences by comparing subject IC loadings between the mTBI and control groups for each IC within each ICA. This comparison was accomplished via a Wilcoxon rank-sum test, with Bonferroni multiple comparisons correction applied to all tests within each ICA (e.g. Bonferroni was applied for 6 comparisons in J-6, and for 7 comparisons in J-7). Through this method, C-4, C-5, C-6, C-7, J-4, J-5, J-6, and J-7 were all found to have an IC that presented a statistically significant group difference in subject loadings ($p < 0.01$, corrected). These 8 ICs were compared using a Pearson correlation and all were found to be highly correlated ($p << 0.001$, corrected), as shown in Figure III-3A. The most correlated IC out of this group (C-5) was selected to demonstrate the significant group difference in subject loadings (Figure III-3B) and to visualize this IC's connectome (Figure III-3C).

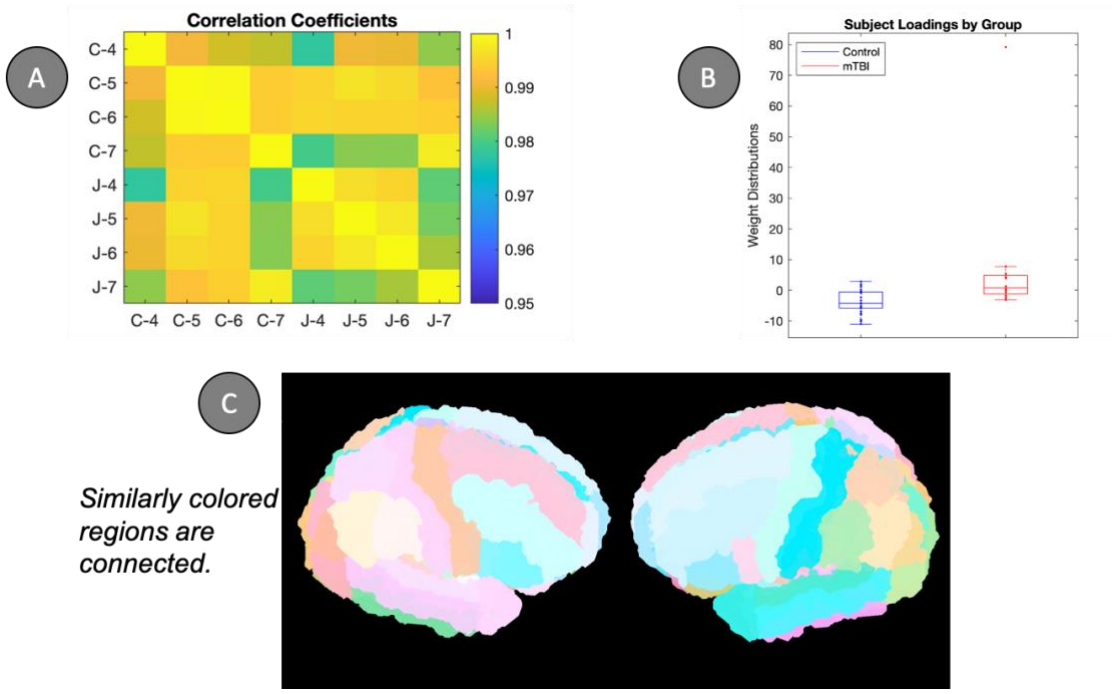


Figure III-3 An independent component presenting significant group differences. A singular independent component was consistently found across both the connectivity-only and joint ICAs. A) Extremely high correlation coefficients are seen for this component across connectivity-only and joint ICAs with X = {4,5,6,7} ICs. (Rows/columns of this matrix are denoted by "ICA type – X", with C connectivity-only and J = joint; so "C-4" means connectivity-only ICA with X=4.) All correlations are statistically significant ($p <<< 0.001$, corrected). B) The subject loadings for this component present a statistically significant ($p < 0.05$, corrected) difference between the control and mTBI populations. C) The independent component back-projected into connectivity data space; the component is visualized on a representative T1w volume, where similarly colored regions are connected (for a full description of the visualization procedure, see section 2.4).

The stability of this significant IC across the 8 ICAs suggests that the pattern of connectivity seen in Figure III-3C is a legitimate feature, not an artifact of ICA's optimization scheme. Furthermore, the IC's subject loadings show that this connectivity pattern is more prominent in mTBI subjects than in controls, implying that the connections may be indicative of mTBI pathology. Table III-1 presents the seven region-to-region connections with the largest weights in the IC from C-5. The strongest connection involves the right parietal operculum, which is believed to be involved with touch and pain perception [152], and the left occipital fusiform gyrus, a region involved in high level visual processing [153]. Several other regions in this list are associated with visual processing (right occipital fusiform gyrus [154], right calcarine cortex [155]), working memory (right/left superior frontal gyrus medial segment [156][157]), and other sensory functions (right angular gyrus [158], left opercular part of the inferior frontal gyrus [159]), all cognitive functions which have been known to be negatively impacted by mTBI [130].

Table III-1 Top seven most highly weighted connections in an IC from C-5 associated with significant group differences

Weight Rank	ROI 1	ROI 2
1	Right parietal operculum	Left occipital fusiform gyrus
2	Right occipital fusiform gyrus	Left superior frontal gyrus medial segment
3	Left lateral orbital gyrus	Right angular gyrus
4	Right posterior orbital gyrus	Right central operculum
5	Right calcarine cortex	Left anterior orbital gyrus
6	Right superior frontal gyrus medial segment	Left frontal operculum
7	Left opercular part of the inferior frontal gyrus	Right opercular part of the inferior frontal gyrus

4.2. Symptom correlations

To further explore the relationship between these IC features and mTBI symptomology, Spearman rank correlations were calculated between mTBI subject symptom scores and IC loadings for all ICs across all ICAs. Out of all tests, a single IC in S-5 was found to be significantly correlated with two symptoms after Bonferroni correction: forgetfulness and slowed thinking ($p < 0.01$, corrected for 110 comparisons). Figure III-4 shows these correlations between S-5 subject IC loadings and symptom scores for the two symptoms, along with the IC back-projected to the cortical thickness metric. This positive correlation between symptom severity scores and subject loadings indicates that the IC presents most in subjects with more

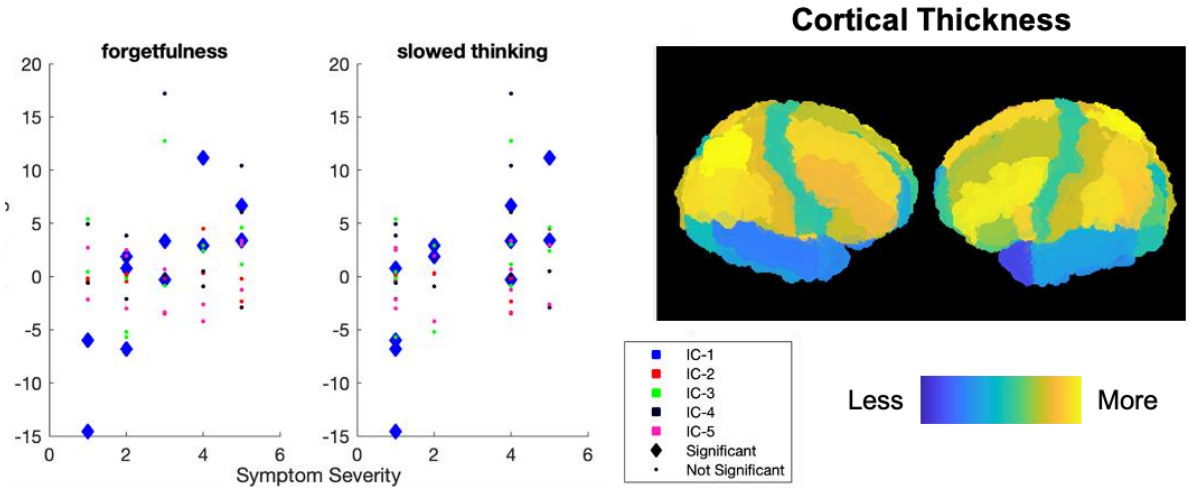


Figure III-4 An independent component presenting significant symptom correlations. Across all ICAs performed, only a single IC was found to be significantly correlated to mTBI symptom scores. This IC is from the surface-only ICA performed with 5 ICs. The two plots on the left in this figure show the statistically significant ($p < 0.01$, corrected) positive correlations between mTBI subject IC loadings and symptom severity scores for forgetfulness and slowed thinking. The IC was back-projected into the cortical surface metrics dataspace. The cortical thickness metric is visualized here on a representative T1w volume; the magnitude of region IC coefficients are encoded via color, with blue corresponding to lower absolute thickness and yellow corresponding to increased absolute thickness.

severe forgetfulness and slowed thinking. Combining this with the IC's pattern cortical thickness pattern suggests that these mTBI symptoms may be related to increased bilateral cortical thickness. This finding is promising, since other studies have also found increased prefrontal cortical thickness to be associated with poorer outcomes in mTBI subjects one year after the acute injury [94].

5. Discussion

We have presented a joint analysis of structural connectivity and cortical surface structure in mTBI via ICA. We demonstrated the potential of this framework for hypothesis generation regarding mTBI pathologies via two significant IC findings. First, our analysis identified a stable IC that presented significant differences in IC loadings between mTBI subjects and controls, in both connectivity-only and joint ICA. This finding revealed structural connectivity changes in individual cortical surface regions that are potentially linked to mTBI pathologies. Additionally, we found a strong correlation between mTBI IC loadings and symptom severity scores for the slowed thinking and forgetfulness symptoms; examining the surface metric representation of this IC suggests that the severity of these symptoms is linked to a relative increase in cortical thickness among mTBI subjects.

In summary, the proposed framework was found to be effective for detecting potential structural connectivity and cortical surface abnormalities in mTBI. A major advantage of this framework was the ability to back-project ICs into data space, allowing for anatomical interpretations of group differences and symptom correlations. This explanatory power makes ICA a desirable method for exploring disease pathologies compared to "black box" neural networks.

Chapter IV

pyPheWAS: A Phenome-Disease Association Tool for Electronic Medical Record Analysis

1. Overview

Along with the increasing availability of electronic medical record (EMR) data, phenome-wide association studies (PheWAS) and phenome-disease association studies (PheDAS) have become a prominent, first-line method of analysis for uncovering the secrets of EMR. Despite this recent growth, there is a lack of approachable software tools for conducting these analyses on large-scale EMR cohorts. In this article, we introduce *pyPheWAS*, an open-source python package for conducting PheDAS and related analyses. This toolkit includes 1) data preparation, such as cohort censoring and age-matching; 2) traditional PheDAS analysis of ICD-9 and ICD-10 billing codes; 3) PheDAS analysis applied to a novel EMR phenotype mapping: current procedural terminology (CPT) codes; and 4) novelty analysis of significant disease-phenotype associations found through PheDAS. The pyPheWAS toolkit is approachable and comprehensive, encapsulating data prep through result visualization all within a simple command-line interface. The toolkit is designed for the ever-growing scale of available EMR data, with the ability to analyze cohorts of 100,000+ patients in less than 2 hours. Through a case study of Down Syndrome and other intellectual developmental disabilities, we demonstrate the ability of pyPheWAS to discover both known and potentially novel disease-phenotype associations across different experiment designs and disease groups. The software and user documentation are available in open source at https://github.com/MASILab/pyPheWAS.

2. Introduction

Since the early 2000s, the introduction of computers in healthcare has led to the adoption of Electronic Medical Records (EMR) in healthcare systems across the globe. Initiatives such as the National Institutes of Health's Clinical and Translational Science Awards have advanced this electronic healthcare landscape by providing funding for institutions to generate, store, and share healthcare information with the ultimate goal of improving patient care [160]. Many institutions, such as Intermountain Healthcare and Vanderbilt University Medical Center (VUMC), have risen to the challenge, building large EMR repositories that encompass patient demographics, insurance billing data, genetic sequences, medication records, laboratory testing, and more [161], [162]. These rich EMR repositories create opportunities for "secondary use" of health data, meaning the utilization of health data outside of direct patient care. In medical research, this translates to opportunities for investigators to study disease progression and comorbidities, treatment

efficacy, genetic factors, systemic problems, and biases in the medical system, among other goals [163]. Yet, taking advantage of these complex databases is not a simple task; the EMR is often biased, incomplete, and inaccurate [54]. Consequently, rapid increases in the size and availability of EMR resources have led to a surge in the development of EMR analysis methods, particularly in the area of deriving and studying EMR phenotypes [54], [164], [165].

A particularly successful type of EMR phenotype analysis is the phenome-wide association study (PheWAS). This analysis is closely related to the genome-wide association study (GWAS), a framework in which a single phenotype is tested for associations with many genotypes [166]. In contrast, a PheWAS tests the association between a single genotype and many EMR-derived phenotypes. This method was pioneered by Denny et al [47] with a proof of concept study that examined the associations between five single nucleotide polymorphisms (SNPs) and 776 EMR phenotypes; this PheWAS both replicated five previously reported SNP-disease associations and identified nineteen potentially novel associations, presenting PheWAS' potential for supporting often-underpowered GWAS investigations. Three years later, the same group performed a large-scale trans-institutional validation of PheWAS, confirming its use as an unbiased phenotype interrogation technique and hypothesis generation tool [167]. The 776 phenotypes used in the proof-of-concept study were derived from International Classification of Disease (ICD) version 9 billing codes; these phenotypes were designated PheWAS Codes, or PheCodes, and have since been publicly released and expanded to a cover a total of 1,866 EMR phenotypes [167], [168].

Since its conception, this groundbreaking technique has inspired many investigations of different sections in the genome. In a similar vein as its initial proof-of-concept, PheWAS has been used to examine the phenotype signature of the HLA-DRB1*1501 haplotype (a genetic variant linked with Multiple Sclerosis) [169], the major histocompatibility complex region of chromosome 6 [170], 31 SNPs associated with serum uric acid [171], and other genome regions of interest revealed via GWAS [172]. Other interesting applications of this technique include examining the contribution of Neanderthal genetic variants to the phenotypes of modern humans [173], and evaluating self-reported ICD-9 records in a large-scale 23andMe database for the purpose of genetic drug targeting [174].

Inspired by PheWAS, an alternative approach has emerged which scans the phenome for associations with non-genetic targets. This extension of PheWAS is advantageous due to the costly nature of genotyping, and therefore, the huge amount of EMR data available when linked genetic data are no longer necessary [175]. This framework has been used to examine linked dental and medical records to identify ICD-9 phenotypes related to periodontitis [176]. In a federated query task, it was used to retrieve records of patients who had a rare condition (multiple myeloma) across multiple institutions, and then further delineate specific subgroups that experienced serious complications [177]. Other examples include scans of ICD-9 phenotype

associations with white blood cell count [178] and non-Hodgkin lymphoma in Medicare claims [179]. Recently, we observed the potential for confusion of study designs with genetic and non-genetic phenome association studies. After consultation with the PheWAS team, we now refer to studies that do not include genetic markers but still use mass univariate regression as Phenome-Disease Association Studies (PheDAS) [48], an example of which is shown in Figure IV-1.

In light of the pervasiveness of this EMR analysis technique, we present pyPheWAS: a comprehensive toolkit for performing PheWAS and PheDAS analyses. The original PheWAS software, written by the team that developed the PheWAS method, is implemented in R and includes core PheWAS functions [180]. The pyPheWAS package reimplements that core functionality in Python, a language that has become more widespread in the machine learning community and adds a collection of easy-to-use command line tools that covers everything from preprocessing EMR data to visualizing results. It includes analysis of ICD-9 and ICD-10 phenotypes, as well as a novel analysis for Current Procedural Terminology (CPT) code phenotypes. It is important to note that pyPheWAS is not a neuro-centric toolkit, although its methods allow investigators to explore the clinical progression of many neurological conditions. Additionally, pyPheWAS is agnostic to the dependent variable, and therefore can be used to implement either PheWAS or PheDAS; for the remainder of this article, we will focus specifically on PheDAS analyses.

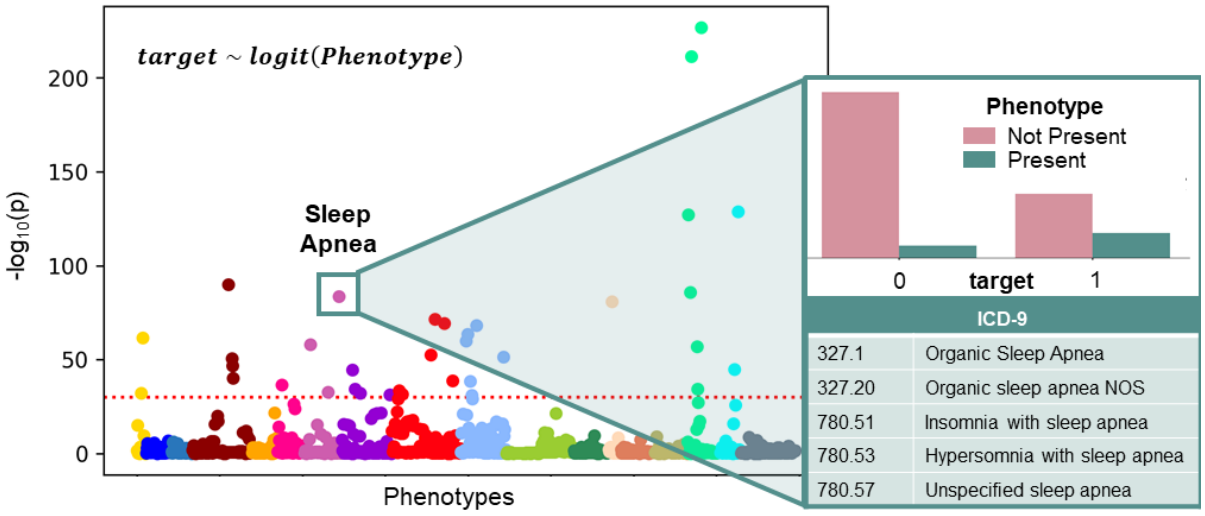


Figure IV-1 Overview of PheDAS. In the background, a Manhattan plot shows the statistical significance of many phenotypes in relation to a single target variable (*target*). Phenotypes are sorted into and colored by category, and the significance threshold for multiple comparisons correction is marked with a dashed horizontal line. These relationships were estimated by individually modeling the target variable as a function of each phenotype using a logistic regression. For a closer look, the significant phenotype *Sleep Apnea* is highlighted. The distribution of subjects from each *target* group that do (not) present the *Sleep Apnea* phenotype is shown, along with the ICD-9 codes that map to this this phenotype.

In the following sections, we first describe the technical details of the pyPheWAS toolkit, including installation instructions, EMR data acquisition, data preprocessing, and analysis methods. Following this, we demonstrate the toolkit in action by performing a PheDAS analysis on a custom synthetic EMR dataset. We then perform a case study on real EMR data, comparing the EMR of Down Syndrome patients to patient with other Intellectual and Developmental Disabilities. Finally, we discuss PheDAS result interpretation and several limitations of the pyPheWAS package.

3. Methods

The overall workflow of a PheDAS analysis is shown in Figure IV-2. EMR events and group demographic data are preprocessed, mapped to meaningful phenotypes, used to model a target variable (such as a disease group), and then visualized for interpretation. Figure IV-3 presents the pyPheWAS toolkit, a collection of command line scripts that aims to make PheDAS-style analysis highly approachable, as this process can quickly become intractable given the sheer scale of EMR data coupled with a lack of easy-to-use software. This section describes the form and function of each tool in detail. Source code for pyPheWAS may be found on GitHub (https://github.com/MASILab/pyPheWAS). The full user documentation may be found at https://pyphewas.readthedocs.io/en/latest/.

3.1. Requirements and installation

pyPheWAS is a Python (version 3.6+) package hosted on pypi.org, making installation quick and easy. On any computer which has Python 3 and the popular package manager pip already installed, the user must simply enter `pip install pyPheWAS` in a terminal or command line to install the software. All tools are accessed

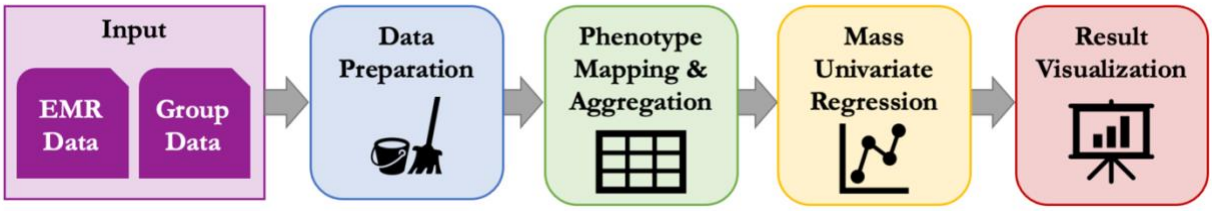


Figure IV-2 PheDAS analysis pipeline. Inputs to the pipeline include EMR data (ICD-9, ICD-10, or CPT codes) and group data (disease group, sex, race, etc.). The data is first prepared for analysis via case-control matching and censoring. Next, the EMR data is mapped to a set of predefined phenotypes (PheWAS or ProWAS Codes) and aggregated across each subject's record. Mass univariate regression is then performed across all phenotypes, where a target variable is modeled as a function of the phenotype plus any relevant covariates (such as sex or race) to determine the relationship between the target variable and each phenotype. Finally, the results are visualized to facilitate interpretation of target variable-phenotype relationship significance and effect size.

via command line. Note that there are no explicit hardware requirements for the pyPheWAS package, but the amount of memory available on the user's system will limit the size of experiment that can be performed.

Beyond software, the only requirements for using pyPheWAS is the format of the input data. Two primary files are expected by pyPheWAS tools: the phenotype file (EMR data) and the group file (demographic data). The phenotype file contains EMR events for all subjects in the group file, with a single line for each event. Events include an ICD or CPT code and the subject's age at the event. The group file contains demographic information, such as sex, and the target response variable which will be used in the logistic regression. The response variable may be pre-defined (such as a diagnosis), or it may be determined based on EMR data using the pyPheWAS data preparation tools. The phenotype and group files are linked by a column labeled 'id' which contains a unique identifier for each subject in the cohort.

3.1.1. EMR data acquisition

Many institutions have spent large amounts of time and resources to build multi-faceted data repositories that include genetic data, clinical records, and demographic information across large swaths of patient populations. A few prominent repositories include the Healthcare Cost and Utilization Project's (HCUP) National Inpatient Sample [181], the eMERGE Network [182], VUMC's Synthetic Derivative

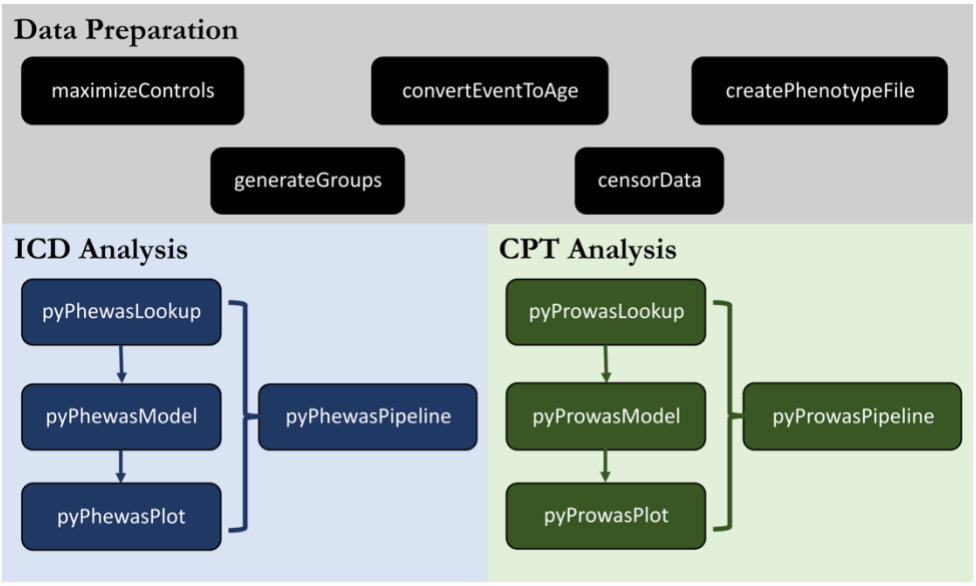


Figure IV-3 pyPheWAS package tools. The package is composed of three main tool sets: data preparation, ICD analysis, and CPT analysis. Data preparation tools focus on preprocessing EMR data, e.g., case/control matching (*maximizeControls*) and censoring events (*censorData*). The ICD analysis tools run PheDAS on ICD code data, while the CPT analysis tools run PheDAS on CPT code data. The function and usage of all tools are described in the *Methods* section.

(VUMC-SD) [161], Intermountain Healthcare's Enterprise Data Warehouse [162], the Utah Population Database [183], and the Rochester Epidemiology Project [184]. Due to the sensitive nature of EMR and protections set forth by the Health Insurance Portability and Accountability Act (HIPAA), an approval process is generally required to obtain access to these repositories. For example, in order to obtain the ICD and CPT records used for this article's Down Syndrome case study from VUMC-SD, we first were required to obtain study approval from Vanderbilt University's Institutional Review Board, sign a data use agreement, and pay a fee for repository use. We then worked with analysts at VUMC-SD to identify our target population using specific ICD codes and other diagnosis information. With our population identified, the VUMC-SD then pulled the requested ICD, CPT, and demographic records. Such processes are common across many EMR repositories. Though these procedures were designed to protect patient information, they also present steep entry barriers for aspiring EMR researchers. Therefore, we have made the synthetic dataset developed for this article publicly available through pyPheWAS's GitHub repository, allowing users to familiarize themselves more quickly with both EMR data and PheDAS methods (see the Results section for details). We hope that this resource will inspire similar accessibility efforts and enthusiasm for large-scale EMR analysis.

3.2. Data preparation

The pyPheWAS package provides several useful data preparation functions so that users do not have to directly manipulate the very large data files often used for PheDAS studies.

3.2.1. Defining case and control groups

The first step in a PheDAS study is defining which subjects are cases and which are controls. In the absence of externally defined group assignments (such as genetic markers [172] or white blood cell count [178]), ICD codes themselves may be used as a proxy for diagnosis [168], [185] (although sources of error for this are well known [49]). The ICD-9 code 758.0 – Down's syndrome, for example, may be used as a proxy for the actual clinical diagnosis of Down Syndrome. Due to the noisy nature of EMR, however, a minimum frequency threshold is applied to codes used for this proxy diagnosis based on the notion that the more frequently a subject is assigned a certain ICD code, the more likely it is that they legitimately have the target condition.

To address this need, the *createPhenotypeFile* function sorts subjects into case and control groups based on the presence or absence of ICD codes in subjects' records. At a minimum, *createPhenotypeFile* requires a phenotype file, a list of ICD-9 and ICD-10 codes that define the case group, and the minimum frequency

of those codes in a subject's record to be considered part of the case group. Users may specify whether this frequency threshold is a daily threshold (code frequency is calculated based on the number of unique days over which a code is recorded; ignores multiple records of a code within a single day) or an absolute threshold (code frequency is calculated based on the absolute number of code events; includes multiple records of a code within a single day). All subjects listed in the phenotype file who have at least the minimum frequency of provided codes in their record are assigned to the case group (target=1). Subjects who have the provided codes in their record but fall below the specified frequency are considered ambiguous and, consequently, excluded. All remaining subjects are assigned to the control group (target=0). These group assignments are saved to a comma-separated values (CSV) file containing A) only subject IDs and target variable assignments, or B) the target variable assignment added to an existing group file specified by the user.

In the basic configuration described above, the control group is comprised of all non-case and non-ambiguous subjects. In some experiments, however, it may be desirable to enforce stricter control group inclusion criteria; *createPhenotypeFile* provides two commonly used practices for narrowing the scope of PheDAS control groups. The first method excludes subjects from the control group based on both the provided case codes and codes *related to* those case codes; this prevents the control group from becoming contaminated by conditions similar to the target condition. The list of related codes may be supplied by the user or pulled from the ICD phenotype map (see the *pyPhewasLookup* section for details on the ICD phenotype map used by pyPheWAS). The second method allows users to target a specific condition for the control group. For example, a PheDAS could be performed comparing Alzheimer's disease patients (case) to Vascular Dementia patients (controls). In this case, the user would supply *createPhenotypeFile* with lists of ICD-9 and ICD-10 codes for both the case group and the control group. The control group is then composed of subjects not in the case group that have at least the minimum frequency of provided control group codes in their record. Optionally, a second argument may be provided to the code frequency input; if this is specified, the second frequency value is applied to the control group.

3.2.2. Converting dates to ages

EMR event data is usually tagged with dates. In certain cases, a researcher may choose to study EMR records only within a specific period of time, or they may want to use age as a covariate. For convenience, the *convertEventToAge* script allows users to quickly convert dates associated with CPT and ICD events to subject ages at the events. This function requires the phenotype file for which event dates are to be converted and a corresponding group file that contains each subjects' date of birth. Optionally, the user may specify the level of precision with which ages are saved in the output phenotype file.

3.2.3. Censoring event data

A common aim of medical studies is to examine specific periods of time in patients' lives. For example, one may be interested in the EMR signature for the five years leading up to an Alzheimer's Disease diagnosis or for children ages 10 to 18 who have Autism/Autism Spectrum Disorder. Data censoring such as this is incorporated into the pyPheWAS toolkit with the *censorData* function. Similar to other tools, this function requires a phenotype file containing the events to be censored and a group file containing subject information, along with user-specified censoring start and/or end years. Censoring can be applied to the data in two distinct ways. The first method censors the absolute value of event ages (e.g. the age at CPT or ICD code events) to only those that fall within the user-defined start and end years, such that all preserved events fulfill the equation

$$start \leq eventAge \leq end \quad (1)$$

The second method instead censors event ages relative to an external event, such as subject age at diagnosis or surgery. In this case, the interval between the events is considered such that all preserved events fulfill the equation

$$start \leq (externalEventAge - eventAge) \leq end \quad (2)$$

The censored events are saved to a new phenotype file, and all subjects with event data remaining after censoring are written to a new group file.

3.2.4. Case-control matching

Another common practice in case-control studies such as PheDAS is matching a certain number of control subjects to each case subject based on specified group variables. The pyPheWAS toolkit includes case-control mapping through its *maximizeControls* tool. This tool requires a group file containing group variables and case/control assignments, a list of variables to match on, tolerance intervals for each of those matching variables, and the desired ratio of controls to cases. It constructs a bipartite graph from the cohort in which subjects are the vertices, matching variables are edges, and the case and control groups are two disjoint independent vertex sets. To find a first set of matches, it uses the Hopcroft-Karp algorithm [186] to find a mapping between the case and control sets that results in maximal cardinality (i.e., matches). If the desired matching ratio is larger than 1:1, the first set of matched controls are removed from the graph, and the Hopcroft-Karp algorithm is applied again to find a second set; this repeats until either the desired matching ratio is satisfied or there are no more possible matches. A new group file is saved containing all matched subjects, along with a separate matched pairs file containing the explicit mapping between each individual case and its control(s).

3.3. Scanning the ICD phenome

As outlined in Figure IV-2, the core of PheDAS analysis may be broken up into three distinct phases: 1) mapping EMR data to phenotypes, 2) mass univariate regression of phenotypes, and 3) result visualization. The ICD analysis tools in the pyPheWAS package focuses on processing ICD-9-CM and ICD-10 codes, with individual functions devoted to each of the three phases: *pyPhewasLookup*, *pyPhewasModel*, and *pyPhewasPlot*, respectively. This section describes each of those functions in detail.

3.3.1. pyPhewasLookup

The *pyPhewasLookup* function transforms individual ICD code records into feature matrices ready to be processed by the *pyPhewasModel* function; Figure IV-4 provides a detailed view of this function. It requires as input a phenotype file containing the ICD records of each subject and a group file containing the target and covariate variables. The feature matrices are constructed in two phases: 1) mapping and 2) aggregation. In the mapping phase, each ICD code in the phenotype file is mapped to its corresponding phenotype. The phenotype mapping used by *pyPhewasLookup* includes 1,866 hierarchical phenotype codes (PheCodes); it was originally constructed solely for ICD-9 codes by Denny et al [167], with later improvements to the ICD-9 mapping [168] and the addition of an ICD-10 code mapping [187]. It should be noted that these mappings are not complete. They do not cover the full range of ICD-9 and ICD-10 codes, so ICD events in a subject's record which are not included in the mapping are removed from the study. When these removals occur, *pyPhewasLookup* notifies the user regarding the number of removed events; optionally, the user may choose to export the list of removed events for further inspection.

The aggregation phase next reformats the mapped data from longitudinal events to subject-by-PheCode feature matrices. Three types of feature matrices are created, in which the columns are PheCodes and the rows are subjects from the group file. The first matrix is the core of the PheWAS analysis; denoted the *aggregate measure* matrix, it contains a single aggregate measure for each PheCode across all subjects. To allow researchers to investigate different aspects of the EMR, three distinct types of aggregation may be performed: binary, count, and duration. Binary aggregation investigates the relationship between the target variable and the presence or absence of a PheCode. Its feature matrix contains only zeros (the PheCode was *absent* in the subject's record) and ones (the PheCode was *present* in the subject's record). Count aggregation investigates the relationship between the target variable and the number of occurrences of a PheCode. Its feature matrix contains positive integers that correspond to the total number of times each PheCode occurred in a subject's record. Duration aggregation investigates the relationship between the

target variable and the interval of time over which a PheCode is experienced. Its feature matrix contains the time in years between the first and last occurrences of each PheCode in a subject's record.

The second and third feature matrices are independent of aggregation type and are created as optional covariates for *pyPhewasModel*. The *ICD age* feature matrix contains the maximum age recorded for each PheCode in a subject's record; if the subject has no records of that PheCode, the subject's overall maximum recorded age is reported. The *PheWAS covariate* matrix allows researchers to use the presence/absence of a specified PheCode as a covariate in the regression. Across all columns, it records a one if the specified PheCode is present in a subject's record or zero if the specified PheCode is absent. All three feature matrices are saved as CSV files in preparation for the *pyPhewasModel* step.

3.3.2. pyPhewasModel

The *pyPhewasModel* function performs the mass logistic regression which is the focal point of PheDAS analyses. It requires the feature matrix files generated by *pyPhewasLookup* in addition to the group file. For each PheCode, *pyPhewasModel* computes a univariate logistic regression of the form

$$Pr(target) \sim logit(A_{phe} + covariates) \tag{3}$$

where the target variable and covariates are specified by the user, and A_{phe} is the aggregate measure vector for a particular PheCode *phe* taken from the aggregate measure matrix.

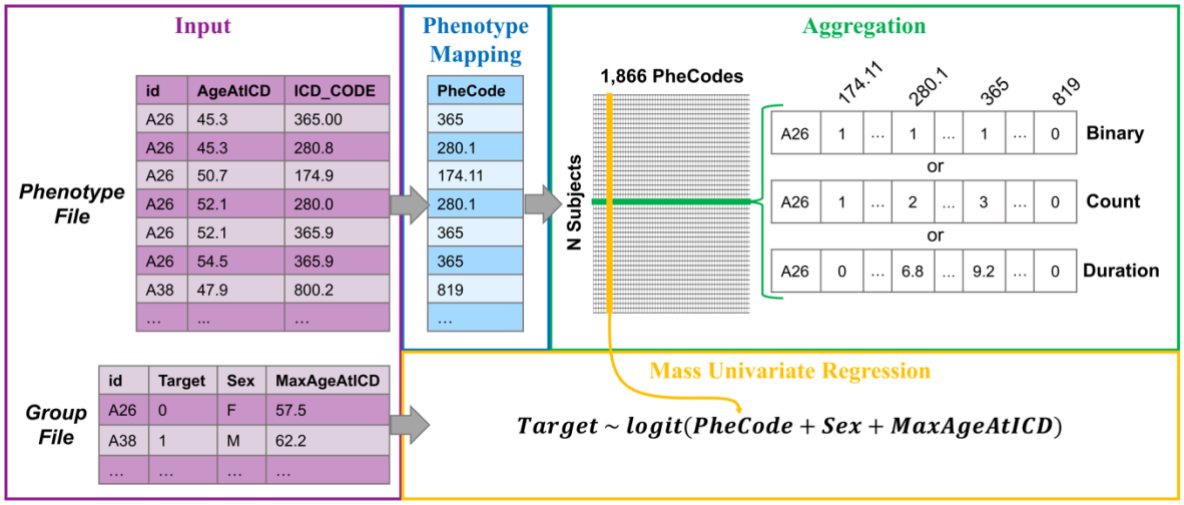


Figure IV-4 Detailed look at phenotype mapping, aggregation, and regression in pyPhewasLookup. On the far left, excerpts from input phenotype and group files containing data from subjects A26 and A38 are shown. ICD codes from the phenotype file are mapped to corresponding PheCodes. These codes are then aggregated via one of three possible methods for each subject; binary, count, and duration aggregations for subject A26 are shown. Finally, the aggregated EMR data is combined with group data (in this case, the target variable Target, and covariates Sex and MaxAgeAtICD), and univariate regressions are computed for each PheCode.

These regressions are only computed on PheCodes for which $A_{phe}$is non-zero in at least X subjects, where X is a user-defined threshold that defaults to 5. This requirement cuts out PheCodes which lack sufficient statistical power. The model is fit to the data via regularized maximum likelihood optimization. The Python library statsmodels is used to generate and fit the logit model to the PheCode data [188]. Regression results are again saved in a CSV file for the user to review and visualize. This file reports the log odds ratio, confidence interval, standard error, and uncorrected p-value estimated from A_{phe} for each PheCode *phe*.

3.3.3. pyPhewasPlot

Visualization of the PheDAS mass regression is performed by the *pyPhewasPlot* function. It requires the regression file produced by *pyPhewasModel* and the user's desired multiple comparisons correction method; both False Discovery Rate (FDR) and Bonferroni are available. From these inputs, it creates three complementary views of the PheDAS analysis using the Python matplotlib library [189]. The first is a Manhattan plot, a classic GWAS plot which compares statistical significance across PheCodes. This view presents PheCodes across the horizontal axis, with negative $\log_{10}$(p-value) along the vertical axis; PheCode markers on the plot are colored and sorted according to 18 general categories (mostly organ systems and disease groups, e.g. "circulatory system" and "mental disorders"), allowing users to distinguish related PheCodes. To enhance legibility, the plot only labels PheCodes which are significant after the chosen multiple comparisons correction is applied.

The second view is a Log Odds plot, which compares effect size across PheCodes. In this plot, the log odds of each PheCode and its confidence interval are plotted on the horizontal axis, with PheCodes plotted along the vertical axis. Similar to the Manhattan plot, PheCode markers are sorted and colored by category; only PheCodes which are significant after multiple comparisons correction are shown.

The final view is a Volcano plot. This view combines the previous two, presenting an overview of the entire experiment. In the Volcano plot, significance, negative $\log_{10}$(p-value), is represented by the vertical axis, and effect size, log odds, is represented by the horizontal. All PheCodes in the regression file are included on this plot, with marker color corresponding to each PheCodes's level of significance (none, FDR, Bonferroni). To ensure legibility, only PheCodes that are significant after FDR or Bonferroni correction are labeled.

These three views together provide a comprehensive visualization of the PheWAS analysis. The Volcano plot allows the user to see an overview of the entire experiment, with the Manhattan and Log Odds

plots then providing a detailed view for closer examination of significant results. The user has the option of either opening the plots in an interactive window or immediately saving them as image files.

3.3.4. pyPhewasPipeline

pyPhewasPipeline is a streamlined combination of *pyPhewasLookup*, *pyPhewasModel*, and *pyPhewasPlot* created for convenience. Its required inputs are the phenotype file, group file, and the regression type. All intermediate results (feature matrices, regressions) are saved. In addition to the Volcano plot, Manhattan and Log Odds plots are created for both FDR and Bonferroni corrections by default. Optional arguments allow users to modify every step of the pipeline (adding covariates, specifying significance level, etc.).

3.4. Scanning the CPT phenome

Procedure wide association studies (ProWAS) are nearly identical to PheDAS, with one critical difference: the EMR data. While PheDAS investigates ICD code phenotypes, ProWAS investigates CPT code phenotypes. Examining ICD codes may provide insight into patient diagnoses; in a similar vein, examining CPT codes may reveal patterns in how patients are treated. As such, these tools are identical to their PheDAS counterparts, with the exception of the EMR-phenotype mapping. As with PheDAS, ProWAS consists of three main stages: 1) mapping EMR data to phenotypes, 2) mass univariate regression of phenotypes, and 3) result visualization. The CPT analysis tools for each of these stages are analogous to the ICD analysis tools: *pyProwasLookup*, *pyProwasModel*, and *pyProwasPlot*.

ProWAS employs a custom procedural phenotype map, linking 10,396 CPT codes to 1,681 ProWAS Codes (ProCodes) [50]. This map is based on the Clinical Classification System for CPT codes provided by the Healthcare Cost and Utilization Project (HCUP) Agency for Healthcare Research and Quality [190]. Starting with 236 of the HCUP clinically meaningful CPT categories, additional granularity was added to the mapping with guidance from medical experts, until 1,681 ProCodes were defined. For example, the HCUP category 66 (Procedures on spleen) was split into ProCodes 66.1 (Splenectomy), 66.2 (Splenorrhaphy), and 66.3 (Laparoscopy). The full CPT-ProCode map may be found at https://github.com/MASILab/pyPheWAS.

4. Results

In this section, we demonstrate the utility of the pyPheWAS package via two example PheDAS experiments. In Experiment 1, we evaluate the package by analyzing a synthetic EMR dataset which contains several hand-crafted PheCode associations. In Experiment 2, we perform a case study on real EMR data, in which we compare subjects with Down Syndrome (DS) to controls with other Intellectual or Developmental Disabilities (IDD).

4.1. Experiment 1: Synthetic dataset

4.1.1. Dataset construction

Our synthetic dataset consists of 10,000 individuals, split evenly into 5,000 case (Dx=1) and 5,000 control (Dx=0) subjects, where Dx is the target variable. Other demographic variables include biological sex and maximum age at visit (MAV). Sex was intentionally made a confounding variable by skewing the female:male ratios between the case and control groups. MAV was calculated as the maximum age recorded from ICD records generated for each individual. These synthetic demographic variables are summarized in Table IV-1.

Table IV-1 Synthetic dataset demographic summary

	Subjects	Sex [% Female]	Max Age At Visit [mean (std.)]
Case (Dx=1)	5,000	70%	59.946 (9.563)
Control (Dx=0)	5,000	40%	60.802 (9.448)

While curating ICD code events for each individual, three types of PheCode associations were created. *Primary* PheCode associations were true associations between Dx and the PheCode. ICD events were generated such that each of these PheCodes would have a unique pre-specified effect size (log odds ratio) across the full cohort; individuals' ages for each event were randomly generated using a uniform distribution over the range [30, 50]. pyPheWAS should accurately estimate each primary association's effect size and determine that the association is statistically significant. We generated nine primary PheCode associations, including six positive associations and three negative associations (Table IV-2). In contrast, *background* PheCode associations were insignificant associations between Dx and the PheCode. ICD events were generated such that each background PheCode would have a small pre-specified effect size, randomly generated via a uniform distribution over the range [-0.1, 0.1]; again, individuals' ages for each

event were randomly generated using a uniform distribution over the range [30, 50]. pyPheWAS should accurately estimate each background association's effect size but determine that the association is insignificant. Twenty background PheCode associations were generated for the synthetic dataset.

Finally, *confounded* PheCode associations were false positives caused by the confounding effect of either sex or age. Without controlling for the confounding variable, pyPheWAS should identify a significant association with these confounded PheCodes; including the confounding variable as a covariate, however, should reduce (or eliminate) the confounded association. PheCode 174.1 (*Breast cancer [female]*) was used as a sex-confounded PheCode (Table IV-2). To produce the confounding effect, ICD events were generated such that all females in the dataset had equal odds of having PheCode 174.1 in their record; event ages were generated in the same way as primary PheCodes. Because females were disproportionally represented across the case and control groups, however, the PheCode's cohort-wide effect size is positively skewed to a 0.6 log odds ratio. Additionally, PheCode 292.2 (*Mild cognitive impairment*) was used as an age-confounded PheCode (Table IV-2). ICD events were generated such that PheCode 292.2 would have a -0.2 log odds ratio; however, event ages were randomly generated using a uniform distribution over the higher age range [65,70]. This resulted in PheCode 292.2 being highly associated with larger values of MAV. This synthetic EMR dataset has been made freely available on pyPheWAS's GitHub.

4.1.2. PheDAS analysis

The synthetic EMR dataset was analyzed in a single command via *pyPhewasPipeline*. We first ran Reg A, a minimal PheDAS with no covariates (Figure IV-5a, Table IV-2). Reg A successfully estimated the log odds ratios of all nine primary PheCodes and determined that they were statistically significant after Bonferroni multiple comparisons correction. The twenty background codes were accurately identified as insignificant. Reg A also correctly estimated the apparent effect sizes and significance of the two confounded PheCodes, 174.1 and 292.2; this was expected since Reg A did not properly control for the confounding variables. To remedy this, we next ran Reg B, a PheDAS that included both sex and MAV as covariates (Figure IV-5b, Table IV-2). With this modification, pyPheWAS recognizes the confounded PheCodes and now correctly determines that they are insignificant.

Table IV-2 PheDAS regression results for the primary and confounded PheCodes in the synthetic dataset.

				Reg A		Reg B	
	PheCode	**Phenotype**	**Actual LOR[a]**	**LOR[a]**	**p-val[b]**	**LOR[a]**	**p-val[b]**
Primary	338.2	Chronic pain	1.50	1.500	**	1.490	**
	340	Migraine	1.10	1.099	**	1.128	**
	1011	Complications of surgical and medical procedures	0.70	0.700	**	0.700	**
	296.22	Major depressive disorder	0.60	0.600	**	0.579	**
	530.11	GERD	0.30	0.300	**	0.302	**
	401	Hypertension	0.25	0.249	**	0.257	**
	041	Bacterial infection NOS	-0.20	-0.200	**	-0.194	**
	1009	Injury, NOS	-0.60	-0.599	**	-0.604	**
	495	Asthma	-1.00	-1.000	**	-0.991	**
Confounded	174.1	Breast cancer [female]	0.66 / 0.00[c]	0.662	**	0.004	-
	292.2	Mild cognitive impairment	-0.2	-0.199	**	-0.500	-

[a] log odds ratio; [b] significant after Bonferroni correction (**), insignificant (-); [c] male+female log odds ratio / female-only log odds ratio

4.2. Experiment 2: Down syndrome case study

4.2.1. Dataset acquisition

This case study and its procedures were carried out in accordance with the Institutional Review Board of Vanderbilt University and VUMC. Our EMR dataset was obtained from the Synthetic Derivative at Vanderbilt University Medical Center as a fully deidentified collection of clinical data via the Vanderbilt Institute for Clinical and Translational Research. All researchers working with this data received proper Human Subjects training. Our initial cohort consisted of 901,883 subjects, each having records of sex, race, and date of birth. Collectively, these subjects had 20,519,770 ICD event records and 19,555,593 CPT event records.

4.2.2. Cohort preparation

We first identified all DS cases and IDD controls in our cohort using the *createPhenotypeFile* tool. For this case study, we defined DS and IDD subjects based on ICD-9 and ICD-10 codes, which are listed in Appendix A. For both the DS and IDD groups, we required that a subject have at least 2 records of the codes listed in this Appendix to be included. From these criteria, we found 2,315 DS subjects and 106,059 IDD subjects. This control group was intentionally designed to cover a broad range of IDDs in order to elucidate phenotypic patterns that are unique to DS. Future investigations with more specific hypotheses, however, may benefit from curating a more targeted comparison group; for example, using PheDAS to

compare autism spectrum disorder with DS could reveal more about the absence of psychiatric comorbid conditions in DS.

After obtaining subject event ages via the *convertEventToAge* tool, we next used the *censorData* tool to restrict both the ICD and CPT data to only those events occurring previous to age 10. After this censoring, we were left with 1,830 DS and 52,138 IDD subjects that had both ICD and CPT events previous to age 10. Finally, due to the highly unbalanced nature of our cohort, we used the *maximizeControls* tool to match our DS cases to IDD controls with a 1:2 ratio. Matching was performed based on sex (exact match), race (exact match), and minimum ICD/CPT event age (± 0.3 years). One DS subject was dropped at this point, as there did not exist a single suitable match in the IDD cohort (even after varying the tolerance for the minimum age matching criterion), leaving us with 1,829 DS subjects and 3,658 IDD subjects.

4.2.3. ICD record analysis

To analyze the ICD signature of DS subjects compared to IDD controls, we performed a binary pyPheWAS analysis. We constructed a binary feature matrix via *pyPhewasLookup*, then performed mass logistic regression across all PheCodes with the maximum ICD age feature matrix as a covariate using *pyPhewasModel*. Applying Bonferroni multiple comparisons correction resulted in 177 PheCodes that were statistically significant; the top five most significant PheCodes in this experiment were found to be Cardiac shunt/heart septal defect (747.11), Muscle weakness (772.30), Hypothyroidism NOS (244.40), Cardiac

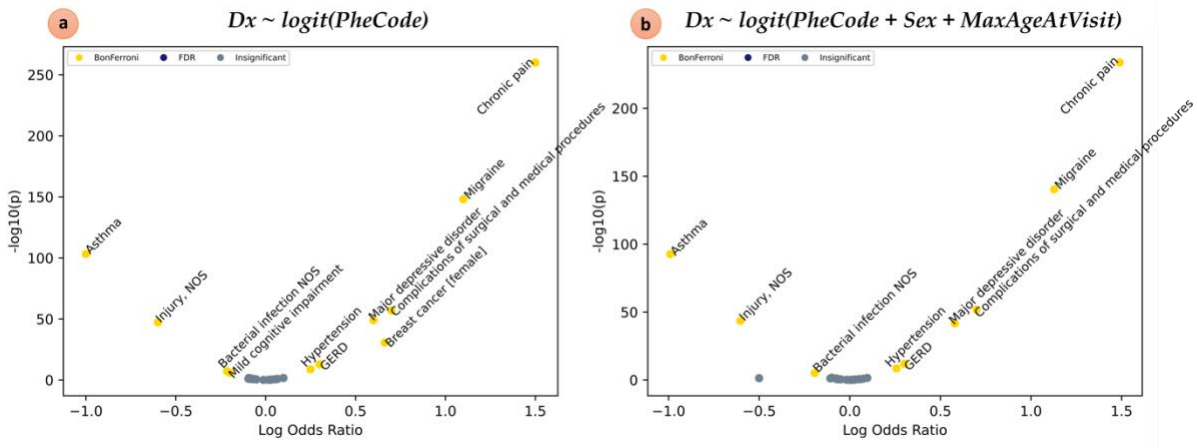
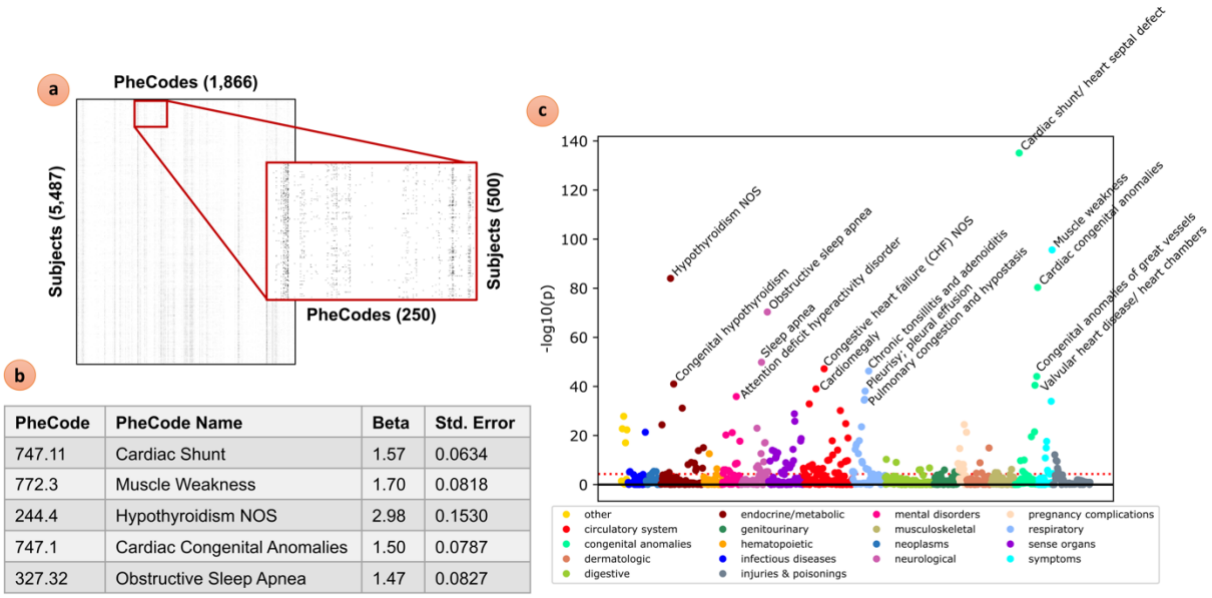


Figure IV-5 PheDAS applied to a synthetic dataset. a) Volcano plot resulting from a PheDAS without covariates. pyPheWAS successfully identified the nine primary PheCode associations in the synthetic dataset and ignored the twenty background associations. The confounded PheCodes (*Breast cancer [female]* and *Mild cognitive impairment*) were also identified as significant. b) Volcano plot resulting from a PheDAS with the *Sex* and *MaxAgeAtVisit* covariates. Controlling for sex and age effects successfully repressed findings from confounded PheCodes (*Breast cancer [female]* and *Mild cognitive impairment*).

congenital anomalies (747.10), and Obstructive sleep apnea (327.32). All regression results were plotted via *pyPhewasPlot* with the Bonferroni threshold. This analysis and the resulting Manhattan plot are presented in Figure IV-6.

4.2.4. CPT record analysis

The CPT signature of DS subjects compared to IDD controls was analyzed in a similar manner. We first constructed a binary ProWAS feature matrix via *pyProwasLookup*. We then performed mass logistic regression across all ProCodes with the maximum CPT age feature matrix as a covariate using *pyProwasModel*. Applying Bonferroni multiple comparisons correction resulted in 109 ProCodes that were statistically significant, of which Spine radiology exam (226.4), Doppler echocardiography (193.5), Clinical nutrition (237.4), Transthoracic echocardiography (193.3), and Occupational therapy (212.4) were found to be the most significant. Due to the large number of significant ProCodes, the results were plotted via *pyProwasPlot* with a much stricter custom threshold ($p_{uncorrected}$ < 1e-30) in order to pare down results for discussion. This ProWAS analysis and its Log Odds plot of significant results are shown in Figure IV-7.



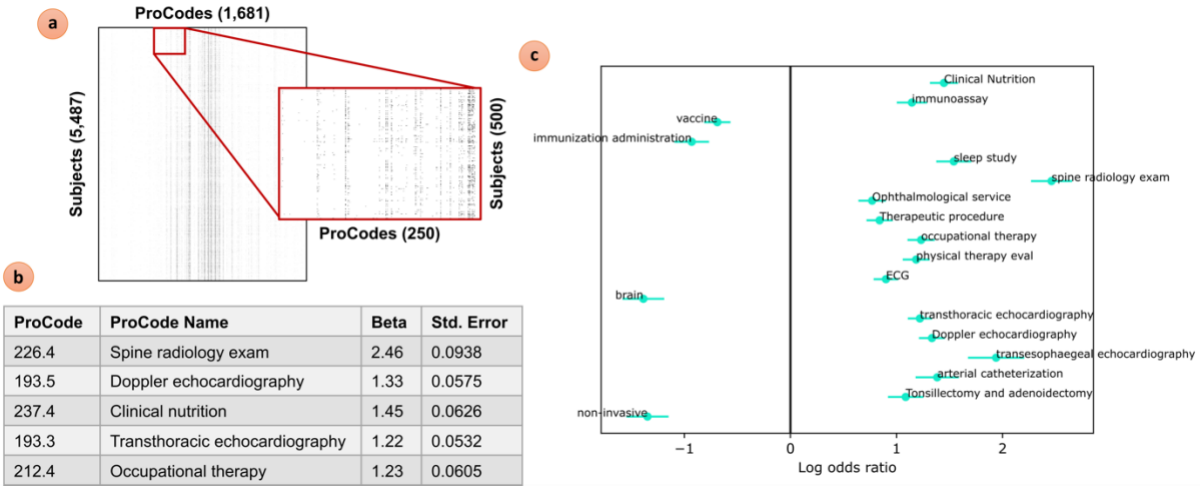
PheCode	PheCode Name	Beta	Std. Error
747.11	Cardiac Shunt	1.57	0.0634
772.3	Muscle Weakness	1.70	0.0818
244.4	Hypothyroidism NOS	2.98	0.1530
747.1	Cardiac Congenital Anomalies	1.50	0.0787
327.32	Obstructive Sleep Apnea	1.47	0.0827

Figure IV-6 Sample PheDAS of ICD records in DS vs. IDD subjects. (a) A binary feature matrix with PheCodes as columns and subjects as rows was constructed from the ICD event records mapped to PheCodes in *pyPhewasLookup*. (b) Mass univariate logistic regression was performed across PheCodes in the feature matrix using *pyPhewasModel*; regression results are listed for the top 5 most significant PheCodes (p <<< 0.001 after Bonferroni multiple comparisons correction). (c) Manhattan plot of all results is shown, with the top 14 most significant PheCodes labeled (p <<< 0.001 after Bonferroni multiple comparisons correction). The Bonferroni threshold is shown as a dotted red line.

5. Discussion

This article presents the pyPheWAS comprehensive toolkit for performing PheDAS analyses on EMR data. We have described the PheDAS process, wherein EMR data, specifically ICD or CPT codes, are first mapped to meaningful phenotypes and aggregated across each patient's record. These aggregate measures are then used along with specified covariates to perform mass univariate regression of a target variable on each phenotype. The results of this mass univariate regression are visualized in several ways to facilitate interpretation. We verified the pyPheWAS package by analyzing a synthetic dataset and then further illustrated its function in a real-world setting via a case study comparing DS subjects with non-DS IDD controls. With the analysis complete, our final consideration focuses on how to interpret PheDAS experiments.

The first question we must ask of a PheDAS is how do we verify its correctness? Since PheDAS is primarily a hypothesis generation method, there is no "correct" set of values we can test the strength, significance, or number of associations against. Despite this, PheDAS has a built-in verification test: expected associations. For practically any disease being tested via PheDAS, there are several previously known phenotype associations. These expected associations may be used as reassuring results in a study; a sanity check that establishes baseline credibility for all regression results [191]. Several such reassuring results are present in the ICD and CPT analyses of our case study. The Manhattan plot in Figure IV-6 shows that the PheCodes for Cardiac Congenital Anomalies, Hypothyroidism, and Obstructive Sleep Apnea were



ProCode	ProCode Name	Beta	Std. Error
226.4	Spine radiology exam	2.46	0.0938
193.5	Doppler echocardiography	1.33	0.0575
237.4	Clinical nutrition	1.45	0.0626
193.3	Transthoracic echocardiography	1.22	0.0532
212.4	Occupational therapy	1.23	0.0605

Figure IV-7 Sample PheDAS of CPT records in DS vs. IDD subjects. (a) A binary feature matrix with ProCodes as columns and subjects as rows was constructed from the CPT event records mapped to ProCodes in *pyProwasLookup*. (b) Mass univariate logistic regression was performed across ProCodes in the feature matrix using *pyProwasModel*; regression results are listed for the top 5 most significant ProCodes (p <<< 0.001 after Bonferroni multiple comparisons correction). (c) The Log Odds plot of top 18 most significant PheCodes (p <<< 0.001 after Bonferroni multiple comparisons correction) is shown, created via *pyProwasPlot.*

found to have positive associations with DS, all of which are known co-morbidities of DS [192] [193]. Similarly, the Log Odds plot in Figure IV-7 shows that the ProCodes for Echocardiography (ECG), Clinical Nutrition, Sleep Studies, and Physical Therapy were found to be significantly positively associated with DS; again, these ProCodes would be expected as they are procedures which could be used to diagnose and treat known co-morbidities of Down Syndrome [192].

With our expected associations established, the next task is identifying unknown or interesting associations in the PheDAS. The volcano plot may serve as a helpful guide in this step, since it provides an overview of all results and directly links statistical significance with effect size. When viewed via *pyPhewasPlot* and *pyProwasPlot*, zooming and panning functions allow users interactively identify results of interest. **Figure** IV-8 shows the volcano plots for both the ICD and CPT analyses described in the Results section; it should be noted that phenotype labels have been removed in this figure for legibility.

An alternative approach for interpreting PheDAS results is assessing the *novelty* of disease-phenotype associations in terms of existing literature. Previous work has presented a formal method for assessing this type of novelty in PheDAS [99]. In brief, a novelty score is calculated for each disease-phenotype

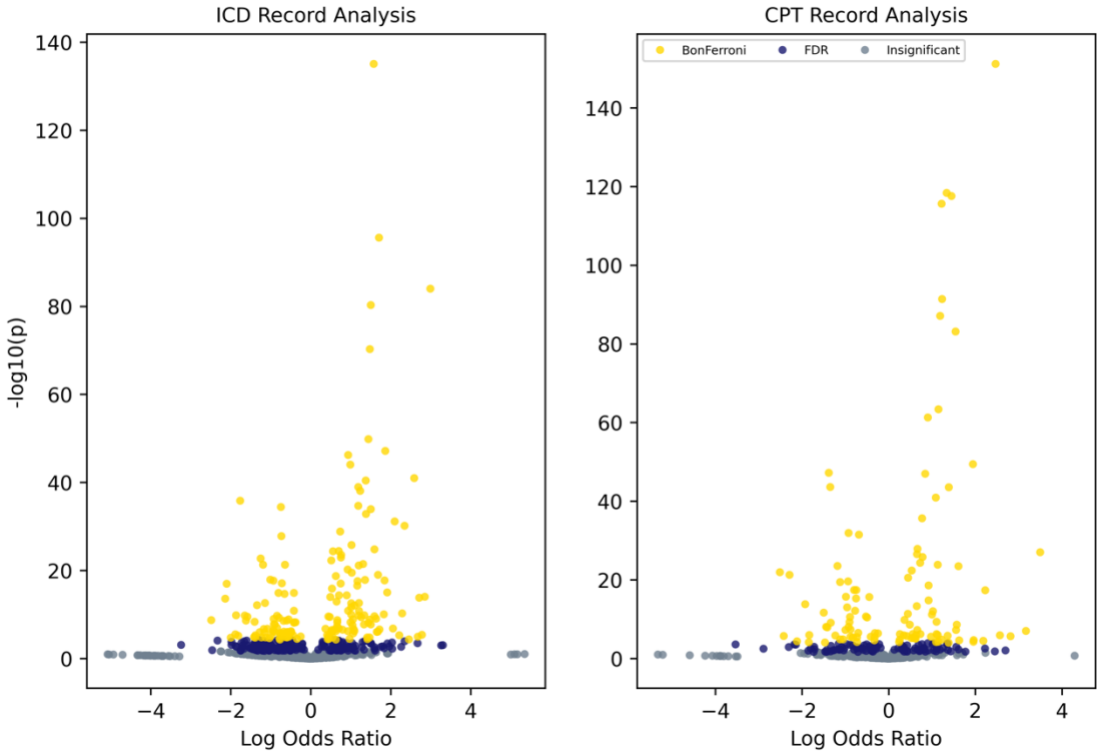


Figure IV-8 Sample volcano plots. Phenotype labels have been removed for legibility. Users may directly interact with these plots via pyPhewasPlot and pyProwasPlot. Zooming and panning across the plot enable users to explore phenotypes with regard to both significance and effect size. Thresholds for multiple comparisons correction are presented visually via color (Bonferroni in yellow, FDR in dark blue, and no significance in gray).

association in a PheDAS that measures the degree to which it is already known based on data mined from PubMed abstracts. If a disease-phenotype pairing is present in a large number of PubMed abstracts, the association is assigned a low novelty score and considered well known. In contrast, if a disease-phenotype pairing is present in only a few PubMed abstracts, the association is assigned a high novelty score and considered unknown. This framework is advantageous for exploratory studies in particular, as it does not require a clinical expert to manually review all results and filters the number of potentially novel or interesting PheDAS results down to a manageable amount. This novelty score framework is also available as part of the pyPheWAS package, though not covered in depth here.

We have shown that PheDAS methods are powerful in isolation, but several studies have also demonstrated their utility as support for other types of analyses. Warner et al performed a proof-of-concept study which employed the PheDAS framework in order to identify subjects for a trans-institutional cohort of multiple myeloma patients [177]. Li et al used PheWAS for hypothesis generation in the context of phenotypes related to the genetic components that drive serum uric acid level, then performed a conventional analysis to investigate causal relevance for the identified phenotypes [171]. In the realm of medical imaging, PheDAS has been used successfully to study diseases of the eye and optic nerve. In one such study, PheCode and ProCode feature matrices were used alongside imaging-derived features in a model of visual function for subjects with glaucoma and thyroid eye disease; inclusion of the EMR data was found to improve the explained variance of disease outcomes [50]. Another study used PheDAS to identify PheCodes associated with several optic nerve diseases, then used the identified phenotypes combined with optic nerve imaging features to classify disease subjects and controls. Again, combining the PheCode feature vectors with imaging-derived features produced the most accurate classifiers [48]. This framework could be extended to the domain of neuroimaging, allowing researchers to support their models of neurological disease with EMR context.

There are several limitations to keep in mind when working with EMR data and the pyPheWAS package. Inherent variability in EMR data is well documented [49]. For example, the ICD coding system's primary function is to bill insurance companies, not to serve as a proxy for diagnosis. ICD codes are generated by a coding specialist who translates clinician notes into insurance billing codes; this process has many opportunities for noise to enter the system, including at the patient-physician interface (patient-physician communication, physician training, expertise, and attention to detail), at the physician-coder interface (variations in clinical practices, coder training and expertise, facility quality assurance), and from simple human errors [49]. Additionally, EMRs suffer from broader issues of record fragmentation (such as when a patient moves between institutions) and a bias toward sicker populations (EMR events are usually recorded during illness) [54]. Some of this error may be mitigated while creating case and control groups

with the *createPhenotypeFile* tool. Users may specify a code frequency threshold which must be met for a subject to be considered a "true" case or control; enforcing higher temporal thresholds on ICD code events reduces the possibility that mis-coded subjects are mistakenly included in the case or control groups. Additionally, the mapping from ICD codes to PheCodes further reduces EMR variability by consolidating large groups of highly-related ICD codes into a single PheCode [168].

Another common challenge with large-scale association methods such as GWAS and PheDAS is confounding. Users have several options for addressing this issue within the pyPheWAS toolkit. The case-control matching tool, *maximizeControls*, allows users to match the distributions of potentially confounding variables, such as sex or age, between the case and control populations. Confounding variables may also be added as covariates in the mass univariate regression step; users may specify both primary variables (height or weight) and combined terms (height divided by weight) via the group file to control for various confounding effects. Furthermore, after completing a PheDAS experiment, users should carefully consider the verification of their results by identifying plausible biological links for identified associations and replicating their analysis in an independent population [194].

These strategies may be used to control for common confounding factors, but investigators should also carefully consider more subtle confounders that might influence their group composition. Individuals suffering from chronic diseases, for example, tend to have more hospital visits and therefore higher numbers of secondary medical diagnoses than individuals with acute ailments; because of this, comparing a chronic disease case group to an acute disease control group may result in false positive phenotype associations unrelated to the chronic disease of interest. This common but challenging scenario could be mitigated in several ways, such as including visit frequency as a matching criterion or redefining the control as a comparable chronic disease. Ultimately, it falls to the investigators using pyPheWAS to precisely select case and control group populations so that their study design properly addresses their specific research question.

A few additional limitations are related directly to the pyPheWAS toolkit. As was previously stated, the ICD-phenotype maps do not cover the full range of possible ICD codes; specifically, the map includes 15,558 ICD-9 codes and 9,505 ICD-10 codes [167], [168], [187]. Users are notified when their datasets contain ICD-9 and ICD-10 codes which are not in the mapping and may choose to save the excluded ICD events for inspection. Relatedly, the pyPheWAS map is limited to processing only ICD-9 and ICD-10 codes; newer coding systems such as ICD-11 are not yet supported. To work with an expanded set of ICD-9 and ICD-10 codes or to incorporate ICD-11, users may wish to use a custom phenotype map with pyPheWAS. Though this feature is currently not supported, pyPheWAS is an open source tool, allowing researchers to customize its functionality. To incorporate a custom phenotype map, users may clone the

pyPheWAS project from GitHub and replace the default map within the source code. This modification would require that the user first edit their custom map's headings to match the default map's headings, and then point the map loading function in the source code to their local custom map. In a similar vein, the pyPheWAS package currently performs only mass logistic regression. Other regression methods have proven interesting in PheDAS analyses, however; for example, one study used of a linear regression to study phenotypic associations with white blood cell count [178]. Again, though this feature is not currently supported, the open source nature of the pyPheWAS toolkit provides the opportunity for other researchers to build in new capabilities. The key modification required for a custom regression type would involve replacing the logistic regression in *pyPhewasModel* with an alternate regression model from the statsmodels python package [188] and specifying which output values to pull from the fitted model. An alternative statistical python package such as scikit-learn [195] may also be used, but would require more modifications to the modeling input and output structure. The pyPheWAS website contains more detailed directions for users wishing to implement either a custom phenotype map or regression modifications.

In this work, we have presented pyPheWAS, a command line toolkit for implementing PheDAS analyses. We have demonstrated a typical PheDAS analysis of children with Down Syndrome compared to children with other intellectual and developmental disorders, complete with suggestions for verifying and interpreting the large amount of statistically significant results. Whether on its own or in combination with other analyses, the pyPheWAS toolkit provides an approachable method for taking advantage of the EMR and integrating this rich resource into our studies of neurological disease.

Chapter V

pyPheWAS Explorer:
A Visualization Tool for Exploratory Analysis of Phenome-Disease Associations

1. Overview

To enable interactive visualization of phenome-disease association studies (PheDAS) on electronic health records (EHR). Current PheDAS technologies require familiarity with command-line interfaces and programming. pyPheWAS Explorer allows users to examine group variables, test assumptions, design PheDAS models, and evaluate results in a streamlined graphical interface. A cohort of attention deficit hyperactivity disorder (ADHD) subjects and matched non-ADHD controls is examined. pyPheWAS Explorer is used to build a PheDAS model including sex and deprivation index as covariates, and the Explorer's result visualization for this model reveals known ADHD comorbidities. pyPheWAS Explorer may be used to rapidly investigate potentially novel EHR associations with conditions such as ADHD. Broader applications include deployment for clinical experts and preliminary exploration tools for institutional EHR repositories. pyPheWAS Explorer provides a seamless graphical interface for designing, executing, and analyzing PheDAS experiments, emphasizing exploratory analysis of regression types and covariate selection.

2. Introduction

The past few decades have seen a surge in the availability of EHR data [196] and, unsurprisingly, numerous methods for making sense of this rich data source [197], [198]. When genetic data is available, phenome-wide association studies (PheWAS) are often used to identify associations between a genotype and many EHR phenotypes, often derived from ICD billing codes [47]. This technique has discovered novel associations between EHR phenotypes and HLA-DRB1*1501 [169] and determined the contribution of Neanderthal genetic variants to phenotypes of modern humans [173]. Inspired by PheWAS, PheDAS emerged to investigate associations between non-genetic targets and EHR-derived phenotypes. Such studies include a characterization of co-occurring phenotypes in autism spectrum disorder [29] and a scan for phenotype associations with white blood cell count in an intensive care unit cohort [178].

Currently, several tools exist for running these PheWAS and PheDAS experiments, including R PheWAS [180], pyPheWAS [199], and PHESANT [200]. All of these tools obscure the study data-flow, requiring the user to navigate several command-line-style tools (or write their own code) and only providing

visualizations for the final output. These conditions can make verification of model inputs and designs difficult, especially considering the unwieldy file sizes of larger cohorts. These factors present unnecessary barriers for researchers, especially those who are experimenting with PheWAS or PheDAS for the first time.

To bridge this accessibility gap, we present pyPheWAS Explorer, an interactive visualization tool for the analysis of ICD-derived phenotypes. Inspired by RegressionExplorer [201], pyPheWAS Explorer provides visual inspection of model inputs, real-time model building, and multi-faceted result visualization. The Explorer may be used for both PheWAS and PheDAS experiments, as the primary difference is whether or not the target variable is genetic. Given the limited availability of genetic data, however, we focus on PheDAS for the remainder of this article to avoid confusion and encourage more researchers to leverage this method even in the absence of genetic data.

3. Materials and Methods

The pyPheWAS Explorer workflow (Figure V-1) is composed of three phases: input and preprocessing, model building, and model evaluation. In the following sections, we describe each of these phases in detail, after briefly outlining PheDAS experiments in general.

3.1. A brief description of PheDAS

A PheDAS aims to identify associations between a non-genetic binary target and ICD-derived phenotype codes (PheCodes). These associations are found via the following procedure: 1) ICD data are mapped to PheCodes, 2) PheCodes are aggregated across each patient's record, 3) mass univariate logistic regression [202] is performed on each PheCode, and 4) the regression results are visualized for interpretation [199]. pyPheWAS Explorer performs steps 1 and 2 of this procedure in the background during the input and preprocessing phase, after which the user may interactively build and run the PheDAS logistic model (step 3). Finally, the user may interpret target-PheCode associations by directly interacting with both tabular and visual representations of the PheDAS model results (step 4).

3.2. Input and preprocessing

The first phase of the Explorer workflow involves transforming the longitudinal EHR into PheCode feature matrices. Two data files are required: an EHR file and a group demographics file. The EHR file contains ICD records, while the group file contains the target variable (disease status) and other group

variables (in this example, Body Mass Index (BMI) and sex) (Figure V-1). pyPheWAS Explorer first maps all ICD-9 and ICD-10 codes to a set of 1,866 PheCodes [168], [187]. It is important to note that while these tables are extensive, they are incomplete; any ICD-9 or ICD-10 codes not included in the mapping are removed from the study. The Explorer then uses three different aggregation methods to summarize the existence of each PheCode across each patient's EHR. The *binary* aggregation method considers the relationship between the target variable and the presence of each PheCode; this feature matrix contains only zeros (the PheCode was absent in the patient's record) and ones (the PheCode was present in the patient's record). The *count* aggregation method considers the relationship between the target variable and the number of occurrences of each PheCode; this feature matrix contains positive integers corresponding to the

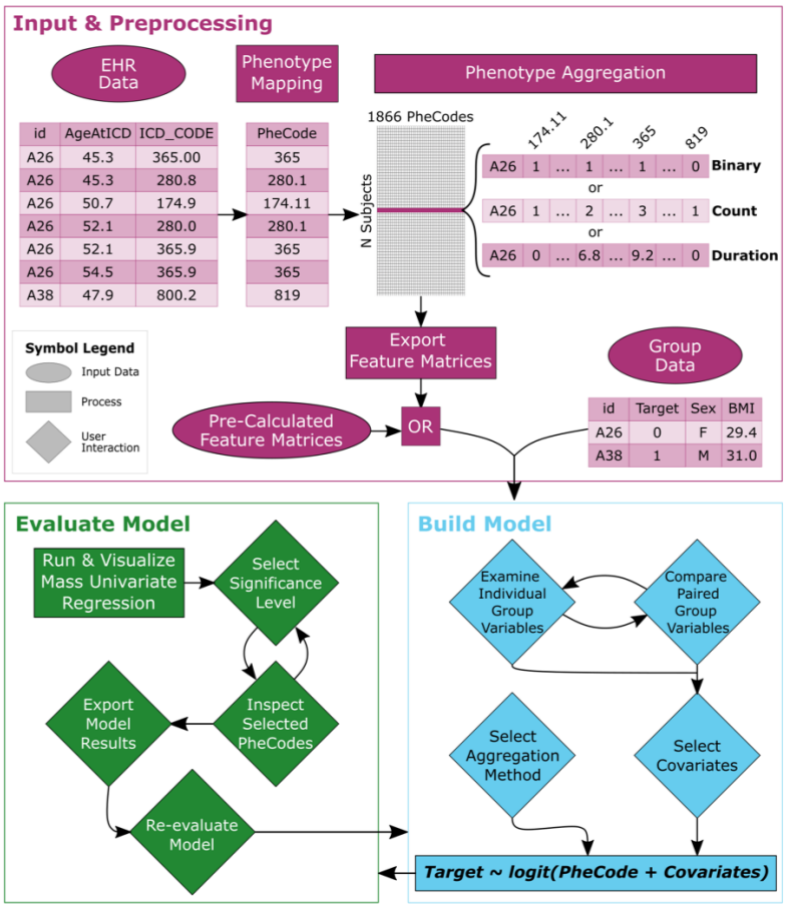


Figure V-1 pyPheWAS Explorer Workflow. All data preprocessing is done automatically in the background; feature matrices are saved for faster startup in subsequent sessions. In the model building phase, the user may examine group variables and compare them to each other before adding them to the PheDAS model. Additionally, users may specify the type of PheCode aggregation (binary, count, or duration). In the model evaluation phase, the user examines mass univariate regression results at configurable significance levels. Based on these results, the user may move back into the model building phase to re-evaluate their model design.

total number of instances of each PheCode in each patient's record. The *duration* aggregation method considers the relationship between the target variable and the span of time over which a PheCode was present; this feature matrix contains the time in years between the first and last occurrences of each PheCode in each patient's record.

3.3. Building a PheDAS model

Model building in pyPheWAS Explorer is facilitated by the interactive panel (Figure V-2). For each group variable γ, the individual view presents the correlation coefficient between γ and the target variable (represented by a colored block) and a histogram of γ values (separated by target group). The overlapping histogram allows the user to check case/control matching for each potential covariate, while the correlation coefficient allows the user to identify potential covariate biases.

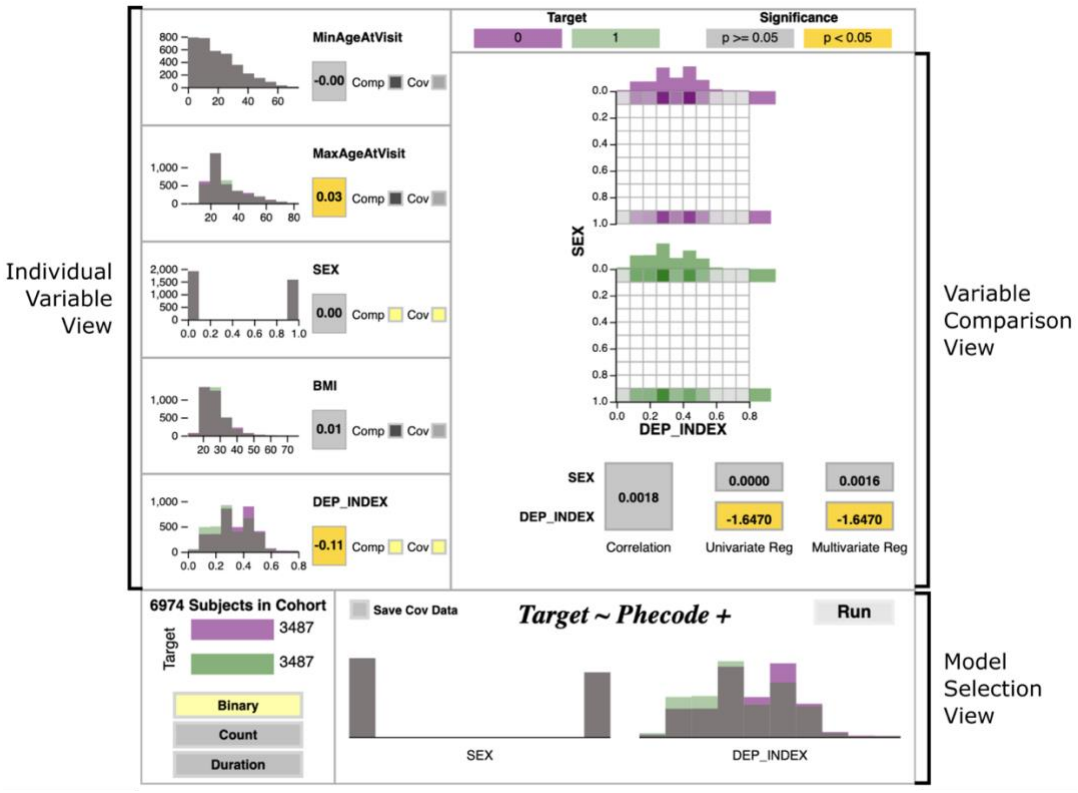


Figure V-2 pyPheWAS Explorer Regression Builder Panel. For demonstration, a cohort of ADHD cases (Target 1) and non-ADHD controls (Target 0) is shown. Group variables in this dataset included minimum/maximum age at visit (MinAgeAtVisit/MaxAgeAtVisit), biological sex, body mass index (BMI), and deprivation index (DEP_INDEX). The right side of this panel shows the variables sex and deprivation index loaded into the variable comparison view, while the model selection view shows both variables added to a binary PheDAS model. Color encodings for the case and control groups, correlations, and regression coefficients are shown along the top bar.

The variable comparison view allows the user to test the independence assumption for a pair of covariates before adding them to the logistic model. Variables are added to the comparison by selecting the *Comp* button from the individual variable panels. For a qualitative independence assessment, the joint distribution of these variables is provided, again separated by target group; this includes a grid that captures the overlap of selected variable distributions, along with the individual histograms for each variable at the top and right of the grid. Hovering over the joint distribution grid allows the user to query the value of each bin. A quantitative assessment is also provided via two statistical tests. The first of these is the correlation coefficient between the selected variables. The second is a multicollinearity test, wherein the target variable is regressed as a function of each variable individually and by both variables together. The coefficients calculated from the correlation and regressions are all overlaid on colored blocks, where the color indicates statistical significance. In general, if the two variables are not correlated and their regression coefficients remain constant across the individual and combined multicollinearity models, the independence assumption holds, and they may be safely included in the PheDAS model together [203].

The model selection view allows users to build their PheDAS model. A list of buttons in this panel allows the user to select a PheCode aggregation type, while the *Cov* button in the individual group variable view allows the user to add covariates.

3.4. Evaluating a PheDAS model

Selecting the *Run* button in the model building panel triggers a real-time estimation of the user's model; the results of this estimation are automatically displayed on the evaluation panel in three linked views (Figure V-3). Selecting a data point in any of these views highlights the corresponding data point in the other two views.

The volcano plot presents an overview of the entire experiment. PheCode-target association significance, represented by -log(p-value), is shown on the y-axis, and association effect size, represented by log(odds ratio), is shown on the x-axis. Marker color in this plot corresponds to which multiple comparisons correction threshold a PheCode exceeds (Bonferroni [204], FDR [205], or insignificant). This view serves as a starting point for deeper investigations, as users can see interesting PheCode relationships in a single glance.

The effect size plot presents PheCode-target association effect size (with confidence interval) on the x-axis, with PheCodes listed down the y-axis. Only PheCodes that exceed the user-selected multiple comparisons correction significance threshold (either FDR [205] or Bonferroni [204]) are included in this plot. Marker color and shape in the effect size plot corresponds to 18 unique PheCode categories.

Finally, the data table provides the most detailed view, listing each PheCode's category, odds ratio, p-value, and the number of subjects that have at least one record of that PheCode. These tabular results are sorted so that the most significant results are at the top of the table. This data table is automatically saved so that users may reference it after closing pyPheWAS Explorer.

3.5. Installation and Use

pyPheWAS Explorer is available in open source as part of the pyPheWAS Python package [199], available at https://github.com/MASILab/pyPheWAS. An installation and usage video is included in the online documentation.

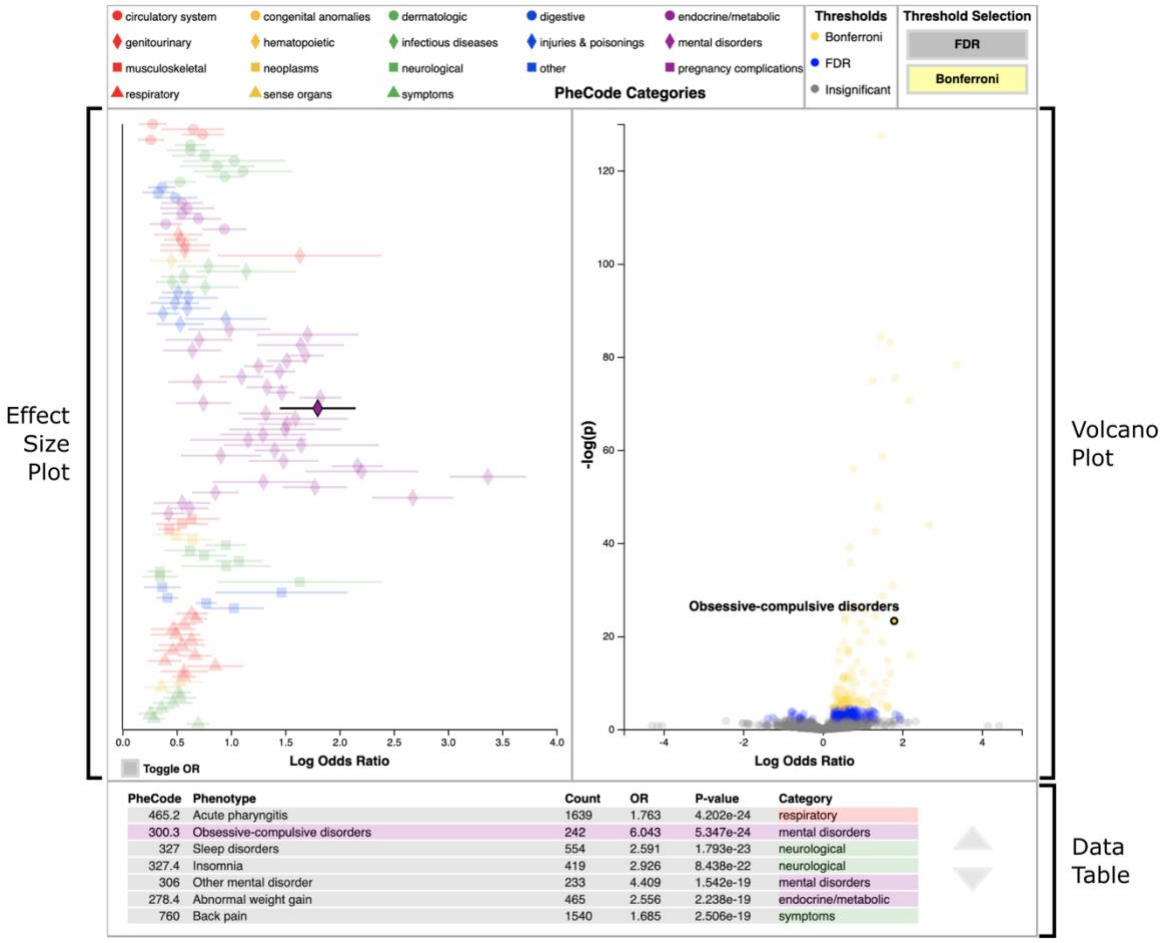


Figure V-3 pyPheWAS Explorer Regression Evaluation Panel. PheDAS results from the binary ADHD model are shown in three linked views: an effect size plot, volcano plot, and data table. Selecting a PheCode in any view highlights it in the other two views; PheCode 300.3, Obsessive-compulsive disorders, is selected for demonstration. The significance threshold for the effect size plot may be toggled between FDR and Bonferroni multiple comparisons correction by selecting the corresponding buttons at the top of the panel; here, Bonferroni is applied. Color legends for the effect size plot (PheCode categories) and volcano plot (significance thresholds) are shown along the top bar.

3.6. Software evaluation

To evaluate the Explorer software, we conduct an exploratory analysis of attention deficit hyperactivity disorder (ADHD) subjects compared to matched controls. A de-identified EHR dataset was acquired from the Synthetic Derivative at Vanderbilt University Medical Center [161]. 3,487 ADHD subjects were identified as those with at least 3 records of ICD-9 code 314.01 or ICD-10 codes F90, F90.0, F90.1, F90.2, F90.8, or F90.9. ADHD subjects were matched one-to-one with non-ADHD controls based on biological sex and minimum age at visit (± 0.1 years).

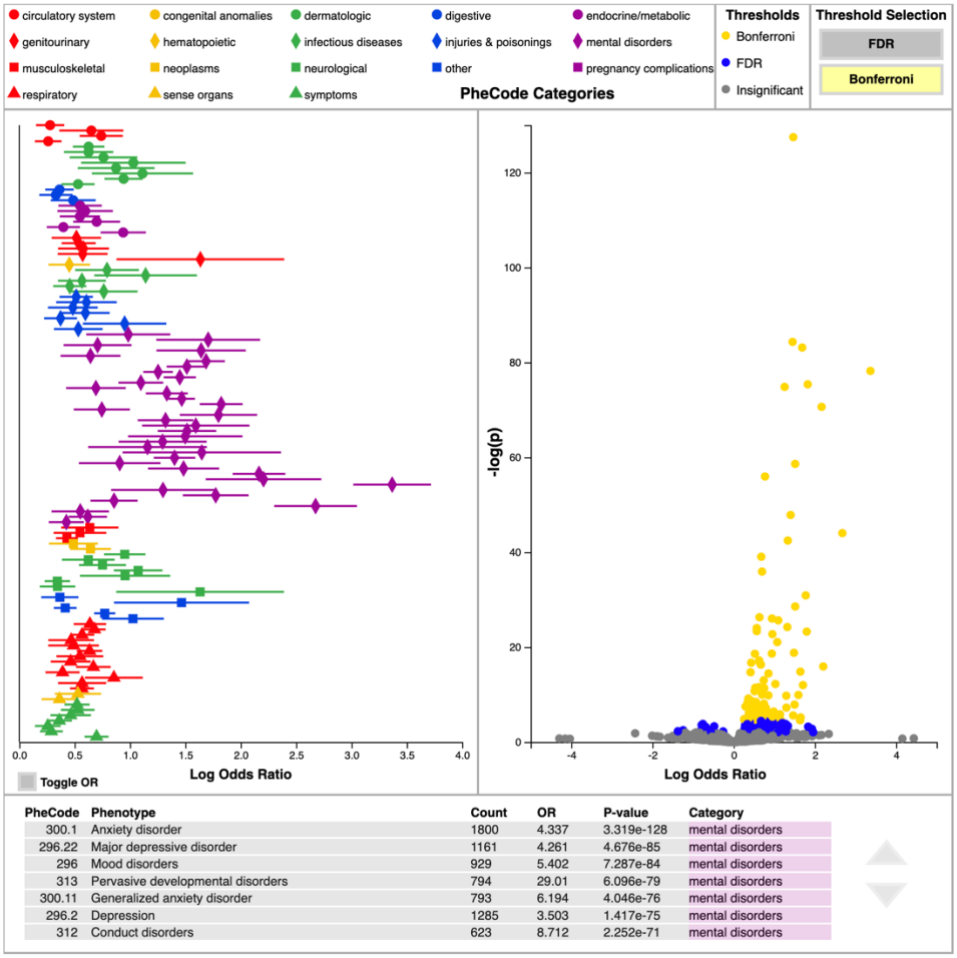


Figure V-4 pyPheWAS Explorer Regression Evaluation panel without a selected PheCode. PheDAS results from the binary ADHD model (Figure V-2) are shown without any PheCodes highlighted and with Bonferroni multiple comparisons correction applied to the effect size plot. There are many significant ADHD-PheCode associations, a large portion of which fall into the "mental disorders" category. This is clear as well from the data table; the top seven most significant associations are listed, all of which are categorized as "mental disorders".

4. Results

Figure V-2 shows the ADHD cohort loaded into pyPheWAS Explorer's regression builder interface. From the individual variable views, we confirm that the case and control groups were matched on sex and minimum visit age due to their histograms' perfect overlap, while the other group variables show diverging distributions. Additionally, we see that most variables are uncorrelated with ADHD, but interestingly, deprivation index (DEP_INDEX), a measure of socio-economic status where a higher value corresponds to higher deprivation [206], has a slightly negative correlation (i.e. ADHD subjects tend to be less deprived than controls). Based on this, we are interested in building a PheDAS model that includes both deprivation index and sex as covariates. To ensure that these covariates may be used together safely, we first examine them using the variable comparison view. The two variables are not highly correlated, and their regression coefficients remain constant across the individual and combined multicollinearity tests; so we conclude that they are sufficiently independent and add them to our model. Finally, we select the *binary* model type and compute our PheDAS model.

We use the volcano plot (Figure V-3, Figure V-4) to examine PheCode associations with the highest combined effect and significance; these include anxiety disorder, mood disorders, learning disorder, and obsessive-compulsive disorders, all of which are known comorbidities of ADHD [207]–[209]. We select Bonferroni correction for the effect size plot and find that all significant results had positive effects, with many in the "mental disorders" category.

5. Discussion

As shown by the ADHD case study, PheDAS studies typically excel in confirming known comorbidities; more interesting, however, are the less prominent PheCode associations that a PheDAS can reveal: "dermatologic" (acne; rash and other nonspecific skin eruption), "endocrine/metabolic" (abnormal weight gain), and "respiratory" (bronchitis; acute sinusitis). To investigate these further, we change the PheCode aggregation type to *duration* and re-compute our PheDAS model (Figure V-5); each of these "interesting" PheCodes remain in the effect size plot. This re-evaluation suggests that these three PheCode categories may be candidates for deeper study, as they presented significant effects in both EHR presence and duration and are potentially less known to be associated with ADHD.

pyPheWAS Explorer is a straightforward visual interface that captures a comprehensive summary of the entire PheDAS experiment, making rapid prototyping and interpretation of PheDAS models possible. The Explorer does currently have several limitations. Preparing EHR for the Explorer is not trivial. The full pyPheWAS package contains some data preparation tools, but these require familiarity with a command-

line, as does installing and launching pyPheWAS Explorer. Though the presented ADHD use case demonstrates the functionality of pyPheWAS Explorer, the efficacy of the tool in practice should be further tested via a user-centered evaluation, though this is out of scope for the current article.

Despite these limitations, several areas of opportunity exist for the Explorer to enhance EHR analysis. If packaged into a standard application, pyPheWAS Explorer could enable non-technical clinical experts to easily interact with PheDAS models and identify potentially novel disease associations. Similarly, institutional EHR repositories may benefit from deploying pyPheWAS Explorer as a data exploration and hypothesis generation tool for researchers building EHR analysis cohorts.

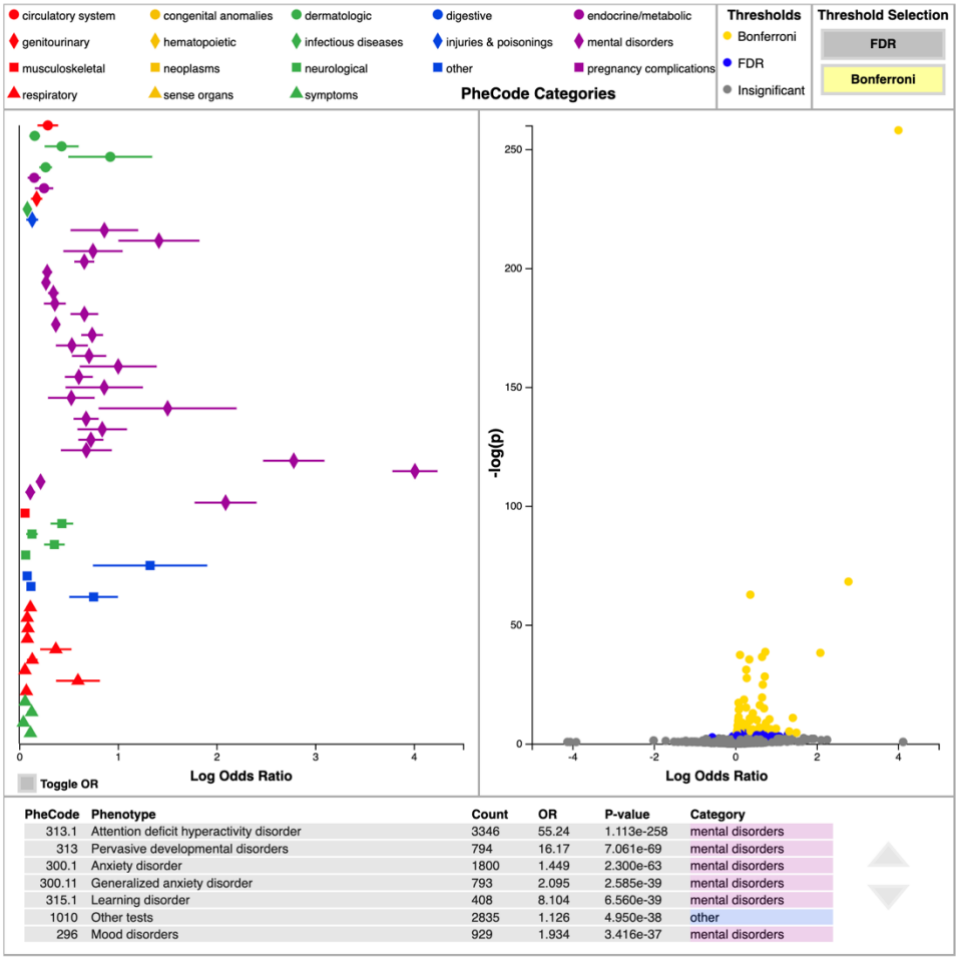


Figure V-5 PheDAS results from the duration ADHD model are shown in the pyPheWAS Explorer Regression Evaluation panel with Bonferroni multiple comparisons correction applied to the effect size plot. Compared to the binary model, there are overall fewer significant PheCode associations. Again, the "mental disorders" category is the most prominent. Many of the PheCode associations from the "interesting" categories in the binary model ("dermatologic", "endocrine/metabolic", and "respiratory") are also significant in this model.

6. Conclusion

pyPheWAS Explorer is a comprehensive tool for exploratory analyses of new EHR datasets. The interactive workflow enables users to quickly answer questions about a dataset's demographic space and potential for novel phenotypic signatures. We hope that pyPheWAS Explorer's approachable interface and comprehensive visualizations will empower a broader range of users to delve into the intriguing domain of EHR data.

Chapter VI

Phenotyping Down Syndrome: Discovery and Predictive Modeling with Electronic Medical Records

1. Overview

Individuals with Down syndrome (DS) are reported to have a heightened risk for various co-occurring health conditions, including congenital heart disease (CHD). In this two-part study, electronic medical records (EMRs) were leveraged in **Study 1** to examine as-yet unidentified co-occurring health conditions among individuals with DS; and in **Study 2**, to investigate health conditions liked to surgical intervention among DS cases with CHD.

De-identified EMRs were acquired from the Synthetic Derivative at Vanderbilt University and facilitated creating a cohort of N = 2,282 DS cases (55% females), along with comparison groups for each study. In **Study 1**, DS cases were one-by-two sex- and age-matched with samples of case-controls and of individuals with *other* intellectual and developmental differences (IDDs). The phenome-wide association study (PheDAS) strategy was employed to reveal co-occurring health conditions in DS versus comparison groups, which were then ranked for how often they are discussed, or as-yet unidentified, in relation to DS using the PubMed database and *Novelty Finding Index*. In **Study 2**, a subset of DS individuals with CHD (N = 1098 [48%]) were identified to create longitudinal data for N = 204 cases with surgical intervention [19%] versus 204 case-controls. Data were included in predictive models and assessing which model-based health conditions, when enriched, would increase the likelihood of surgical intervention.

In **Study 1**, relative to case-controls and those with *other* IDDs, co-occurring health conditions among individuals with DS were confirmed to include, e.g., hypertension, atrioventricular block, and heart failure (circulatory); hypothyroidism (endocrine/metabolic); and sleep apnea and Alzheimer's (neurological/mental). Findings with high *Novelty Finding Index* were pacemaker in situ, atrioventricular block, and valve stenosis (circulatory); gastritis (digestive); and elevated C-reactive protein (symptoms). In **Study 2**, among individuals with DS and CHD, model-based explanatory health conditions revealed, e.g., congestive heart failure (circulatory), valvular heart disease and cardiac shunt (congenital), and pleural effusion and pulmonary collapse (respiratory) linked to the likelihood of surgical intervention. Research efforts using EMRs and rigorous methods, including our study, could shed lights on the complexity in health profile among individuals with DS and other IDDs and motivate precision-care development.

2. Background

Down syndrome (DS; or trisomy 21) is a genetic condition characterized by the presence of an extra copy of chromosome 21 [97]. Individuals with DS are reported to have a heightened risk for various co-occurring intellectual, developmental, and health differences, including congenital heart disease [98], [210]. Greater and more in-depth understanding of this range of co-occurring health conditions could facilitate promising opportunities for both precision-medicine development as well as characterization of unmet clinical needs among individuals with DS. This two-part study leveraged an extant dataset of EMRs, in ***Study 1***, to investigate as-yet unidentified co-occurring health conditions among individuals with DS; and in ***Study 2***, to evaluate longitudinal predictors of known clinical outcomes in a subset of individuals with DS and CHD (see Table VI-1 for a list of terms and respective acronyms used repeatedly throughout this study).

Table VI-1 List of acronyms and terms used repeatedly throughout this study

Acronym	Term
DS	Down syndrome
IDD	Intellectual and developmental difficulties
CHD	Congenital heart disease
EMR	Electronic medical record
PheDAS	Phenome-disease association study
Phecode	Phenotype (or diagnostic) code
NFI	Novelty Finding Index
ICD	International Classification of Diseases

Aside from its substantial association with CHD and other circulatory and congenital (cardiac) differences, DS has been linked to a range of co-occurring health conditions in, e.g., metabolic and endocrine, neurological and sensory, as well as respiratory systems. For example, hypothyroidism is among the most common co-occurring endocrine conditions among individuals with DS [211]. A malfunctioning thyroid gland reportedly leads to hypothyroidism because it complicates a number of metabolic activities (e.g., protein synthesis, energy conservation, hormone regulation) as seen among many individuals with DS [212]. In terms of neurological and sensory conditions, increased prevalence of hearing difficulties and sleep apnea have been observed in DS [213]. Similar to their typically developing counterparts, increased hearing difficulties among individuals with DS have been linked to lower proficiency in understanding spoken language and auditory (phonological) memory [214], [215]. Sleep apnea has been similarly shown as highly prevalent in DS, as it also relates to cardiac (e.g., hypertension), respiratory (breathing failure), and other mental and neurological conditions (depression, ADHD) [216]. Scant evidence, however, reports

these various health conditions in concert, as it is rare to obtain large cohorts and comprehensively study associated clinical features among individuals with DS. The current study leveraged an extant EMR dataset to address this issue and, at the same, reveal co-occurring health conditions potentially unrecognized in limited sample sizes of individuals with DS.

Research efforts using de-identified EMR are gaining traction, given that such an approach enables large sample sizes and sufficient creation of comparison groups to evaluate co-occurring health conditions among individuals with intellectual and developmental difficulties (IDD) (e.g., in autism [29]). Particularly relevant to the current study, Alexander and colleagues [211] employed this strategy of analyzing retrospective EMRs which contain de-identified clinical features (e.g., diagnostic codes) among 6,430 individuals with DS from the United Kingdom Clinical Practice Research Datalink. Recently, Valentini and colleagues [217] gathered a one-year subset of EMRs from 763 individuals with DS after carefully matching for age and sex. In addition to corroborating previous findings in small sample sizes, these candidate EMR studies in DS offer a nuanced examination into the prevalence and incidence rates across a range of co-occurring health conditions concertedly. For instance, individuals with DS were shown to exhibit heightened prevalence for, e.g., CHD (cardiac), dementia (mental), and hypothyroidism (metabolic/endocrine) conditions, along with interesting results on lower incidence of anxiety and depression (mental) in DS relative to case-controls. While promising, the repertoire of methodologically sound tools is in its infancy as to comprehensively analyze these data and capture what health conditions may have been unrecognized in smaller samples. Such tools, when combined with and interrogating the complexity of EMRs, could underscore individual differences across multiple clinical features.

Capitalizing on the rich EMR dataset, in ***Study 1***, two nuanced strategies were used to comprehensively analyze clinical features and capture as-yet unidentified co-occurring health conditions among individuals with DS. First, the phenome-disease association study, or *PheDAS*, strategy was employed to investigate phenotypic differences between individuals with and without DS. Derived from phenome-wide association study (PheWAS) and the long-standing genome-wide association study (GWAS), PheDAS has been demonstrated to have the propensity of assess the "whole phenome" [48], [218]. The application of PheDAS enabled a method of revealing a range of co-occurring health conditions in DS using "phenotype" codes, or *Phecodes*, derived from EMRs [47], [48], [219]. Findings from this approach were confirmatory with previous works using EMRs that aimed to evaluate clinical features in DS and contribute to precision-medicine development. Departing from other similar analyses, our study aimed to characterize within these PheDAS findings which co-occurring health conditions were either not yet identified among individuals with DS or less discussed in the relevant literature. For this second goal, the *Novelty Finding Index* (NFI) strategy was employed since it was developed to assess the level of statistical and clinical significance

among health conditions found in PheDAS and other EMR studies [99]. *NFI* ranked the co-occurring health conditions found in relation to DS by how "well studied" they were in the PubMed database. Less "well studied" (or more novel) co-occurring health conditions could provide insights into potentially unmet health needs among individuals with DS.

Knowing the likelihood of various co-occurring health conditions in DS presents an incomplete story; some health conditions may drive the severity of medical needs in other clinical features. For instance, cardiac malformations both predispose individuals with DS to CHD, which affects over 40% of these individuals, and are a major cause of mortality in DS [210], [220]. CHD in DS has been shown to be pervasive as it could relate to the accelerated severity of other cardiac and circulatory conditions (e.g., pulmonary arterial hypertension [221]). More than 20% of individuals with DS and CHD have been reported to undergo surgical intervention to correct these cardiac malformations [222], [223]. Because of this, it remains crucial to achieve a greater understanding of which health conditions may co-occur with CHD and DS, and how those conditions may increase the likelihood of surgical intervention. In ***Study 2***, two specific steps were taken to evaluate longitudinal predictors, or health conditions, of known outcomes associated with DS and CHD. First, diagnostic codes drawn from EMRs were used to build a predictive model which estimated the likelihood of surgical intervention among individuals with DS and CHD. Second, the predictive model was used to characterize the relative importance of different diagnostic codes in the predicted likelihood of surgical intervention. These findings and methodological strategies could be crucial to precision-care development as well as maximization of treatment effectiveness for individuals with DS.

In this two-part study, EMRs were leveraged to investigate (***Study 1***) as-yet unidentified co-occurring health conditions and (***Study 2***) explanatory health conditions in a known outcome among individuals with DS. Specific goals and strategies are as follows. Extant EMR data facilitated creating comparison groups of sufficient sizes, including a cohort of individuals with DS who were then one-by-two sex- and age-matched with a sample of case-controls. To carefully assess the clinical features that may uniquely associate with DS versus *other* IDDs (e.g., autism, ADHD, dyslexia), a matched group of individuals with *other* IDDs was also created for comparison of co-occurring health conditions in relation to DS cases. In ***Study 1***, the PheDAS approach was applied to reveal co-occurring health conditions in DS versus case-controls to confirm previous findings, and then in DS versus *other* IDDs. Then, the found co-occurring health conditions were ranked for their clinical relevance and how often they are discussed in relation to DS using the PubMed database and *NFI*; these steps highlighted which clinical features were as-yet unidentified. In ***Study 2***, individuals with CHD were identified from the initial cohort of DS cases. EMR data were used to

build predictive models and evaluate the extent to which different health conditions longitudinally contribute to the likelihood of receiving heart-related surgical intervention.

3. Methods

The schematic overview of this two-part study design can be found in Figure VI-1. Specific steps included data acquisition and cleaning, sampling and creating comparison groups, and converting clinical features from EMRs to analyzable data (e.g., phecodes). These exhaustive steps mined and prepared analyzable data from the rich EMRs, which enabled this study to then pursue nuanced strategies proposed in ***Studies 1*** and ***2*** and evaluate our specific questions.

This study was approved by and its procedures were carried out in accordance with the Internal Review Board of Vanderbilt University and Vanderbilt University Medical Center. N = 1,025,321 de-identified EMRs were acquired from the Synthetic Derivative at Vanderbilt University Medical Center. From the initial repository, EMRs were excluded based on the following criteria: (a) containing no information on International Classification of Diseases (Ninth and Tenth Revisions [ICD-9 and ICD-10]); (b) invalid age, date-shifting and anonymization errors; and (c) invalid or unknown biological sex.

3.1. Study 1: Characterizing as-yet unidentified co-occurring health conditions in DS

In ***Study 1***, we aimed to characterize as-yet identified health conditions that co-occur with DS. This was done by first identifying DS patients in the EMRs, along with suitable control groups. A PheDAS was then performed on the EMRs, revealing which health conditions co-occur with DS. Finally, the PheDAS output was examined via a novelty analysis to identify which co-occurring conditions were not well-characterized yet. The following sections explain each of these steps in detail.

3.1.1. Cohort selection

Individuals with DS were identified as those who had *at least* two instances of an ICD-9 or ICD-10 code for DS: 758.0, Q90.0 (meiotic nondisjunction), Q90.1 (miotic nondisjunction), Q90.2 (translocation), and/or Q90.9 (others/unspecified). This process yielded N = 2,282 DS cases. Extant data were also leveraged in the creation of comparison groups, which consisted of case-controls and of individuals with *other* IDDs. The case-control (or "typically developing") group was composed of individuals *without* any records of DS or other IDD ICD codes; this group was intended to replicate and confirm previous findings in our sample. Of particular and unique interest to this study is the comparison of co-occurring health conditions in DS cases versus individuals with *other* IDDs; this group was composed of individuals with at

least two records of a given ICD code for other IDDs (see Appendix A). Both the case-control and IDD groups were matched two-to-one with the DS individuals based on biological sex, minimum age at visit (± 0.5 years), and maximum age at visit (± 1.9 years). This process resulted in N = 4,565 subjects in the case-control group, and N = 4,565 subjects in the other IDD group.

3.1.2. Phenome-disease association study (PheDAS)

PheDAS was performed to compare both the DS group to the case-control and the IDD groups. Briefly, this analysis (as shown in Figure VI-1) involved first mapping ICD-9 and ICD-10 codes to a set of 1,866 phecodes [168], [187]. Next, phecode instances were aggregated across each subject's record, resulting in

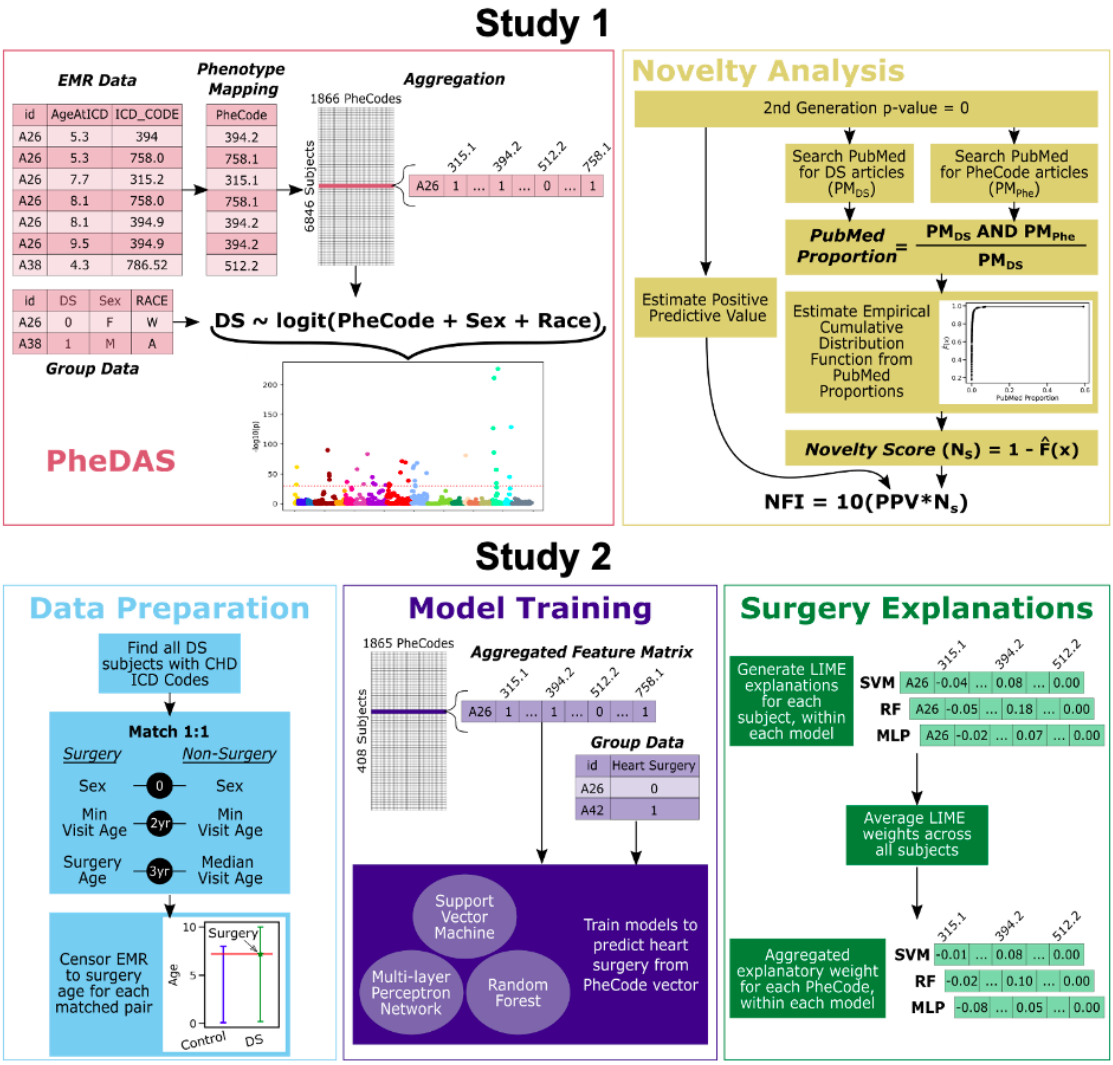


Figure VI-1 Overall two-part study design and flow charts for Studies 1 and 2.

a 1x1,866 binary vector for each subject; phecodes present in the record were represented by a 1, while phecodes not present in the record were represented by a 0. A mass univariate logistic regression was then performed for each phecode such that the DS status (whether/not a subject was in the DS group) was modeled as a function of the aggregated phecode, biological sex and race. Bonferroni correction was applied to the regression results to adjust for multiple comparisons. All PheDAS functions were performed via the *pyPheWAS* Python package [224].

3.1.3. Novelty analysis

We next determined which DS-phecode associations were as-yet identified by calculating the *Novelty Finding Index* (NFI) for each PheDAS result. Introduced by Chaganti and colleagues [99], the *NFI* aims to assist researchers in evaluating the extent to which empirical relationships are both clinically meaningful and well-studied in scientific literature. It incorporates second generation p-values to identify phecode associations that are both clinically and statistically meaningful [225], positive predictive value (PPV) to estimate phecode association reliability, and PubMed search results to evaluate phecode associations' literary novelty.

Briefly, the *NFI* is computed by first calculating the second-generation *p*-value for each phecode; in this study, second generation *p*-values were calculated using an odds ratio null interval of [0.3, 1.5]. Phecodes with a second-generation *p*-value of 0 are considered clinically interesting and statistically significant, while all others are removed from the analysis. Next, the PPV is estimated for each phecode using an empirical Bayes approach. The *PubMed proportion*, or proportion of published articles that mention the phecode-disease pairing out of all PubMed articles that mention the disease, is then calculated for each phecode. From this proportion, a novelty score, N_S, is calculated as $N_S = 1 - \hat{F}(x)$, where $\hat{F}(x)$ is the empirical cumulative distribution function estimated from all phecode PubMed proportions. Finally, the *NFI* may be computed: $NFI = (PPV \cdot N_S) \cdot 10$. This index provides a ranking that accounts for both the reliability of phecode-disease associations and their relative novelty. Again, all novelty analyses were performed via the *pyPheWAS* [Python] package [224].

3.2. Study 2: Congenital heart disease and surgical needs in DS

Study 2 aimed to determine longitudinal EMR predictors of surgical interventions for DS subjects diagnosed with CHD. To this end, we first identified which DS subjects from *Study 1* also had CHD, and from the CHD subset, which subjects did or did not receive heart surgery. The EMRs of these subjects were then censored to seven days before the first heart surgery record to isolate only events that occurred before

surgery. Three classifiers, including *Support Vector Machine*, *Random Forest*, and *Multi-layer Perceptron*, were subsequently trained to predict whether or not a subject would receive surgery based only on the pre-surgical records. Finally, a black-box explainability technique was used on the best-performing trained classifier to determine the importance of each phecode toward surgery predictions. This process is described more fully in the following sections.

3.2.1. Cohort selection

Out of the 2,282 DS subjects from *Study 1*, subjects with the additional diagnosis of CHD were identified as those who had at least one instance of a phecode for CHD: 747, 747.1, 747.13, and 747.2. This resulted in 1,098 subjects in the combined DS+CHD group. From this group, 204 surgery subjects were identified as those with at least one record of the phecode for heart transplant/surgery (429.1). Surgery subjects were matched one-to-one with non-surgery DS+CHD controls based on three matching criteria. The first two included biological sex and minimum age at visit (± 2 years). The last criterion matched surgery subjects' age at surgery to control subjects' median visit age (± 3 years); this final criterion ensured that matched controls had EMR events around the same time as their match's surgery, forcing their EMRs to be approximately the same length up until the point of surgery. Finally, the EMRs of all subjects were censored to only records before seven days prior to surgery (control EMRs were censored to their matched surgery subject's age at surgery). The seven-day buffer was added leading up to surgery to prevent EMR contamination around the time of surgery due to reporting delays. This selection process yielded a final cohort of 408 DS+CHD subjects, evenly split between surgery and non-surgery.

3.2.2. Longitudinal predictive modeling

In the same way as *Study 1*, the first step in analyzing the EMRs was converting ICD-9 and ICD-10 codes to phecodes and aggregating those phecode instances across each subject's record. This resulted in a 1x1,866 binary vector for each subject, where again, phecodes present in the record were represented by a 1, while phecodes not present in the record were represented by a 0. Finally, the phecode for heart transplant/surgery (429.1) was removed, yielding a 1x1,865 binary phecode vector describing the EMR fingerprint leading up to surgery for each subject in the DS+CHD group.

Three different classifier models were next trained to predict whether or not a given subject would receive heart surgery based only on the subject's 1x1,865 binary phecode vector; these models included *Support Vector Machine*, *Random Forest*, and *Multi-layer Perceptron*. Four-fold cross validation was used to optimize model and training parameters for all model types. The optimal *Support Vector Machine* model

used a sigmoid kernel and class weightings (0.45 for the non-surgery class, 0.55 for the surgery class). The optimal *Random Forest* model included 60 estimators, employed minimal cost-complexity pruning with $\alpha = 0.007$, and used entropy as the split criterion. Finally, the optimal *Multi-layer Perceptron* model employed rectified linear unit activation, a stochastic gradient decent optimizer, and two hidden network layers (a 100-neuron layer followed by a 50-neuron layer). All models were trained using the *scikit-learn* [Python] package; any model parameters not explicitly mentioned in this article were set to the scikit-learn default values [83].

These optimized parameters were next used to train secondary models of each type, now on the full dataset; 80% of the DS+CHD cohort (326 subjects) were used for training, and the remaining 20% (82 subjects) were used for testing. These secondary models were trained using the full dataset in order to leverage as much information as possible in the subsequent explanatory phecode analysis.

3.2.3. Model-based feature importance and explanatory variables

We next used LIME, an explainability technique for machine learning models, to investigate which phecodes were most important in predictions of surgery. LIME stands for *local interpretable model-agnostic explanations*; in short, it is a method of generating human-interpretable explanations for the predictions of any machine learning model [68]. For a given input X, LIME perturbs the input data and monitors how the perturbations modify the model's prediction. LIME then generates an explanation for the model's prediction based on X, consisting of a weight for each input feature. For the purposes of this study, LIME perturbs the binary phecode vector for an individual subject and monitors how this modifies the surgery prediction. It then generates an explanation for that subject, consisting of a weight for each phecode. The weight, w_{phe}, for a phecode, *phe*, may be interpreted in the following way: on average, the presence of *phe* in the subject's record increases the probability of a surgery prediction by w_{phe}. Such explanations were generated for all 408 subjects in the DS+CHD cohort. Explanatory weights were then averaged across subjects, yielding a single average explanatory weight for each phecode. This procedure was repeated for each classification model type, resulting in three separate sets of explanatory phecode weights. Finally, these weights were considered in relation their prevalence in the surgery cohort; findings were reported only for health conditions that appear in at least 40% of the subjects who received surgery.

4. Results

Extant EMR data enabled a large sample of N = 2,282 individuals with DS, as well as the creation of comparison groups (as specified below) with stringent criteria, to be included in this two-part study. ***Study***

1 comprehensively analyzed a range of co-occurring health conditions and captured as-yet unidentified ones among individuals in DS. ***Study 2*** evaluated health conditions that longitudinally predict known clinical outcomes by looking at surgical intervention for CHD in a subset of individuals with DS.

4.1. Characterizing as-yet unidentified co-occurring health conditions in DS

Previous studies have compared clinical features among individuals with DS relative to typically developing case-controls. No studies to date in DS, though, have employed innovative tools such as PheDAS to comprehensively identify co-occurring health conditions across individuals with this diagnosis, or had ample EMRs to include *other* IDDs as a comparison group to account for their co-morbidities in DS (see Appendix A). To address these matters and extend the current literature, N = 2,282 DS cases in our study were sex- and age-matched with N = 4,564 individuals with *other* IDDs and with N = 4,564 case-controls without DS or any other IDDs.

4.1.1. Compared to individuals with other IDDs

Findings from the PheDAS strategy revealed that a range of co-occurring health conditions (145 *phecodes*) were found to be more prevalent in individuals with DS than those with *other* IDDs ($p < 0.05$ after multiple-comparison [Bonferroni] correction) (Figure VI-2). Among the top findings from each health category, co-occurring health conditions that were significantly more prevalent, or *phecodes* that were enriched, among DS than *other* IDD cases included: atrioventricular block (circulatory; b = 2.950, se = 0.294), hidradenitis (dermatologic; b = 2.087, se = 0.427), celiac disease (digestive; b = 2.232, se = 0.237), congenital hypothyroidism (endocrine/metabolism; b = 2.815, se = 0.175), acute renal failure (genitourinary; b = 0.799, se = 0.107), primary thrombocytopenia (hematopoietic; b = 1.336, se = 0.305), Alzheimer's disease (mental; b = 3.917, se = 0.722), laxity of ligament or hypermobility syndrome (musculoskeletal; b = 1.217, se = 0.162), myeloid leukemia [cute] (neoplasms; b = 2.883, se = 0.528), obstructive sleep apnea (neurological; b = 1.776, se = 0.074), respiratory conditions of fetus and newborn (pregnancy complications; b = 0.489, se = 0.075), pulmonary insufficiency or respiratory failure following trauma (respiratory; b = 1.706, se = 0.123), epiphora (sensory; b = 2.761, se = 0.614), and muscle weakness (symptoms; b = 1.881, se = 0.073) (all corrected p-values < 0.05). Using the PheDAS strategy, these co-occurring health conditions found in heightened prevalence among individuals with DS are consistent with previous studies, including recent ones using EMR datasets. Though, unique to our current study is that these health conditions remain as prevalent in DS cases even when compared to *other* IDDs, highlighting some unique or specific clinical features.

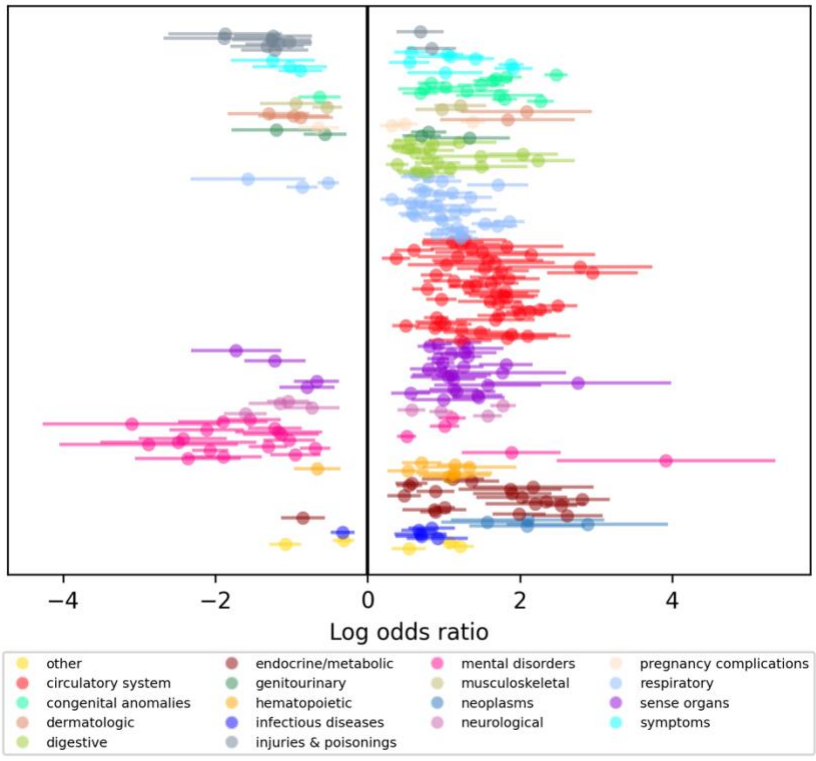


Figure VI-2 Plot of summary PheDAS findings (Log odds ratio) for co-occurring health conditions revealed to be more prevalent in individuals with DS than those with other IDDs (positive values [right panel]; $p < 0.05$ after multiple comparison [Bonferroni] correction).

4.1.2. As-yet unidentified health conditions in DS

Findings on the co-occurring health conditions among individuals with DS were further interrogated to reveal which ones are as-yet unidentified or less discussed within the relevant literature. With this step, a recently innovated *Novel Finding Index* (NFI) approach was employed to rank these health conditions (i.e., *phecodes* found enriched among DS cases) based on how frequently they appear in titles and abstracts containing relevant terms (e.g. "Down syndrome") using PubMed search. Here, main *NFI* results were central to the analyses between individuals with DS versus individuals with *other* IDDs. 33 *phecodes* were revealed with *NFI* > 6 (relative to $m = 3.782$, $sd = 2.508$, *range* = [0.090, 8.201], for the 145 *phecodes* found in confirmatory analyses), which suggests that these particular co-occurring health conditions are less "well studied" or discussed in the DS literature. To refine the clinical scope when interpreting our findings, clinicians with DS specialty were invited to affirm the novelty of these 33 phecodes. In doing so, *phecodes* from each health category included: cardiac pacemaker in situ, first- and second-degree atrioventricular blocks, mitral valve stenosis and aortic valve stenosis, and right bundle branch block (circulatory); duodenitis, atrophic gastritis, and other specific gastritis (digestive); Nonsenile cataract and

eustachian tube disorders (sensory); and elevated C-reactive protein (CRP) (symptoms). These findings could be considered to be as-yet unidentified co-occurring health conditions in DS, as these clinical features appear to be less discussed in this literature. Moreover, these results would motivate future investigations into potentially unmet health needs among individuals with DS.

4.2. Study 2: Longitudinal predictors of surgical intervention among DS cases with CHD

A further step to investigate the relative importance of various co-occurring health conditions is to evaluate among these, which would longitudinally predict known outcomes such as surgical intervention among some DS cases with CHD. Within our sample, N = 1,098 (or 48% of 2,282) DS cases were reported to have CHD diagnoses. Of these cases, N = 220 (or 20%) were found to have a diagnostic code of heart transplant or surgery, whereas N = 878 did not and were utilized in the creation of a 1:1 comparison [case-control] group. The mined set of longitudinal data for N = 204 DS cases with CHD and surgical intervention (versus 204 case-controls) was included in building three predictive models, including Support Vector Machine, Random Forest, and Multi-layer Perceptron, with co-occurring health conditions as the explanatory variables. Then, the best-performing model was evaluated for its precision in predicting the likelihood of having surgical intervention in DS cases with CHD; and the model-based explanatory variables were ranked for relative importance to reveal which health conditions, when enriched, would increase the likelihood of receiving heart-related surgery.

4.2.1. Model-based predictors of surgery in DS cases with CHD

Across the three predictive models, *Random Forest* was the best-performing model based on its precision-recall ratio (Figure VI-3; Table VI-2), meaning that this model provided the best balance between maximizing the true positive rate while simultaneously minimizing the false positive rate. As such, model-based explanatory variables, or health conditions, were characterized using the *Random Forest* classifier to evaluate their relative importance. Results suggested that the presence of congestive heart failure (circulatory; *weight* = 0.048), valvular heart disease or heart chambers (congenital anomalies; *weight* = 0.041), pleurisy or pleural effusion (respiratory; *weight* = 0.022), cardiac shunt or heart septal defect (congenital; *weight* = 0.020), pulmonary collapse or interstitial and compensatory emphysema (respiratory; *weight* = 0.017), cardiomegaly (circulatory; *weight* = 0.017), pulmonary congestion and hypostasis (respiratory; *weight* = 0.014), respiratory failure (respiratory; *weight* = 0.011), cardiac congenital anomalies (congenital; *weight* = 0.009), fever or unknown origin (symptoms; *weight* = 0.009), and congenital anomalies of great vessels (congenital; *weight* = 0.005) each positively predicts the likelihood of surgical

intervention among DS cases with CHD (Figure VI-4; Table VI-3). Findings on these co-occurring circulatory and congenital (cardiac) conditions are expected as they have been shown to exacerbate the severity of CHD in individuals with and without DS [220], [221], [226]. Other health conditions within the respiratory system may relate to (post-)surgical intervention to address medical concerns in or the severity of CHD [227]–[229].

Table VI-2 Performance statistics for predictive models

	Support Vector Machine		**Random Forest**		**Multi-layer Perceptron**	
a) Mean Precision/Recall Curve Samples						
	Recall	*Precision*	*Recall*	*Precision*	*Recall*	*Precision*
	0.01	1.00	0.01	1.00	0.01	1.00
	0.05	1.00	0.05	1.00	0.05	1.00
	0.10	0.89	0.10	0.89	0.10	0.82
b) Training Data						
	Predicted label		*Predicted label*		*Predicted label*	
True label	118	35	144	9	119	34
	43	110	14	139	41	112
c) Testing Data						
	Predicted label		*Predicted label*		*Predicted label*	
True label	36	15	36	15	36	15
	11	40	12	39	12	39

5. Discussion

Genetic circumstances in DS have been linked to various intellectual, developmental, and health differences. Using extant EMR data, our study first comprehensively confirmed this range of co-occurring health conditions in DS, as well as demonstrated those that are less discussed in the relevant literature or potentially as-yet unidentified. A substantial number of co-occurring health conditions among individuals with DS as revealed by our results fell within the circulatory category, followed by endocrine/metabolic, respiratory, and other findings. Given the high prevalence of circulatory differences that include and could exacerbate CHD among individuals with DS, the rich EMR data were further interrogated to reveal longitudinal predictors, or co-occurring health conditions, of surgical intervention. Model-based findings revealed several circulatory, congenital, and respiratory conditions that longitudinally predict the likelihood of surgical needs among individuals with DS and CHD. Put together, while evidence clearly indicates the

range of differences in health conditions associated with DS, the combination of clinical data (EMRs), planned methods, and thorough findings could provide crucial insights into precision-medicine development and determination of unmet medical support for individuals with DS.

Table VI-3 Health conditions related to the likelihood of surgical intervention among DS cases with CHD based on model-based explanatory predictors from best-performing *Random Forest* classifier.

Code	Phecode	Weight	Category
428.1	Congestive heart failure (CHF) NOS	0.048	circulatory system
747.12	Valvular heart disease/ heart chambers	0.041	congenital anomalies
507	Pleurisy / pleural effusion	0.022	respiratory
747.11	Cardiac shunt/ heart septal defect	0.020	congenital anomalies
508	Pulmonary collapse / interstitial and compensatory emphysema	0.017	respiratory
416	Cardiomegaly	0.017	circulatory system
503	Pulmonary congestion and hypostasis	0.014	respiratory
509.1	Respiratory failure	0.011	respiratory
747.1	Cardiac congenital anomalies	0.009	congenital anomalies
783	Fever of unknown origin	0.009	symptoms
747.13	Congenital anomalies of great vessels	0.005	congenital anomalies

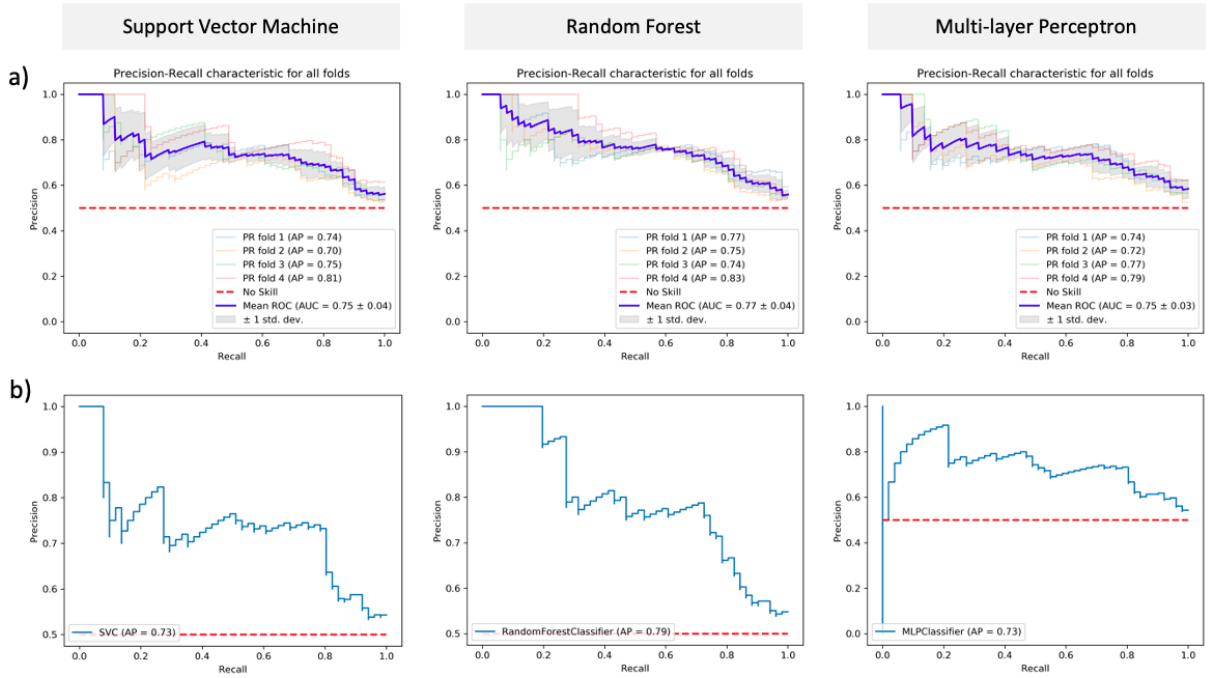


Figure VI-3 Precision-Recall Characteristic plots for predictive models.

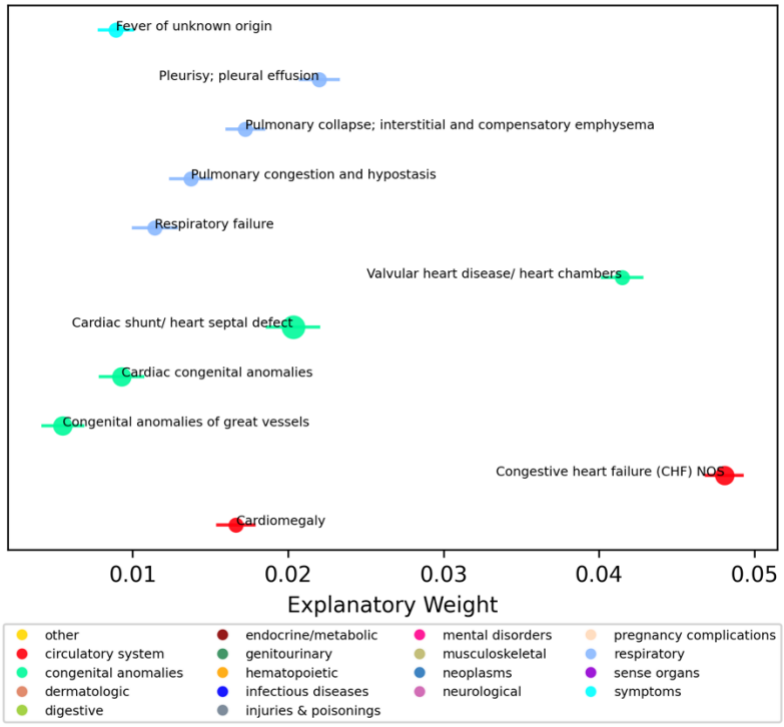


Figure VI-4 Health conditions related to the likelihood of surgical intervention among DS cases with CHD based on model-based explanatory predictors from best-performing Random Forest classifier.

5.1. Known versus as-yet unidentified co-occurring health conditions in DS

Consistent with prior studies in large EMR datasets, our findings using the PheDAS approach revealed a range of co-occurring health conditions among individuals with DS relative to case-controls and those with *other* IDDs (e.g., [211], [217]). Circulatory conditions were found as highly prevalent among DS cases, as differences in heart formation at birth among many individuals with DS have been linked to various cardiac complications, including (pulmonary) hypertension, atrioventricular block, and heart failure, aside from CHD [230]. Our results also confirmed other previously shown clinical features to co-occur in DS such as hypothyroidism [212], sleep apnea [213], [231], and Alzheimer's disease [232]. Many of these health conditions among individuals with DS have been linked to differences and severity in some circulatory complications (e.g., hypertension, CHD [232], [233]). Other co-occurring conditions with vision and hearing, digestion and respiratory, as well as leukemia have previously been found in DS [213], [234]. Notably, our study carefully assessed and replicated these clinical findings associated with DS cases versus case-controls and those with *other* IDDs (e.g., autism, ADHD, dyslexia). Put together, findings on this range of co-occurring conditions highlight the complexity in heath profile among DS cases and, at the same time, emphasize the needs to comprehensively consider multiple clinical features and health outcomes

within the EMRs to better serve individuals with DS. Future studies may consider extending the utilization of EMR datasets and nuanced tools like PheDAS to delineate some imperative clinical features that are potentially unrecognized in smaller samples.

Building on the range of co-occurring health conditions in DS as revealed from the PheDAS approach, these findings were subjected to comparison with what is more versus less well studied in the relevant literature – that is, using "Down syndrome" search term on PubMed [99]. In highlighting which clinical features are likely as-yet unidentified among individuals with DS, this step revealed a number of co-occurring circulatory conditions that are less well studied or discussed in the literature (with $NFI > 6$; and after consulting with clinicians/co-authors). Some of these circulatory conditions included pacemaker in situ, first- and second-degree atrioventricular blocks, mitral valve stenosis and aortic valve stenosis, and right bundle branch block. Atrioventricular block is a known complication linked to surgical intervention on septal defects and CHD [235], which in terms of our study are particularly prevalent to DS [236]. Some studies have further teased out the links between atrioventricular block and DS to pacemaker placement, post-operative complication, and heart valve stenosis [220], [223], [237]. Some other co-occurring conditions among DS cases that were found to be less discussed were: gastritis that could be linked to hypothyroidism [238]; Nonsenile cataract and eustachian tube disorders that could tie to demonstrated visual and hearing differences [239]; and elevated C-reactive protein that could associate with Alzheimer's disease [240], [241]. Further studies may consider confirming our current findings; reports, like ours, on as-yet unidentified co-occurring health features in DS could help researchers and clinicians pinpoint unmet support among individuals with this genetic condition.

5.2. Health conditions in the likelihood of surgery among DS cases with CHD

Across various co-occurring circulatory and other differences in DS, the prevalence of CHD reportedly accounts over 40% of individuals with this genetic condition, more than 20% of whom require surgical intervention to correct the underlying cardiac malformations [222], [223]. Different types of heart defects, such as atrioventricular or tetralogy of Fallot, have been posited to explain the varying degree of severity in CHD, and in turn surgical needs, among some individuals with DS [222]. Congestive heart failure is another circulatory condition that is known to associate with CHD [242], whereby this relation is likely linked to decision for surgical intervention and correction of the underlying cardiac differences in DS [222]. Indeed, our model-based findings confirmed the presence of congestive heart failure in the predicted likelihood of surgical needs among individuals with DS and CHD. Additionally, such likelihood of surgical needs was explained by cardiomegaly, a heart enlargement condition that has been previously demonstrated as attributable to congestive heart failure in CHD (e.g., [243]). Our results found other congenital conditions

related to the heart, including valvular, septal, and vessel issues, a combination of which could relate to the severity of CHD among individuals with DS and lead to needed surgical correction [222], [244]. Finally, respiratory failure and other pulmonary difficulties (e.g., congestion, hypostasis, pleural effusion, emphysema) were observed in relation to the likelihood of surgical intervention among DS cases with CHD in our sample. Previous studies have reported that the prevalence of these different respiratory and pulmonary conditions may surround hospitalization and surgical intervention among individuals with DS and CHD, and that have an impact on their post-operative recovery and life expectancy after correction of cardiac malformations[213], [245]. While promising, further investigation is needed into health conditions that play a role in surgical and medical care among DS cases CHD, as such findings could shed insights onto the effectiveness of interventions and long-term clinical outcomes for individuals with genetic differences.

5.3. Limitations and future directions

While offering promising evidence on health conditions in DS using extant EMR data and methodologically rigorous tools, there are several limitations to our analysis that future studies may consider tackling. First, differences in co-occurring health conditions among individuals with DS have been linked to access to care, health insurance, and socioeconomic differences [246]. Future studies may consider applying tools to derive these metrics from the EMRs [247]–[250], as well as to examine the extent to which environmental and malleable factors relate to clinical features in DS—a genetic condition. Second, an important consideration is the health association between individuals with DS and their immediate caregivers (e.g., parents). Relative to caregivers of individuals *without* DS (or some other IDDs), caregivers of individuals *with* DS have reported varied levels of physical, emotional, and clinical profile of well-being [251], [252]. By applying "dyadic pairing", studies have characterized areas of support (e.g., geographical, social, financial, educational) discussed from the caregivers' perspective [253], or could reveal differences in health conditions between individuals with DS and their caregivers. Third, PheDAS was utilized in the current study given its propensity to evaluate "whole phenome" and to reveal various co-occurring health conditions among individuals with DS. Future studies may consider innovated and streamlined tools such as PheGWAS, a combination of phenotype- and genome-wide association studies [254], and interrogate the phenotypic-genetic underpinnings of co-occurring health conditions in DS. Overall, research efforts using EMRs and rigorous methods, including our current study, could shed crucial light on the complexity in health profile among individuals with DS and other IDDs and further motivate precision-care development.

Chapter VII

Batch Size: Go Big or Go Home?
Counterintuitive Improvement in Medical Autoencoders with Smaller Batch Size

1. Overview

Batch size is a key hyperparameter in training deep learning models. Conventional wisdom suggests larger batches produce improved model performance. Here we present evidence to the contrary, particularly when using autoencoders to derive meaningful latent spaces from data with spatially global similarities and local differences, such as electronic health records (EHR) and medical imaging. We investigate batch size effects in both EHR data from the Baltimore Longitudinal Study of Aging and medical imaging data from the multimodal brain tumor segmentation (BraTS) challenge. We train fully connected and convolutional autoencoders to compress the EHR and imaging input spaces, respectively, into 32-dimensional latent spaces via reconstruction losses for various batch sizes between 1 and 100. Under the same hyperparameter configurations, smaller batches improve loss performance for both datasets. Additionally, latent spaces derived by autoencoders with smaller batches capture more biologically meaningful information. Qualitatively, we visualize 2-dimensional projections of the latent spaces and find that with smaller batches the EHR network better separates the sex of the individuals, and the imaging network better captures the right-left laterality of tumors. Quantitatively, the analogous sex classification and laterality regressions using the latent spaces demonstrate statistically significant improvements in performance at smaller batch sizes. Finally, we find improved individual variation locally in visualizations of representative data reconstructions at lower batch sizes. Taken together, these results suggest that smaller batch sizes should be considered when designing autoencoders to extract meaningful latent spaces among EHR and medical imaging data driven by global similarities and local variation.

2. Introduction

Autoencoders are a class of deep learning models that seek to construct compressed latent representations of input populations by constraining that the latent representations be able to reconstruct the inputs. In practice, they have gained popularity in dimensionality reduction, clustering, denoising, and anomaly detection applications [255]–[258]. Additionally, in the medical domain, ideal autoencoders are not only able to reconstruct inputs but are also able to uncover latent representations that capture meaningful information about the population without supervision (Figure VII-1) [259], [260].

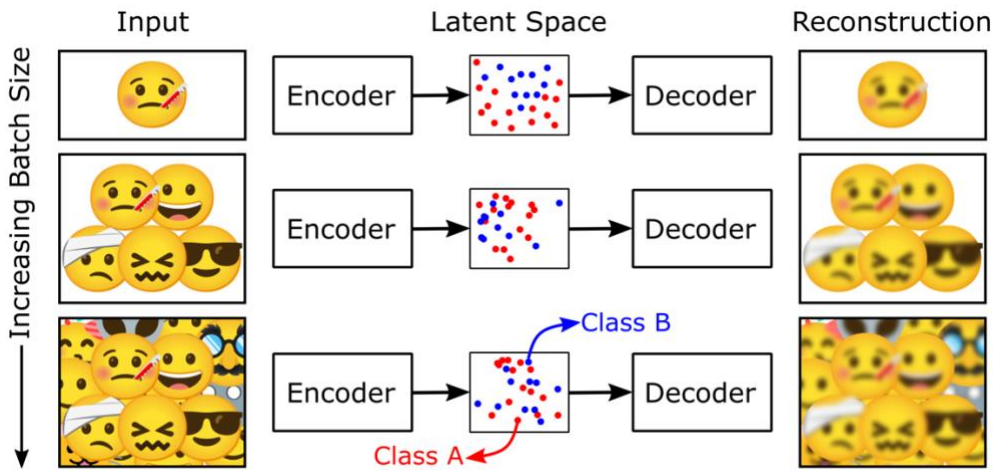


Figure VII-1 Medical autoencoders seek to derive latent spaces that capture clinically or biologically meaningful information about the cohort. Batch size is a key hyperparameter for training these models, but it is unclear how batch size impacts their performance.

The reconstruction loss is the lone constraint on an autoencoder, and thus the performance of the model depends solely on its reconstruction capacity. In practice, model training occurs on the batch (or "mini-batch") level, meaning backpropagation and model weight updates occur after a group of individuals in the training cohort is passed through the network and the reconstruction loss is computed by averaging the reconstructions across the batch. Thus, it is logical that batch size might be an important hyperparameter when designing autoencoder training protocols.

Conventional wisdom suggests larger batch sizes in medical deep learning offer improved performance. However, in reality, the literature surrounding ideal batch size selection remains unclear. For instance, one study suggests that increased batch sizes during training may achieve the same effect as decaying the learning rate, a common practice in deep learning used to improve performance [261]. Others suggest the primary impact of larger batch sizes is a change in training time [262], [263] and yet others still conclude that performance may indeed be impacted by batch size via a so-called "generalization gap" [264], [265]. As such, despite conventional wisdom, there is a gap in the literature regarding the effects of batch size on deep learning training paradigms. Additionally, these studies have primarily been conducted in the natural data domain, but fundamental differences between many natural and medical domain datasets may affect the application of these studies to medical autoencoders, further widening this gap.

Specifically, in medical domains, spatially global similarities between individuals can dominate inputs whereas individual variability can be relegated to the local level. For instance, consider brain MRI, especially those that are co-registered and intensity normalized. While the shape of different gyri and sulci and the presence of lesions may differentiate images locally, the prior probability that any given brain MRI

appears globally as a centered, blurred, gray-scale ellipsoid with ventricles in the middle is high. Thus, one logical solution for autoencoders constrained with a reconstruction loss averaged over individuals in the batch is to simply ignore local individual variability during reconstruction in favor of producing a global average regardless of the input. While this minimum satisfies an averaged reconstruction criterion, this theoretically destroys the network's ability to capture meaningful individual variation on the local level and thus cannot produce a useful latent space.

It follows then that reducing the amount of averaging across a batch during backpropagation might place more emphasis on individual variability on the local level. Thus, in medical domains where global similarities between individuals dominate local differences, we hypothesize, against conventional wisdom, that reducing batch size will not only improve autoencoder reconstruction ability but also the utility of the associated latent spaces by allowing the models to better capture individual variability.

We investigate this hypothesis in both EHR and brain tumor imaging data by training an autoencoder for each dataset at different batch sizes while keeping all other hyperparameters the same. We then inspect the latent spaces of these autoencoders as a function of batch size for their ability to preserve key individual characteristics, such as the sex of the participants and the right-left location of the tumors, respectively, as well as the reconstructions for their ability to capture individual variation beyond a global average.

3. Methods

3.1. Data overview and preparation

To study the effect of batch size on EHR autoencoders, we use International Classifications of Disease version 9 (ICD-9) codes from 3127 participants in the Baltimore Longitudinal Study of Aging [111], [266]. In this study, participants checked in with the data collection team every 1-4 years depending on their age; during these visits, ICD-9 codes were collected via self-report, physical examinations, and medical record history. This yielded a dataset of 321,265 unique ICD-9 events occurring between the ages of 17 and 104 across all 3127 participants. To reduce noise and dimensionality in this data, we mapped all ICD-9 codes to a set of 1866 Phenotype codes (PheCodes) [167], a hierarchical set of meaningful codes that group similar ICD-9 codes together. Each PheCode was then aggregated across each participant's record, such that the PheCode was assigned a 1 if it was present in the participant's record and 0 otherwise. Together, this mapping and aggregation resulted in a binary 1x1866 feature vector for each participant that captured the presence/absence of EHR phenotypes across the participant's lifetime. All PheCode processing was done using the pyPheWAS package [199].

To study the effect of batch size on convolutional autoencoders in brain tumor imaging, we utilized FLAIR MRI from 1251 participants in the BraTS 2021 cohort [267], [268]. These images were made available in 1mm isotropic resolution in Montreal Neurological Institute (MNI) space [269]. For this study, we preprocessed the images by first normalizing the images by the 99th percentile intensity within the brain to rescale them between 0 and 1 and then zero-padded and downsampled them to 3mm isotropic resolution. This produced images of size 81x81x54 voxels in the sagittal, coronal, and axial dimensions respectively.

For both datasets, we followed a random 70/10/20% split on the participant level for both datasets to obtain training, validation, and testing cohorts, respectively. We trained all autoencoder models on these same cohorts.

3.2. Autoencoder training protocols

For the EHR data, we designed a fully connected autoencoder to generate a 32-dimensional latent space (Figure VII-2). For the encoder, we used two blocks, each consisting of a fully connected layer followed by a ReLU activation function, reducing the feature space from 1866 dimensions to 256 to 64. These blocks were followed by a fully connected layer to produce the 32-dimensional latent space. Similarly, the decoder

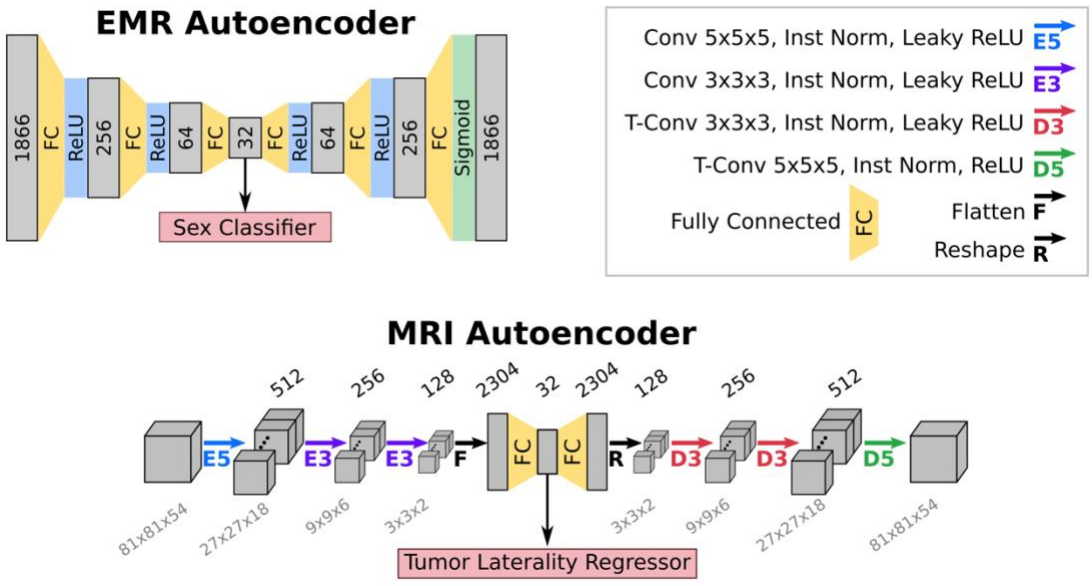


Figure VII-2 Autoencoder architectures. We utilized a fully connected autoencoder for the EHR data, compressing the 1,866-dimensional PheCode feature vectors to 32-dimensions and subsequently reconstructing them. We utilized a convolutional autoencoder for the brain tumor MRI data, compressing images of size 81x81x54 voxels to 32 dimensions before reconstructing them. All convolutional and transpose convolutional layers in the MRI autoencoder utilized a stride of 3 voxels. The derived latent spaces were subsequently evaluated in secondary tasks: sex classification for the EHR data and tumor laterality regression for the MRI data.

consisted of two blocks (a fully connected layer followed by ReLU activation), expanding the 32 dimensions of the latent space to 64 and then 256 dimensions. Finally, another fully connected layer followed by a sigmoid activation function expanded the model output back to the original 1866-dimension feature space. To assess the reconstruction, we used binary cross entropy (BCE) loss optimized via Stochastic Gradient Descent with a learning rate of 0.01 over 5000 epochs on either a NVIDIA Titan Xp or Quadro RTX 5000 GPU. We performed ten different batch size trials, including batch sizes of 1, 2, 3, 4, 5, 7, 10, 25, 50, and 100, maintaining the same training protocol and hyperparameters otherwise. After training, we selected the model with the lowest validation loss as the "best" model for comparison against other batch sizes.

For the brain tumor FLAIR MRI dataset, we designed a convolutional autoencoder to generate a 32-dimensional latent space (Figure VII-2). For the encoder, we utilized three convolutional blocks, each consisting of a convolutional layer, an instance normalization layer, and a LeakyReLU activation with 0.1 slope. The first layer utilized an isotropic kernel of size 5 and stride 3 and the remaining two used an isotropic kernel of size 3 and stride 3. At each of the three layers, the images were encoded to 512, 256, and 128 features, respectively. This resulted in a convolutional encoder output of size 3x3x2 with 128 features. To construct the latent space, the encoder output was flattened and passed through a single fully connected layer. The latent space was then projected and reshaped to match the convolutional encoder output with another fully connected layer and subsequently passed through a convolutional decoder. For the decoder, the encoder was mirrored with three transpose convolutional blocks, each consisting of a transpose convolutional layer, an instance normalization layer, and an activation function. The first and second blocks each used an isotropic kernel and stride of size 3 as well as LeakyReLU activations with 0.1 slope. The third block utilized an isotropic kernel of size 5 and stride of 3 with a ReLU activation. For the reconstruction loss, we utilized a negative log likelihood (NLL) loss (Eq. 1). We define this loss as the negative log likelihood that the input image, $\bar{x}$, of size $D = 81 \times 81 \times 54$ voxels was sampled from a Gaussian, $\mathcal{N}_{\bar{y},\bar{\bar{I}}}$, centered at the reconstruction, $\bar{y}$, with identity covariance, $\bar{\bar{I}}$, averaged across the D voxels:

$$NLL(\bar{y},\bar{x}) = -\frac{1}{D}\sum \log \mathcal{N}_{\bar{y},\bar{\bar{I}}}(\bar{x}) \tag{1}$$

We trained the convolutional autoencoder with batch sizes of 1, 20, 50, and 100, maintaining the same training protocol and hyperparameters otherwise. We used the Adam optimizer with a learning rate of 0.00001 without decay for 1000 epochs on either a NVIDIA Quadro RTX 5000 or A6000 GPU. As with the EHR autoencoder, we selected the model with best validation performance for subsequent testing.

3.3. Latent space evaluation

We examined the utility of each trained autoencoder by leveraging the latent space to perform a secondary task. The EHR latent space was evaluated via classifying participant sex. For each of the 10 batch sizes, we trained a support vector machine (SVM) to predict sex based on the latent space embeddings of the training cohort in 10-fold cross validation. We then projected the EHR testing cohort into the latent space and used the SVMs to generate sex predictions; we evaluated the SVM's classification performance on this test cohort via the area under the receiver operating curve (AUROC) across all folds. Finally, we compare model performances by plotting the testing cohort's AUROC across all batch sizes. As the predictions between folds are not independently sampled, we do not perform statistical testing on AUROCs between batch sizes.

The MRI latent space was evaluated by predicting tumor laterality; for this task, we trained a random forest (RF) regression model on the training cohort to predict the laterality of the tumor centroid, where a laterality of 0 indicated a centroid on the left-most edge of the image and a laterality of 1 indicated a centroid on the right-most edge. We then projected the MRI testing cohort into the latent space and generated tumor laterality predictions from the RF. We plotted the residuals as absolute percent difference across samples as a function of batch size. We evaluated for statistically significant differences between batch sizes with the Wilcoxon sign-rank test at 0.05 significance with Bonferroni correction.

Additionally, we qualitatively examined each latent space by projecting the testing cohort into each latent space and computing a 2-dimensional t-distributed stochastic neighborhood embedding (tSNE) representation [270]. We plotted this 2-dimensional tSNE projection for all batch sizes, coloring each sample by sex for the EHR spaces, or tumor laterality for the MRI spaces.

4. Results

4.1. Training performance

We summarize autoencoder training performance in Table VII-1 as a function of batch size. We observe lower testing and validation losses at lower batch sizes for both datasets. We also observe this improved performance with fewer iterations through the entire dataset (epochs) with decreasing batch size.

Table VII-1 Reconstruction loss performance at best validation epoch across batch sizes for validation and testing cohorts. Best loss performance in bold.

	Batch Size	Best Validation Epoch	Best Validation Loss	Best Testing Loss
EHR*				
	100	4978/5000	0.03706	0.03575
	50	4999/5000	0.03643	0.03520
	25	4990/5000	0.03498	0.03390
	10	3372/5000	0.03446	0.03337
	7	2430/5000	0.03445	0.03333
	5	2270/5000	0.03437	0.03324
	4	1850/5000	0.03433	0.03320
	3	1395/5000	0.03419	0.03309
	2	922/5000	0.03397	0.03287
	1	470/5000	**0.03378**	**0.03274**
Brain MRI**				
	100	989/1000	0.92161	0.92181
	50	963/1000	0.92148	0.92155
	20	986/1000	0.92126	0.92135
	1	241/1000	**0.92121**	**0.92120**

* BCE loss = 0 for a perfect reconstruction
** NLL loss = $-\log\frac{1}{\sqrt{2\pi}} \approx 0.91894$ for a perfect reconstruction

4.2. Qualitative analysis of latent space separability and data reconstruction

In Figure VII-3, we visualize the effect of different batch sizes on the EHR dataset reconstructions and find improved reconstruction of individual variation with smaller batch sizes. We observe a globally uniform signal across subjects with little to no individual variation at a batch size of 100 and steadily improved variation as the batch size decreases to 50 and finally 25 and 10. At batch sizes of 5 and 1, we observed further improved capture of individual variation, though the increases are not as dramatic. Additionally, in Figure VII-3 we present tSNE visualizations of the latent space as a function of batch size for the EHR data. We observe improved separability of the sexes with decreasing batch size, especially as the batch size decreases from 100 and 50 to 25. For batch sizes less than 10, we observe similar separability.

We observe similar trends in the brain MRI autoencoder. We observe the best separability of right-left tumor laterality in the latent space with a batch size of 1 (Figure VII-4C). At a batch size of 20 and 50 we observe some separation but a large amount of mixing. At a batch size of 100 we observe almost total

mixing of right and left. Additionally, in Figure VII-4A and B, we observe improved reconstruction quality at lower batch sizes. Specifically, we find in one representative case that the shape of tumor boundaries and the affected ventricles sharpens from a batch size of 20 to 1 (Figure VII-4A). We find in another representative case that at batch sizes larger than 1, the tumor presence is difficult to detect, whereas a batch size of 1 better identifies the expected hyperintensity (Figure VII-4B).

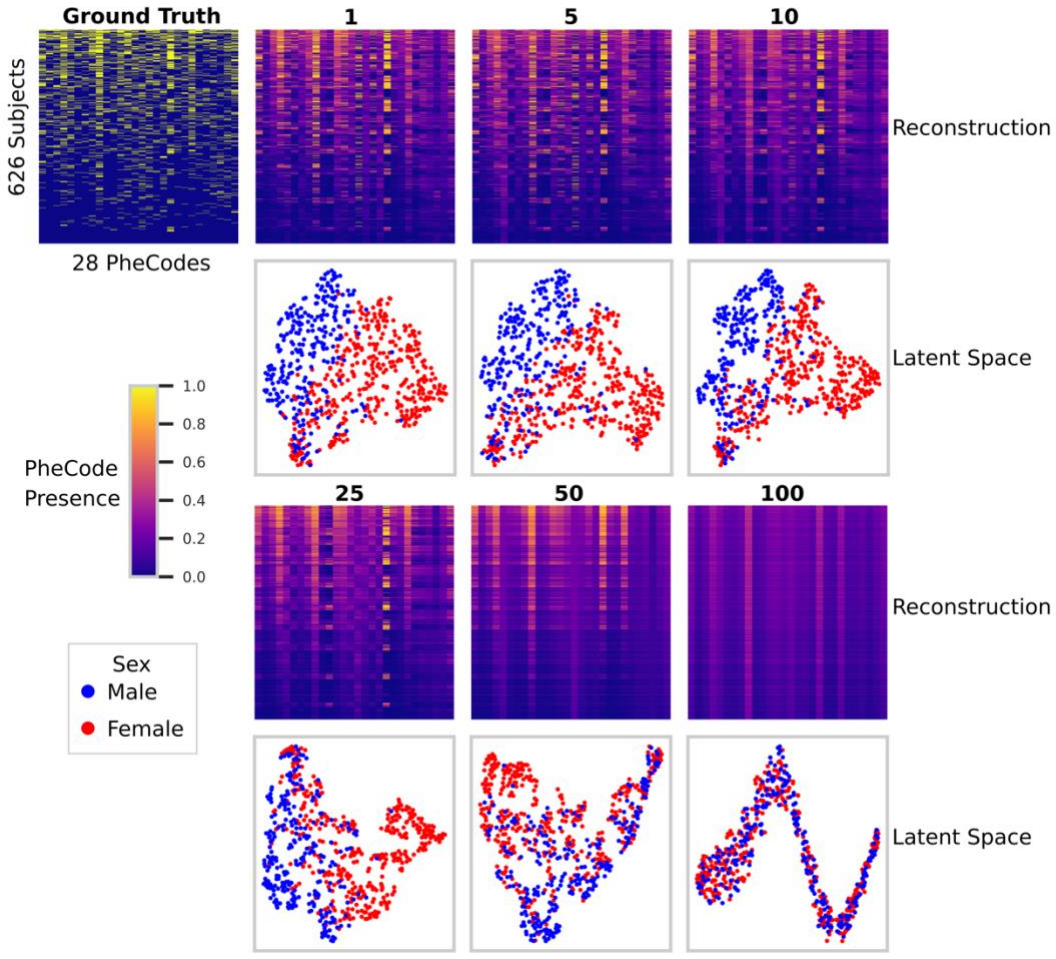


Figure VII-3 EHR autoencoder ground truth, reconstruction, and latent space visualizations for the withheld test set (n = 626) across six batch sizes (1, 5, 10, 25, 50, and 100). The ground truth and reconstructions consisted of a 626 x 28 grid, with individuals on the y-axis and the 28 most representative PheCodes on the x-axis; individuals were sorted so that those with the most PheCode events are at the top of the grid. The reconstructed PheCode value was indicated via color; note that while the ground truth contained only binary values, reconstructions were the output of a sigmoid function and therefore contained intermediate values. Latent space visualizations consisted of a 2-dimensional tSNE projection of the 32-dimensional latent space embeddings of the test cohort; color in this visualization denoted individual sex.

4.3. Latent space performance on secondary tasks

In Figure VII-5A, we plotted the AUROC for sex classification across folds for different batch sizes. We find dramatic improvements in AUROC from batch sizes of 100, 50, 25, to 10. In Figure VII-5B, we plot the absolute percent difference in tumor laterality regression and identify statistically significant improvements under Wilcoxon sign-rank tests with Bonferroni correction from a batch size of 100 to 20 or 1 and a batch size of 50 or 20 to 1.

5. Discussion

In this work, we found qualitatively that (1) reconstructions from larger batch sizes tended to converge toward a global average, whereas smaller batch sizes better preserved individual variation in EHR codes and tumor characteristics and (2) latent spaces from smaller batches better preserved individual sex and tumor laterality. These visual findings were supported quantitatively via smaller reconstruction losses in

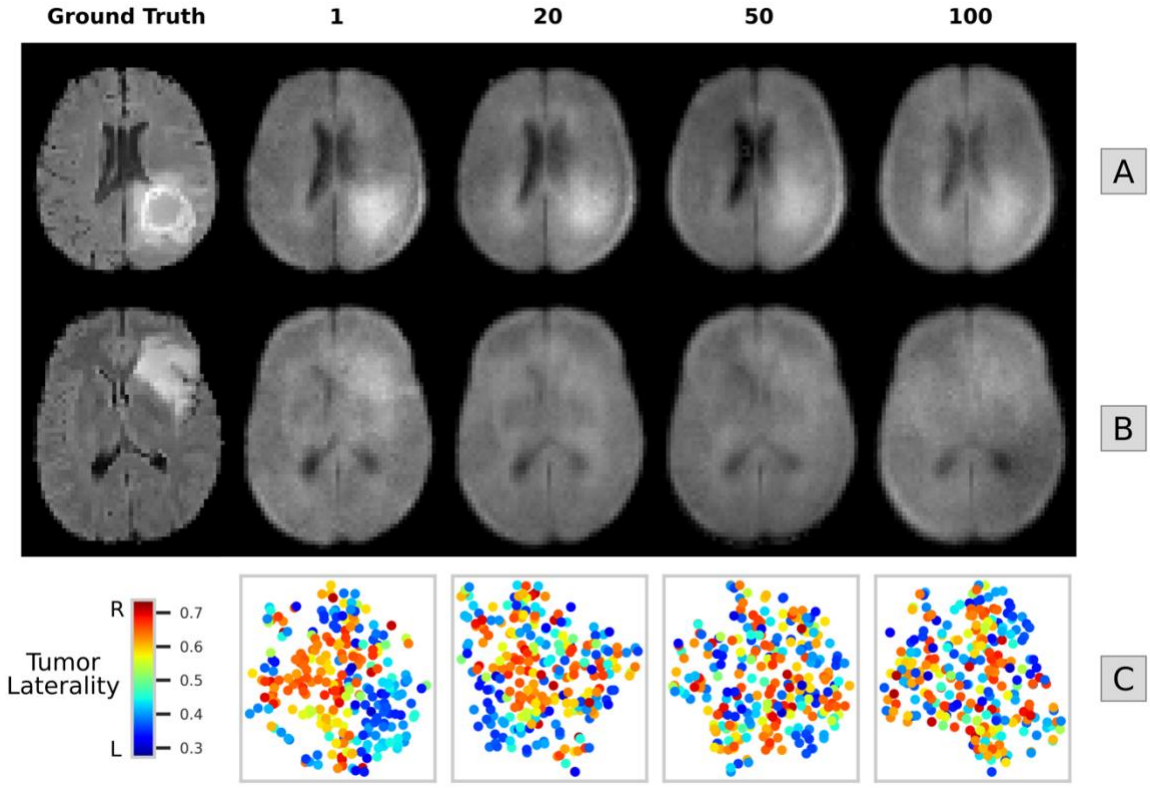


Figure VII-4 MRI autoencoder ground truth, reconstruction, and latent space visualizations across 4 batch sizes (1, 20, 50, and 100). [A, B] Ground truth and reconstructed axial slices were shown for two representative individuals in the withheld test set. [C] Latent space visualizations consisted of a 2-dimensional tSNE projection of the 32-dimensional latent space embeddings of the full withheld test cohort (n = 250); color in this visualization denoted tumor laterality with 0 being the left-most edge of the image and 1 being the right-most edge.

fewer epochs and improved latent space performance on sex classification and tumor laterality regression at smaller batch sizes. These results support our hypothesis that reducing batch size not only improves autoencoder reconstruction ability but also the utility of the associated latent spaces.

We do not robustly explore the theoretical underpinnings for this finding presently, but we offer a potential explanation for this phenomenon. During each batch, model weights are updated following the gradients with respect to the loss, which is averaged across the batch. We posit that though one individual's data may push the gradients to move in one direction, another individual in the batch may push them in a different direction. If both individuals share global features but differ locally, the gradients corresponding to the local features may cancel out when averaged across a batch whereas those corresponding to the global features may compound. Thus, this would effectively cancel out the model's ability to capture individual variation in favor of capturing similar, global features.

Additionally, we posit that this problem may not be present for all deep learning models, such as those with residual connections as they would allow for local context in the input to inform the output. For instance, a U-net can be conceptualized as an autoencoder with skip connections [271]. Since features of the input at multiple scales inform the decoding process, the network would not need to encode local variability, instead relying on the data to do so implicitly during the reconstruction process and would thus not suffer from this problem. That being said, because of the skip connections, a U-net would not be a good replacement for autoencoder applications, as information in the input inherently would not be encoded in the latent, or bottleneck, layer. Thus, the ability to improve the capture of local variability in autoencoders remains an open problem, as addressed presently.

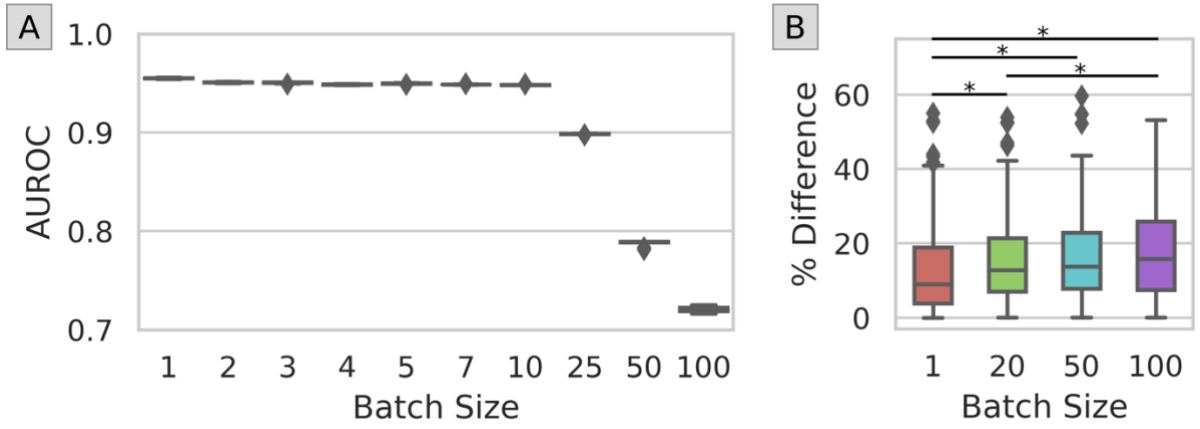


Figure VII-5 Batch size effect on test cohort sub-task performance. (A) AUROC for SVM test cohort sex predictions across 10 cross-validation folds for all EHR autoencoder batch size trials. (B) Percent difference in tumor laterality predictions compared to ground truth for the MRI test cohort across all MRI autoencoder batch size trials. (* p <0.05 with Wilcoxon sign-rank tests after Bonferroni multiple comparisons correction)

This work is not without its limitations. First, we do not consider the effect of other parameters in conjunction with batch size. For instance, prior studies have looked at the effects of learning rate in conjunction with batch size on model performance [272]. Second, we investigate these effects in small datasets. Thus, it is unknown whether the same phenomena would occur with densely sampled medical datasets of tens or hundreds of thousands or even millions of individuals. Third, we do not investigate the effects of batch size on other tasks like in supervised classification or regression where reconstruction is not the primarily constraint, though there is some literature suggesting a similar phenomenon in histopathological image classification [273].

Based on this work, there are many directions to pursue. One is the use of additional latent constraints, such as those used in variational autoencoders [274], to regularize the network, which may offer improved performance. Further investigation into the theoretical underpinnings of this phenomenon would also be beneficial in advancing the field toward the use of autoencoders to develop robust and meaningful latent spaces. Last, considering each batch as a homogeneous group of individuals with a few divergent samples canceling out the gradient directions that capture individual variability, one potentially can view this problem as a case of multi-instance learning, a class of deep learning approaches which may offer additional solutions.

Chapter VIII

Unsupervised Hard Case Mining for Medical Autoencoders

1. Overview

Autoencoders are commonly used for unsupervised dimensionality reduction in medical datasets. One of the most influential hyperparameters in autoencoder training is batch size; choosing an appropriate batch size for a given dataset is imperative for producing a model with high reconstruction accuracy and a rich, interpretable latent space. Previous work has suggested that smaller batch sizes may produce better models, but the longer training time often required for small batches can make this solution impractical for large datasets. In this work, we propose that the primary drawback of large batch sizes is that batch averaging ignores important local variation in favor of dominant global similarities often seen in medical data. This effectively creates a pseudo-class imbalance, with global similarities being the over-represented class and local variations the under-represented class. Therefore, to approximate smaller batch performance while maintaining large batch training time, we propose unsupervised hard case mining, a computationally efficient extension of supervised hard positive/negative mining for unlabeled data. During training, this method considers only the hardest subset of training data within each batch, where "hardest" cases are defined as those with the highest reconstruction loss. We test the proposed method on autoencoders trained for three different datasets: the canonical MNIST digits dataset, an electronic health record dataset, and a whole-brain magnetic resonance imaging dataset. We demonstrate that across these applications, unsupervised hard case mining may reduce reconstruction loss, accelerate network convergence, and improve latent space interpretability.

2. Introduction

Autoencoders have increasingly been used to solve complex problems in medical contexts [275]. This unsupervised deep neural network architecture learns to compress its input data and then reconstruct that input data from the compressed representation [31]. Though seemingly learning a useless task, this architecture is useful primarily due to the latent space, the learned low-dimensional representation of the input data located at the heart of the network. A well-trained autoencoder's latent space can provide an interpretable and compact representation of complex medical datasets [70], [86]. This interpretability makes the autoencoder a promising candidate model for anomaly detection [276], phenotype discovery [197], [277], [278], and disease classification [115].

Typically, autoencoders (and other deep neural networks) are trained via batch-wise (or "mini-batch") learning; this involves iterating over subsets of the training dataset, called *batches*, and updating model weights incrementally at each iteration according to the prediction accuracy averaged across samples in the batch. Choosing an appropriate batch size is imperative for producing a high-performing model but work in this area has resulted in varying recommendations [261]–[265]. In a previous study, we found that smaller batch sizes tend to produce higher quality medical autoencoders than larger batch sizes [279]. There is a tradeoff, however: large batches often train significantly faster than small batch sizes [262], [263]. Herein, we explore a central question arising from this work: how can we achieve small batch model performance while maintaining the fast training times of large batch models?

We propose that the degraded model performance of large batch medical autoencoders is driven by imbalances inherent in medical data. Typically, medical datasets contain a high degree of global feature similarity, with important variations between individuals occurring at the local feature level. Consider, for example, a typical dataset of whole-brain magnetic resonance imaging (MRI) volumes that have been co-registered and intensity-normalized. Globally, all samples in such a dataset would appear to be centered, grayscale, squiggly ellipsoids with dark cavities in the middle, while important local features, such as individual cortical shapes or lesion characteristics, would vary widely across volumes. During large batch training, where the reconstruction loss is averaged across many individuals, this redundancy in global information between volumes effectively creates a pseudo-class imbalance, with global similarities being the over-represented class and local variations the under-represented class. Therefore, a logical solution for the large-batch autoencoder would be to ignore these important individual variations in favor of the fuzzy global average brain.

Class imbalance is a popular ongoing area of research in *supervised* machine learning [280]. *Hard negative mining* refers to a collection of techniques which compensate for class imbalance by over-training on difficult examples. Here, hard negatives are typically defined as true negative samples which the model mis-classifies as positives [281]. In computer vision, this class of techniques has been effective in multi-class classification [281], re-ID [282], and object detection [283]. Seeing this success, medical researchers have also applied hard negative mining to boost learning in breast cancer [284], [285], lesion [286], [287], tooth decay [287], and symptom detection [288]. However, this method requires labeled data, and therefore cannot be directly implemented for unsupervised frameworks like autoencoders.

In this paper, we present a *batch-wise unsupervised hard case mining* technique for training medical autoencoders. Our proposed method is a computationally efficient extension of supervised hard negative mining for unlabeled medical data. In the following sections, we describe the proposed hard case mining

method and demonstrate its utility for training autoencoders on the canonical MNIST digits, electronic health records (EHR), and whole-brain MRI.

3. Methods

3.1. Unsupervised Hard Case Mining

In this work, we aim to train unsupervised medical autoencoder models with interpretable latent spaces. Previous work has found that this may be achieved with small batch sizes [279], but as training time typically increases exponentially with decreasing batch size, the small batch size approach is impractical for large datasets. To maintain these small batch size results while leveraging the faster training times of a large batch size, we propose an unsupervised batch-wise hard case mining method for training autoencoders. Within a given training batch, we consider **hard cases** to be the top *K* samples with the largest reconstruction loss values, and we calculate a hard case loss by averaging the individual losses across the hard cases. We then update model weights based only on the hard case loss. The number of hard cases being considered, *K*, is calculated according to

$$K = \eta * \beta \tag{1}$$

where β is the batch size and η is the **hard case proportion**. The hard case proportion is a decimal value (0.0, 1.0] which controls the number of samples within each batch that are identified as hard cases. Setting $\eta = 1.0$ is equivalent to traditional batch-wise learning where all samples within the batch are considered. Setting $0.0 < \eta < 1.0$ implements unsupervised hard case mining, where only the most lossy subset of each batch is considered.

Focusing on only the hard cases should allow the model to more quickly refine reconstructions of difficult features while ignoring well-known features. This approach may appear to throw out a large proportion of the training data, but since this subsampling is performed *independently* within every batch, the randomization of batch construction over the course of many training epochs incorporates these seemingly ignored samples into the model. Additionally, medical datasets tend to be dominated by high levels of global similarity with important inter-subject variation occurring at the local level. Therefore, even as the model focuses on the hard cases, it learns global feature representations which may apply to all samples in the dataset. In these ways, learning from the most difficult samples allows the network to refine harder representations without sacrificing reconstruction accuracy for the easier samples.

3.2. Autoencoder Experiments

To explore the efficacy of the proposed unsupervised hard case mining method, we demonstrate its application to autoencoders trained on three experimental datasets: the canonical MNIST digits dataset, a tabular EHR dataset, and a whole-brain MRI dataset. All autoencoders were constructed and trained using PyTorch [289]. MNIST and EHR autoencoders were trained on either an NVIDIA Titan Xp or Quadro RTX 5000 GPU, while the MRI autoencoders were trained on an NVIDIA A6000 GPU.

3.2.1. MNIST

The MNIST dataset is a collection of 2-dimensional 28x28 grayscale images depicting handwritten numeric digits 0-9 [290]. To simulate the common pattern of global similarity with local variation seen in medical data, we only used MNIST digits 4 and 9 for our experiments, yielding 11,791 and 1,991 samples from the standard train and test splits, respectively. This subset of the standard MNIST training dataset was further randomly split 80:20 into training and validation sets for autoencoder training. All MNIST images were linearly scaled to the intensity range [0.0, 1.0] and flattened into a 1x784 vector prior to training.

We used the fully connected autoencoder architecture shown in Figure VIII-1 for the MNIST experiment. Over a series of three fully connected layers with ReLU activation, the encoder compressed the 784-dimension input feature space to a 32-dimension latent space. Three mirrored decoder layers then

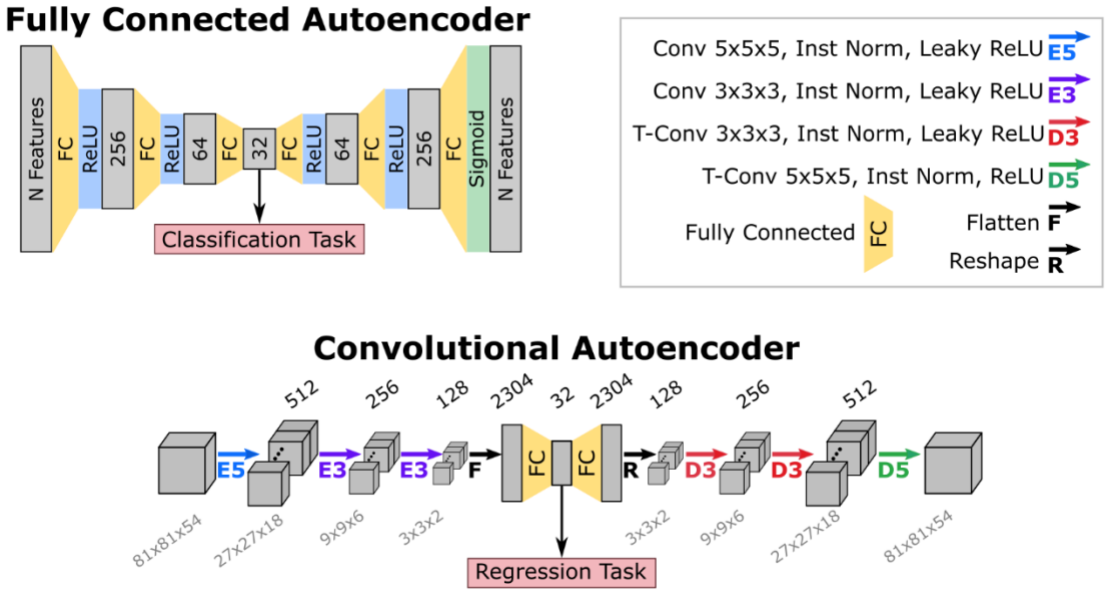


Figure VIII-1 Network architectures for fully connected and convolutional autoencoders. The 32-dimension latent spaces from the fully connected networks were used for classification tasks, while the 32-dimension latent spaces from the convolutional networks were used for regression tasks. Figure adapted with permission from (Kerley et al. SPIE Medical Imaging: Image Processing 2023)

expanded the latent space back to 784-dimensions. Sigmoid activation was used for the output layer only, followed by L1 loss to assess reconstruction accuracy. This network was trained over 5,000 epochs on the training data split and optimized via Stochastic Gradient Descent with learning rate 0.01 and momentum 0.9. Model performance was evaluated at every epoch via reconstruction loss averaged across the validation split. We tested 18 different hard case mining configurations with this MNIST autoencoder design, including all possible combinations of $\beta = \{25, 50, 100\}$ and $\eta = \{1.0, 0.5, 0.4, 0.3, 0.2, 0.1\}$.

3.2.2. EHR

The EHR dataset used for this study covered 3,127 individuals between the ages of 17 and 104 in the Baltimore Longitudinal Study of Aging [291]. Participants were visited at least once every 1-4 years, during which researchers gathered health information and evaluated the participant for various cognitive impairments (mild cognitive impairment, dementia, etc.). Available data included biological sex, cognitive impairment diagnoses, and International Classifications of Disease version 9 (ICD9) codes. Participants were randomly split 70:10:20% into training, validation, and withheld test datasets, respectively.

The pyPheWAS package [199] was used to transform each individual's longitudinal ICD9 record into a compact vector of clinical phenotypes. Briefly, this involved first mapping each ICD9 code to one of 1,866 clinical phenotype codes (PheCodes). PheCodes were then aggregated across each participant's record, yielding a 1x1,866 binary vector representing the presence (1) or absence (0) of each PheCode in the participant's health history. This PheCode vector was used as the input data for the EHR experiment.

We used an almost identical fully connected autoencoder architecture (Figure VIII-1) and training procedure for the EHR experiment as was used for the MNIST experiment. The only modifications made for the EHR models were 1) the autoencoder's input and output layers each contained 1,866 features; 2) binary cross entropy (BCE) loss was used to assess reconstruction accuracy; and 3) the EHR models were trained for 40,000 epochs each. We again tested 18 different hard case mining configurations for the EHR experiment, including all possible combinations of $\beta = \{25, 50, 100\}$ and $\eta = \{1.0, 0.5, 0.4, 0.3, 0.2, 0.1\}$.

3.2.3. MRI

For our MRI experiment, we used 3-dimensional FLAIR MRI volumes from the 1,251 participants in the 2021 BraTS cohort [268], a study focused on high-grade glioma imaging. Each participant in this dataset had a high-grade brain tumor at the time of MRI acquisition. As with the EHR dataset, participants were randomly split 70:10:20% into training, validation, and withheld test datasets, respectively.

The BraTS MRI were available registered to the standard Montreal Neurological Institute reference space in 1mm isotropic resolution [292]. For the current study, all image intensities were rescaled to the range [0,1] by normalizing in reference to the 99th percentile intensity within the brain; volumes were then zero-padded and down-sampled to 3mm isotropic resolution. The preprocessed images were of size 81x81x54 voxels in the sagittal, coronal, and axial dimensions respectively.

The MRI experiment used the convolutional autoencoder architecture seen in Figure VIII-1. An encoder consisting of three convolutional layers followed by a single fully connected layer reduced the input 81x81x54 volume to a 32-dimension latent space; a mirrored decoder expanded the 32-dimension latent space back to the 81x81x54 input space. Instance normalization and the Leaky ReLU activation function with 0.1 slope were used throughout the network's hidden layers. The output convolutional layer was followed by instance normalization and ReLU activation. Reconstruction was evaluated via negative log likelihood (NLL) loss (Eq. 2). NLL loss was defined as the negative log likelihood that the input image $\bar{x}$ of size $D = 81 \times 81 \times 54$ voxels was sampled from the Gaussian distribution $\mathcal{N}_{\bar{y},\bar{\bar{I}}}$ centered at the reconstruction $\bar{y}$ with identity covariance $\bar{\bar{I}}$ averaged across the D voxels:

$$NLL(\bar{y},\bar{x}) = -\frac{1}{D}\sum \log \mathcal{N}_{\bar{y},\bar{\bar{I}}}(\bar{x}) \tag{2}$$

The MRI network was trained on the training data split for 5,000 epochs using the Adams optimizer with learning rate 0.00001 and no decay. As with the MNIST and EHR experiments, model performance was evaluated at every epoch via reconstruction loss averaged across the validation data. We tested 2 different hard case mining configurations with this MRI autoencoder design: $\beta = 20$ and $\eta = \{1.0, 0.1\}$.

3.2.4. Model Evaluation

After training was complete, the "best epoch" was found for each autoencoder by determining which epoch had the lowest validation loss; the model weights for this best epoch were used for all subsequent model evaluations. The reconstruction accuracy of all autoencoders was evaluated by examining the reconstruction loss for the withheld test data splits.

The interpretability of each model's latent space was evaluated *quantitatively* by using the 32-dimension latent space projection for a secondary classification or regression task. The MNIST models were evaluated by training a digit classifier (classes 4/9), and the EHR models were evaluated by training both a biological sex classifier (classes male/female) and a cognitive impairment classifier (classes yes/no). Both the digit and sex classifiers were intended as relatively easy benchmarks due to class balance and the presence of class-specific signals in the original data space. Meanwhile, the cognitive impairment classifier

was intended to be a more difficult task due to class imbalance and the weak label resulting from collapsing varying cognitive impairment diagnoses into a single "yes" class. All three classifiers were multi-layer perceptron models (MLPs) trained with 10-fold cross-validation on the latent space projection of the training data. Each MLP had with a 32-dimension input layer, a 100-dimension hidden layer with ReLU activation, and a 1-dimensional output layer with sigmoid activation; they were trained for 200 epochs and optimized via stochastic gradient descent with learning rate 0.001 and momentum 0.9. Classifier performance was assessed via area under the receiver operating curve (AUROC) for predictions generated from the latent space projection of the withheld data across all 10 cross-validation folds.

The MRI autoencoder latent spaces were evaluated by regressing a tumor laterality score. The tumor laterality score was a decimal in the range (0,1) representing how far to the left (0) or right (1) the brain tumor centroid was in the axial plane of the MRI volume. A random forest (RF) regressor was trained for this task on the 32-dimension latent space projection of the training data split. Regressor performance was evaluated via the absolute percent difference between predicted and actual tumor laterality score for predictions generated from the latent space projection of the withheld test data. Additionally, we looked for

statistically significant differences in regressor performance across η values by computing a Wilcoxon sign-rank test between the percent difference distributions.

Finally, all models were evaluated *qualitatively* by visualizing the latent space projection of the withheld data split in a 2-dimensional t-distributed stochastic neighborhood embedding (t-SNE)[270]. The t-SNE representation of each model's latent space was colored according to each dataset's respective class or regression label to show variation in qualitative interpretability across hard case mining configurations.

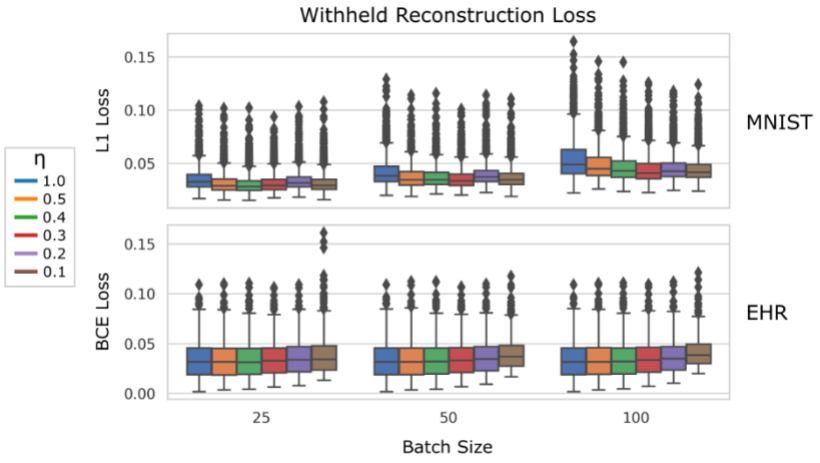


Figure VIII-2 Per-image and per-participant reconstruction loss box plots for the withheld data split evaluated at the best epoch for all MNIST and EHR hard case mining configurations. Points marked above boxes denote outliers.

4. Results

For both the MNIST and EHR datasets, we tested 18 unsupervised hard case mining configurations across 6 hard case proportions and 3 batch sizes; for the MRI dataset, we tested 2 hard case proportion values across a single batch size. For the remainder of this article, we referred to models using the shorthand notation $\eta_X * \beta_Y$, where the subscripts X and Y denoted the values for hard case proportion and batch size, respectively.

4.1. MNIST

Figure VIII-2 summarizes the per-image reconstruction loss for the withheld MNIST dataset across each of the 18 hard case mining configurations. Smaller β values generally produced better reconstruction performance. Within each β value, however, η values less than 1.0 tended to produce better reconstructions with fewer high-loss outliers. Interestingly, there appeared to be an inflection point in this performance improvement at $\eta_{0.3}$ for each value of β; reconstruction loss was higher for η values both greater than and less than 0.3.

To examine these effects more closely, we identified the 5 MNIST images with the *worst* reconstruction losses from the $\eta_{1.0} * \beta_{100}$ model and compared the reconstructions of those 5 images across η values for all β_{100} MNIST models. The same comparison was made for the 5 MNIST images with the *best* reconstruction losses from the $\eta_{1.0} * \beta_{100}$ model. Reconstructions of the "hardest" images (Figure VIII-3)

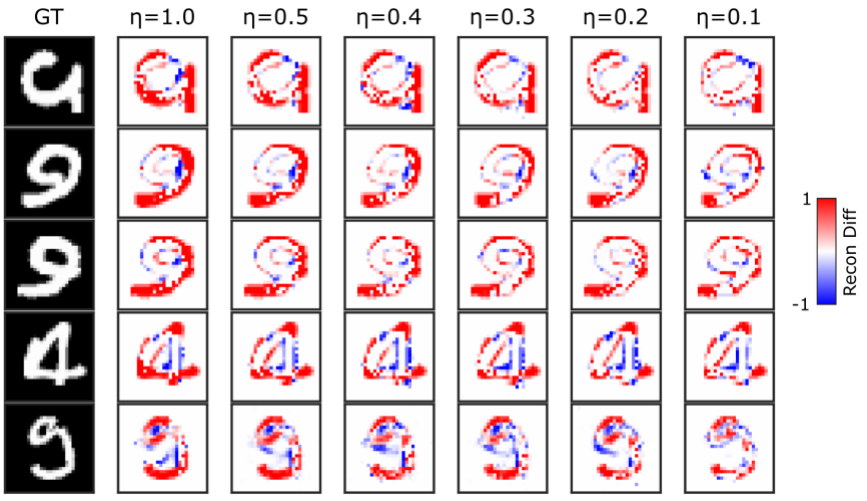


Figure VIII-3 Top 5 worst MNIST reconstructions from the $\eta_{1.0} * \beta_{100}$ model compared to reconstructions from all other β_{100} MNIST models and the image ground truth (GT). Comparisons shown as the absolute difference between GT and the reconstruction.

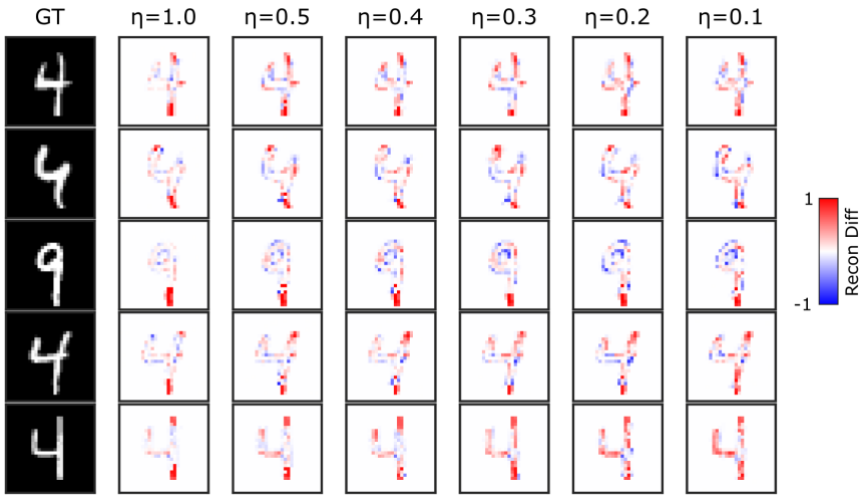


Figure VIII-4 Top 5 best MNIST reconstructions from the $\eta_{1.0} * \beta_{100}$ model compared to reconstructions from all other β_{100} MNIST models and the image ground truth (GT). Comparisons shown as the absolute difference between GT and the reconstruction.

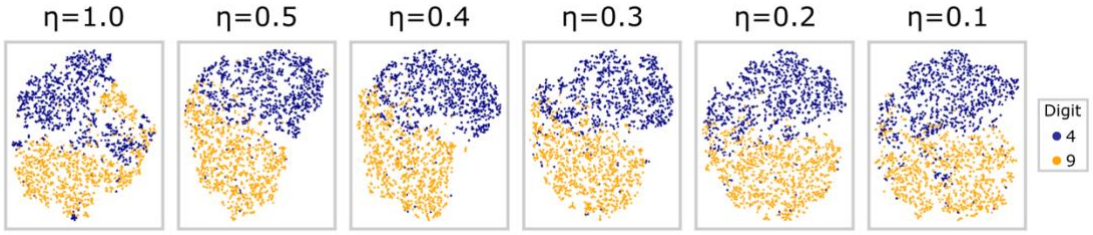


Figure VIII-5 Latent space projections of the withheld dataset for all MNIST β_{100} autoencoders represented in 2-dimensional t-SNE space. Each point represented an individual image. Points are colored according to their digit label.

showed substantial improvement with decreasing η. Conversely, reconstructions of the "easiest" images (Figure VIII-4) showed minor deterioration with decreasing η.

Examining the latent space projections of the β_{100} MNIST models revealed similar, though less dramatic, improvement with unsupervised hard case mining (Figure VIII-5). The t-SNE latent space representations mirrored the effects seen in the reconstruction loss plots from Figure VIII-2: the $\eta_{1.0} * \beta_{100}$ latent space showed some mixing between digits; all $\eta < 1.0$ models showed more separability than the $\eta_{1.0} * \beta_{100}$ model; and there was an inflection point in latent space improvement at the $\eta_{0.3} * \beta_{100}$ model. As expected, classifying MNIST digits was an easy task; all classifiers achieved near perfect AUCROC on the withheld dataset (Table VIII.1).

4.2. EHR

The per-participant reconstruction loss for the withheld EHR dataset across all 18 unsupervised hard case mining configurations was shown in Figure VIII-2. Unlike the MNIST experiment, reconstruction

performance was similar across all β values and this performance worsened with decreasing η. However, there were substantial differences in training time required for each of these models. Figure VIII-6 shows the validation loss curves for the β_{100} models over the full 40,000 epoch training period. Lower η models converged significantly earlier in training than higher η models, with the most extreme difference between the $\eta_{0.1} * \beta_{100}$ (epoch 9,386) and $\eta_{1.0} * \beta_{100}$ (epoch 34,093) models. Similar convergence patterns were also seen across η values in the β_{50} and β_{25} models. Examining the latent space projections of the β_{100} models (Figure VIII-7) showed that the EHR autoencoders had similar latent space separation for both the sex and cognitive impairment classes across η values, despite the decrease in reconstruction losses for lower η models seen in Figure VIII-2.

Finally, the EHR experiment's latent space classifier performances are shown in Figure VIII-8. As expected, sex classification was a relatively easy task; all 18 unsupervised hard case mining configurations achieve greater that 0.92 AUROC averaged across the 10 cross-validation folds. Cognitive impairment, however, was a more difficult task with AUROC values ranging from 0.64 to 0.81. Interestingly, classifier performance for both tasks was similar for corresponding η values regardless of β. This pattern was only broken for the $\eta_{0.1}$ models in the cognitive impairment task where classifier performance across batch sizes diverged, with the β_{25} and β_{100} models having the best and worst performances, respectively. Across both tasks, classifier performance for the $\eta_{1.0}$, $\eta_{0.5}$, and $\eta_{0.4}$ models was largely equivalent; classifier performance then decreased for lower η values.

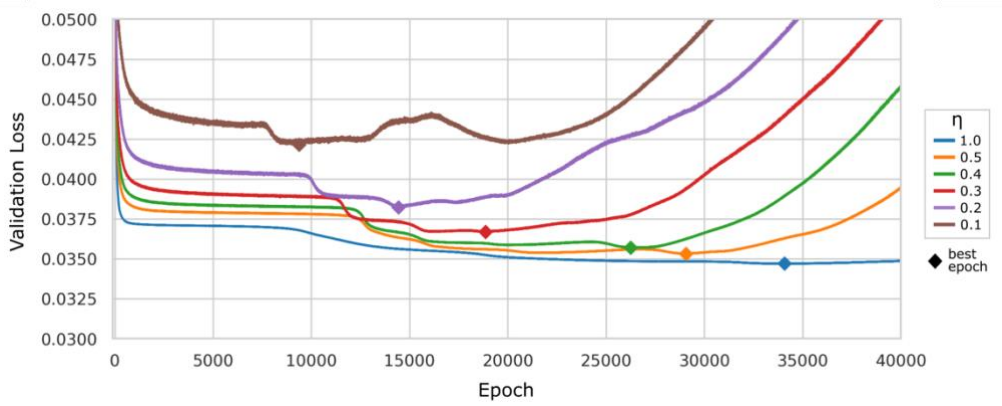


Figure VIII-6 Validation loss curves during training for all β_{100} EHR autoencoders. The best epoch (lowest validation loss) is marked for each model.

Table VIII.1 Reconstruction Loss and Latent Space Classifier Performance for MNIST Withheld Dataset

β	η	Best Epoch	L1 Loss	Classifier AUROC
25	1.0	4991	0.0353 (0.0111)	0.9992 (0.0000)
	0.5	4999	0.0314 (0.0099)	0.9995 (0.0000)
	0.4	4981	0.0305 (0.0092)	0.9991 (0.0000)
	0.3	4981	0.0314 (0.0093)	0.9992 (0.0000)
	0.2	4992	0.0340 (0.0099)	0.9988 (0.0000)
	0.1	4993	0.0315 (0.0094)	0.9988 (0.0000)
50	1.0	4999	0.0415 (0.0132)	0.9987 (0.0001)
	0.5	4994	0.0374 (0.0115)	0.9993 (0.0000)
	0.4	4996	0.0372 (0.0107)	0.9992 (0.0000)
	0.3	4999	0.0361 (0.0103)	0.9989 (0.0000)
	0.2	4994	0.0395 (0.0102)	0.9991 (0.0000)
	0.1	4993	0.0367 (0.0097)	0.9990 (0.0001)
100	1.0	4997	0.0535 (0.0185)	0.9952 (0.0001)
	0.5	4996	0.0489 (0.0149)	0.9979 (0.0001)
	0.4	4989	0.0464 (0.0140)	0.9979 (0.0001)
	0.3	4999	0.0443 (0.0131)	0.9975 (0.0001)
	0.2	4999	0.0455 (0.0119)	0.9973 (0.0001)
	0.1	4988	0.0442 (0.0110)	0.9972 (0.0001)

4.3. MRI

For the MRI experiment, we trained one unsupervised hard case mining autoencoder at $\eta_{0.1}$ and a baseline at $\eta_{1.0}$, both with a batch size of 20. According to validation loss, the best $\eta_{0.1} * \beta_{20}$ model was at epoch 2955, while the best $\eta_{1.0} * \beta_{20}$ model was at epoch 1283. The per-participant reconstruction losses for the withheld MRI dataset across both models are shown in Figure VIII-9A. Similar to the EHR experiment, reconstruction loss was worse for the $\eta_{0.1}$ model compared to the $\eta_{1.0}$ model. Despite this decrease in reconstruction accuracy, however, the latent space for the $\eta_{0.1}$ model showed a more interpretable structure than the $\eta_{1.0}$ model in terms of tumor laterality (Figure VIII-9B). This improved separation in the latent space was reflected in the latent space-trained tumor laterality regression models (Figure VIII-9C). A modest, yet statistically significant ($p<0.05$), improvement was observed in the absolute percent difference between predicted and actual tumor laterality score for the $\eta_{0.1}$ regression model compared to the $\eta_{1.0}$ model.

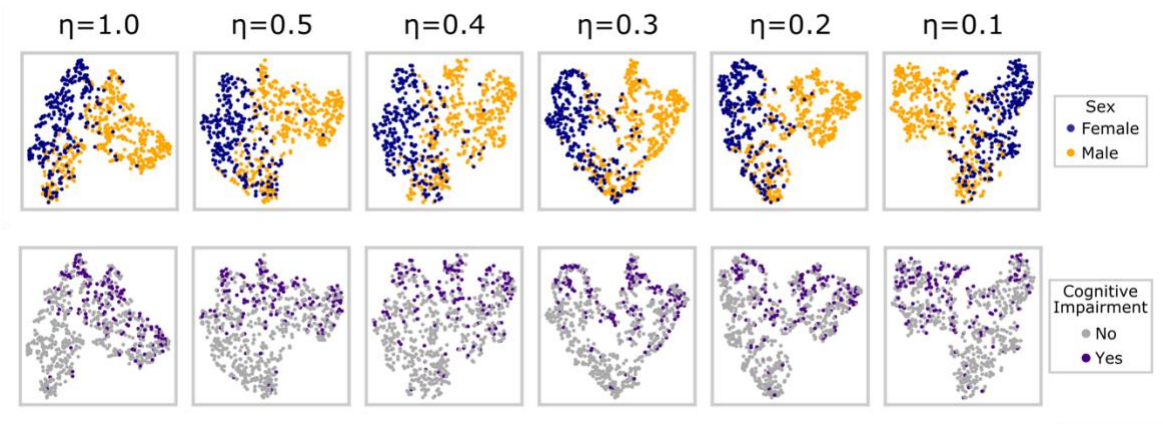


Figure VIII-7 Latent space projections of the withheld dataset for all EHR β_{100} autoencoders represented in 2-dimensional t-SNE space. Each point represented an individual participant. Points are colored according to biological sex (top row) and cognitive impairment (bottom row).

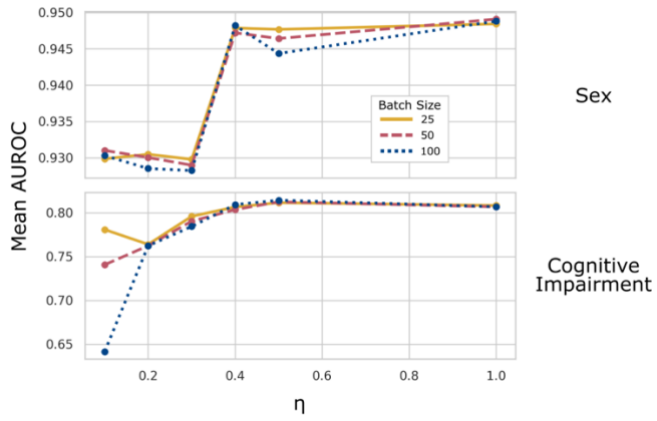


Figure VIII-8 Sex and cognitive impairment classifier performance as a function of η. Performance is expressed as the mean classifier AUROC evaluated on the withheld EHR dataset across 10 cross-validation folds.

5. Discussion

In this article, we presented unsupervised hard case mining, a batch-wise approach for optimizing medical autoencoder models. We tested this method in experiments on natural image data (MNIST), EHR, and whole-brain MRI. Taken together, the results from these experiments demonstrated the varied benefits of the proposed method. In the MNIST experiment, we saw that unsupervised hard case mining improved the mean reconstruction accuracy of the autoencoder models; in the EHR experiment, we saw that unsupervised hard case mining significantly accelerated model convergence; and in the MRI experiment, we saw that unsupervised hard case mining improved the interpretability of the autoencoder's latent space both qualitatively and quantitatively. These benefits were not uniformly applied, however. In contrast to MNIST, both the EHR and MRI experiments suffered a decrease in reconstruction accuracy because of hard case mining; neither the MNIST nor MRI experiments showed accelerated model convergence like

the EHR experiment; and counter to the improvement seen in MRI, the MNIST and EHR latent spaces remained largely consistent with hard case mining.

These varied effects may partially be due to variation in the underlying characteristics and distributions of these disparate datasets. After all, each experiment focused on a fundamentally different data type, and each dataset contained varying levels of feature and class imbalances. Another contributor to the varied effects, however, may be related to how unsupervised hard case mining affects model convergence. Given identical batch size, we saw in the EHR experiment that unsupervised hard case mining accelerated the convergence of models with smaller hard case proportion values; but all models, regardless of hard case proportion, eventually converged to a similar solution. Consider the EHR experiment's β_{100} validation loss curves from Figure VIII-6; if training had ended after 20,000 epochs, the $\eta_{1.0}$, $\eta_{0.5}$, and $\eta_{0.4}$ models would not have converged. At this point in training, both reconstruction accuracy and latent space interpretability for these three higher η models would likely have been poor compared to the fully converged lower η models. This effect may have contributed to the differences in reconstruction loss and latent space interpretability seen in the MNIST and MRI experiments as well. Given sufficiently long training times, the MNIST and MRI models may also have converged to similar solutions regardless of η value. Exhaustive training such as this is often not feasible, though. The EHR models were intentionally trained for an atypically long time; based on the obvious validation loss divergence at lower η values seen in Figure VIII-6, we were convinced that if trained for long enough, the high η models would eventually converge.

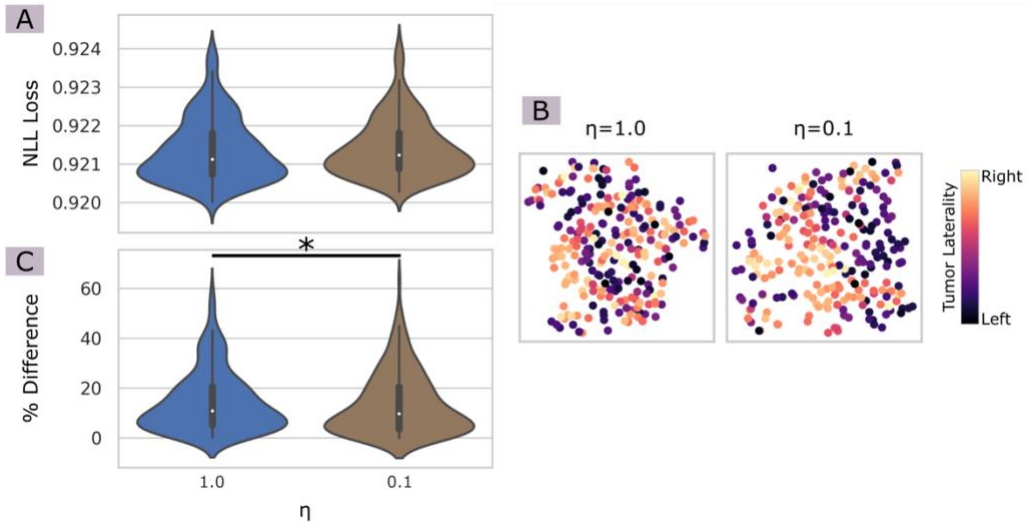


Figure VIII-9 MRI experiment results in unsupervised hard case mining. (A) Per-participant reconstruction loss violin plots for the withheld data split evaluated at the best epoch for both MRI models. (B) Latent space projections of the withheld dataset represented in 2-dimensional t-SNE space. Each point represents an individual participant; color denotes tumor laterality score. (C) Absolute percent differences between predicted and actual tumor laterality score for regression models trained on the latent spaces of both MRI models.

We allowed them extensive training time, but this was only possible because of the EHR model's relatively fast per-epoch wall time. Given the consistency in latent space interpretability and classification accuracy seen at the lower η values, though, this extra training time amounted only to wasted time and resources.

Besides the heterogeneity seen in experimental results, the proposed unsupervised hard case mining method has several areas which require further investigation. First, it is currently unclear how to select values for β and η given an arbitrary dataset, though the results of this study tend towards smaller η values, regardless of β. Second, the simple definition of "hard cases" as those samples within a batch having the largest reconstruction losses has not been tested against other alternatives. It is possible that a more optimal unsupervised hard case mining method could include other information in this definition, such as gradient information or similarity between samples. Finally, a key assumption throughout this article has been that the data being encoded is characterized by a high degree of global similarity paired with important local-level variation; the generalization of this method to datasets with more global heterogeneity, such as the full MNIST dataset or an MRI dataset including multiple body parts, should be explored.

Regardless, the potential application areas for unsupervised hard case mining as proposed are quite broad. This method has both a low computational cost and simple implementation (in our PyTorch implementation, it required only two additional lines of code); additionally, it is both dataset and architecture agnostic. As such, it may easily be incorporated into existing training protocols. This unsupervised framework could also be implemented in supervised network training. As discussed previously, supervised networks have benefited from label-specific hard positive and hard negative mining [281], [284], [285], [288], so it is reasonable to assume that an unsupervised approach could boost learning without the extra computational cost of comparable supervised methods.

6. Conclusion

In this paper, we presented a batch-wise unsupervised hard case mining approach for training medical autoencoders. In experiments on MNIST digits, EHR, and brain MRI, we demonstrated that the proposed method increased reconstruction accuracy, accelerated model convergence, and improved the latent space both qualitatively and quantitatively. This simple and computationally inexpensive approach to optimizing autoencoders provides small-batch model performance while maintaining large-batch training times and is particularly useful for medical datasets with combined high global similarity and local variation. Many potential applications may benefit from this method, such as large-scale medical image retrieval algorithms that must be trained on large amounts of data. Such optimization efforts will become increasingly more

relevant as larger medical projects are funded and more robust GPU systems that can accommodate large batch sizes are employed in research settings.

Chapter IX

EHR-Defined Subtypes of Autism in Children and their Associations with Structural MRI

1. Overview

Autism is a developmental disorder characterized by social communication deficits, repetitive and restrictive behaviors, and a heterogeneous presentation of both severity and co-occurring conditions. Many studies have investigated patterns of symptoms and co-occurring conditions or differences in brain morphometry, but few studies have attempted to connect findings of clinical heterogeneity with specific structural brain differences. In this study, we analyzed joint electronic health record (EHR) and structural brain features in 124 autistic individuals who each had a clinical T1-weighted magnetic resonance image (T1w MRI) acquired between the ages of 2 and 10. From each patient's EHR, we extracted 1865 "diagnostic" phenotypes from International Classification of Diseases records and 1681 "procedural" phenotypes from Current Procedural Terminology records. After dimensionality reduction, we concatenated these sets of phenotypes and ran a clustering analysis. Each T1w MRI was segmented into 132 anatomical regions, and the volume of each region was modeled as a function of MRI age, biological sex, and cluster membership. Using this framework, we identified four clinical autism sub-types primarily characterized by (1) cognitive delays and auditory dysfunction; (2) physiological development delays; (3) convulsions; and (4) epilepsy with high prevalence of other co-occurring conditions. Six brain regions were found to be weakly associated ($p<0.05$, uncorrected) with the identified EHR subtypes: three in the left basal ganglia, two in the left temporal lobe, and one in the cerebellum. Across all regions the epilepsy cluster had the lowest region volumes. This work is the first study to our knowledge that jointly examines EHR subtypes and MRI-derived brain region volumes in autism.

2. Introduction

Approximately 1 in 44 children are diagnosed with autism by age 8 according to the most recent CDC estimate [293]. This condition is typically identified in early childhood, with primary characteristics including developmental delays, social and communication deficits, and repetitive or restrictive behaviors [294]. Despite its high prevalence, the underlying causes of autism are not yet well-understood [294]. In some respects, the most certain feature of autism is the heterogeneity of its symptoms and co-occurring conditions [294]. Developing a better understanding of sub-types within the autism spectrum and the

underlying neuroanatomy that contributes to those sub-types could lead to improved patient care and individualized support for autistic individuals.

The big data era ushered in many opportunities and tools necessary for leveraging clinical EHR systems to study complex patterns of co-occurring conditions in psychiatric disorders, including autism [295]–[298]. Two large-scale EHR cohort studies (n=14,000 and n=47,000) in autistic individuals found abnormally high levels of co-occurring conditions in autism patients compared to non-autism patients [299], [300]; though each study also found varying levels of these conditions within the autism groups, particularly with respect to age and sex. Other studies have used EHR clustering to better elucidate potential sub-types within the autism population. For example, a study of time-series EHR phenotypes uncovered three distinct clinical trajectories for autistic children under age 15: seizures, psychiatric disorders, and a mixture of gastrointestinal/infections/auditory disorders [301]. In another study, a clustering analysis identified different gradients of autism symptom severity corresponding to whether patients were primarily characterized by social communication deficits or by fixated interests and repetitive behaviors [302].

Given the cognitive dysfunction seen in autism and its many co-occurring conditions, another area of research focuses on identifying variations in brain anatomy in autism, typically via structural magnetic resonance imaging (MRI) [303]. Consistently, many such studies find that autistics exhibit increased gray matter volume compared to both typically developing controls [304], [305] and patients with attention-deficit hyperactivity disorder, which commonly co-occurs with autism [306], [307]. A recent study aimed to separate the contributions of gray matter volume and density to these findings; ultimately, their results showed no associations with density, but further confirmed this trend of increased volume in several gray matter regions for autism patients [308]. To further investigate whether brain differences correspond with symptomatic heterogeneity, one recent study compared brain region volumes of autistic people with high vs. low support needs; they found that while both groups tended to have increased gray matter volumes in the temporal lobe, people with higher support needs exhibited this increase in more widespread regions than those with lower support needs [221].

Despite the significant effort being poured into both the phenotypic comorbidity and brain structures of autism, these two areas do not often overlap. Studies using the EHR have the advantage of large sample sizes to capture a wide range of autism presentations but cannot assess the underlying brain anatomy which might contribute to any identified patterns. Conversely, studies focused on cortical structure in autism often link their findings with intelligence quotient scores [308], [309], or a small set of clinical assessments [310], [311], but these measures are limited; such brain structure comparisons cannot capture the full heterogeneity of autism.

In this study, we aimed to bridge this gap by performing a joint analysis of EHR and brain MRI for autistic children. We examined a cohort of 124 children, each with EHRs and a clinical structural whole-brain MRI between the ages of 2 and 10. In the EHR, we identified four unique sub-types of autism symptoms and co-occurring conditions. We then examined differences in brain volume for 132 anatomical regions across EHR sub-types, while accounting for age and sex effects. To our knowledge, this was the first cross-modality analysis of cortical structure and longitudinal EHR patterns in autism.

3. Material and Methods

This study and its procedures were carried out in accordance with the Institutional Review Board of Vanderbilt University and Vanderbilt University Medical Center (VUMC). Clinical EHR and MRI data were obtained in fully deidentified form from the Synthetic Derivative and ImageVU at VUMC via the Vanderbilt Institute for Clinical and Translational Research. All researchers working with this data received proper Human Subjects training. The initial cohort included 1016 individuals with clinical MRI sessions, International Classification of Diseases (ICD) records, and Current Procedural Terminology (CPT) records.

From this cohort, we selected autistic patients who had a 3D T1w turbo field echo MRI performed between the ages of 2 and 10; T1w MRIs with obviously abnormal pathology (tumors, resections, etc) were excluded. We additionally specified that each MRI session had to occur within ± 1 year of at least one ICD code for autism: [ICD-9] 299, 299.0, 299.00, 299.01, 299.1, 299.10, 299.11, 299.8, 299.80, 299.81, 299.9, 299.90, 299.91, [ICD-10] F84, F84.0, F84.1, F84.3, F84.5, F84.8, or F84.9. For individuals with multiple T1w MRIs meeting these criteria, the session in closest proximity to an ICD code for autism was chosen. For all patients, any ICD and CPT records prior to age 2 and after age 10 were removed. This selection procedure resulted in a final cohort of 124 children with autism (Table IX-1), having a collective 9748 CPT records and 13928 ICD records.

Figure IX-1 presents the joint EHR and MRI analysis pipeline. The pipeline was divided into four main phases: *EHR Processing*, *MRI Processing*, *Clustering*, and *Brain Volume Models*. Each of these steps is described in detail in the following sections.

Table IX-1 Autism Cohort Demographics

N	124
Male (%)	95 (76.6%)
Race (%)	
White	98 (79.0%)
Black	10 (8.1%)
Asian	3 (2.4%)
Unknown/Not Specified	13 (10.5%)
Mean MRI age (std)	5.44 (2.26)
Mean first autism code age (std)	4.64 (2.06)
Mean length of EHR in years (std)	3.08 (2.15)
Median N EHR codes	69

3.1. EHR processing

The first step in the EHR processing phase was extracting "diagnostic" phenotypes from the ICD records and "procedural" phenotypes from the CPT records using the pyPheWAS package [199]. Briefly, this involved mapping related groups of ICD or CPT codes to a single phenotype code. For example, ICD-9 codes 493.0 (*Extrinsic asthma*), 493.1 (*Intrinsic asthma*), and 493.82 (*Cough variant asthma*) all map to the single diagnostic phenotype 495 (*Asthma*). pyPheWAS maps ICD-9 and ICD-10 codes to a set of 1866 "diagnostic" phenotype codes (PheCodes), and CPT codes to a set of 1681 "procedural" phenotype codes (ProCodes).

The mapping step was then followed by record aggregation. PheCodes were aggregated across each patient's record, such that a 1x1866 binary PheCode vector was created for each patient, representing the presence (1) or absence (0) of each PheCode in the patient's record. All patient vectors were then stacked, and the PheCode for Autism (313.3) was removed, yielding a 124x1865 binary PheCode feature matrix. This same aggregation procedure was also applied to the ProCode data, yielding a 124x1681 binary feature matrix of ProCode presence/absence across all patient records.

Next, dimensionality reduction via principal component analysis (PCA) was performed on both the PheCode and ProCode feature matrices using the scikit-learn package [195]. The final EHR clustering dataspace was then created by trimming both PCA spaces to an equal number of PCA components and concatenating the trimmed spaces. The number of components chosen, M, was determined according to

$$M = \max(m_{phe}, m_{pro}) \quad (1)$$

where m_{phe} and m_{pro} were the number of components required to explain at least 70% of the overall variance in the PheCode and ProCode PCA spaces, respectively. For this experiment, $m_{phe} = 29$ and

$m_{pro} = 20$, so $M = m_{phe} = 29$, yielding a unified EHR dataspace of size 124 patients x 58 PCA components. This PCA procedure was used to reduce noise inherent in the EHR [49] and to balance the contributions of the PheCode and ProCode data for the subsequent clustering analysis.

3.2. MRI processing

The SLANT pipeline [139] was used to segment each T1w MRI into 132 anatomical regions from the BrainColor atlas [312]. Briefly, for each target T1w MRI, this deep learning pipeline registered the target MRI volume to the common Montreal Neurological Institute (MNI) reference space and divided the registered volume into 27 small overlapping patches. Each patch was individually segmented using a separate convolutional neural network. The overlapping segmentation patches were then combined via

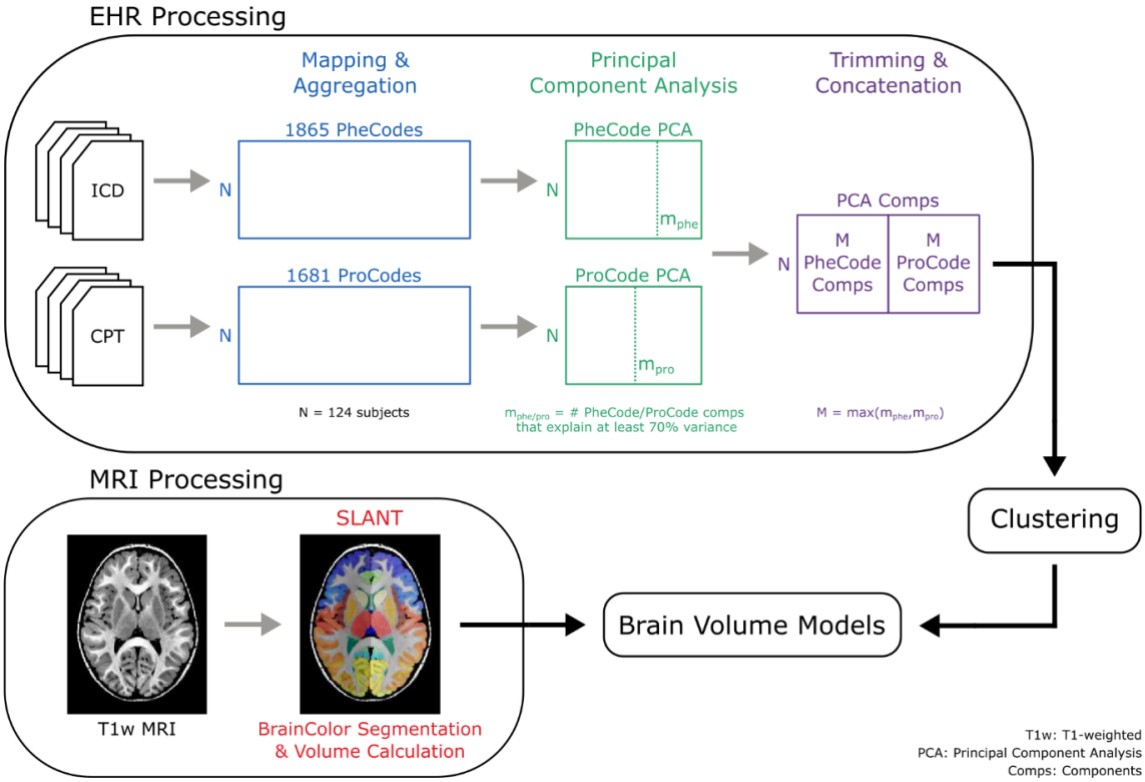


Figure IX-1 Overview of autism subtype analysis. ICD and CPT data from all patient EHRs were mapped to clinically relevant phenotypes (PheCodes and ProCodes, respectively) and aggregated across each patient's record. PCA was then performed, and *M* components from both the PheCode and ProCode PCA spaces were concatenated to form a unified EHR dataspace. A clustering analysis was performed on this unified space to identify subtypes within the EHR of autistic patients. Separately, a T1w MRI from each autistic patient was processed via SLANT, which segmented the volume into 132 anatomical regions and calculated the volume of each region. Finally, these region volumes were modeled as a function of age, sex, and EHR cluster via a general linear model.

majority vote label fusion to attain the whole-brain 132-region segmentation. The segmentation was then transformed back to the original target volume space, and the volume of each region was calculated in mm^3.

3.3. Clustering and brain volume models

With the EHR and MRI processing phases both complete, a clustering analysis was performed on the concatenated PCA spaces. Though many EHR clustering studies use hierarchical clustering with a Euclidean distance measure [278], [301], [313], we chose the spectral clustering algorithm [314] as we did not expect Euclidean distance to be a useful measure of similarity in the concatenated EHR PCA spaces. Our implementation of spectral clustering searched for 4 clusters using a nearest neighbors affinity matrix and discrete labels assignment. The scikit-learn python package [195] was again employed for the clustering analysis. Each patient was successfully assigned to a unique cluster.

After clustering, a general linear model (GLM) was estimated for each of the 132 brain regions found in section 3.2. Each GLM took the form

$$volume = Intercept + \beta_0 age + \beta_1 female + \beta_2 (age * female) + \beta_3 C_1 + \beta_4 C_2 + \beta_5 C_3 \quad (2)$$

where $volume$ was the region volume in mm^3, age was the patient's age in years at the time of their MRI scan, $female$ was a binary variable indicating whether or not the patient's biological sex was female, and $\{C_1, C_2, C_3\}$ were one-hot encoded binary variables indicating which cluster the patient belonged to. For modeling purposes, C_0 was assumed to be the reference state ($C_1 = C_2 = C_3 = 0$), and thus was not explicitly included as a predictor. After fitting the GLM for each region, an F-test was performed to determine if β_3, β_4, and β_5, were all jointly significantly not equal to zero; in other words, this tested whether or not including the EHR cluster assignments in the GLM improved the model of brain region volume. Both the GLM estimation and the F-test were performed using the statsmodels package [188].

4. Results

4.1. EHR clustering

As described in section 3.2, four unique clusters were found in the EHR data for our autism cohort (Figure IX-2); these clusters will be referred to as C_0, C_1, C_2, and C_3 throughout the rest of this article. The clusters had slightly unbalanced sizes with 22 patients in C_0, 35 patients in C_1, 37 patients in C_2, and 30 patients in C_3. Despite this size imbalance, the clusters had fairly equal sex distributions. The MRI ages for all clusters spanned nearly the entire 2 to 10 year-old range; however, C_0 and C_1 tended to have younger MRI ages and C_2 tended to have older MRI ages, while C_3 had a more even distribution. Figure IX-2

includes a 2-dimensional UMAP embedding [315] of the unified EHR space to qualitatively examine cluster separability. This embedding shows that the identified clusters were mostly well-separated in the EHR dataspace, though there was some noticeable mixing between C_0, C_1, and C_2 in the bottom right, in addition to some overlap between C_1 and C_3.

To investigate the underlying characteristics of the EHR clusters, we next examined the prevalence of PheCodes and ProCodes within them. Codes with a prevalence greater than or equal to 0.5 were considered primary cluster conditions, while codes with prevalence between 0.25 and 0.5 were considered secondary conditions. All clusters contained the ProCodes *MRI* and *Medical Service* with prevalence higher than 0.8; these codes were therefore deemed uninformative and ignored for the remainder of the analysis.

Figure IX-3 depicts prevalence for all PheCode and ProCode categories which contained at least one primary or secondary condition in any cluster. Clusters C_0, C_1, and C_2 all contained well-defined patterns of co-occuring categories with four to six primary conditions each. C_0 focused primarily on **mental disorders**, **ophthalmologic/otologic services**, and **psychiatric evaluation/therapy**. C_1 focused primarily on the **endocrine/metabolic**, **pathology**, **therapeutic procedures**, and **mental disorders** categories. C_2 focused primarily on **neurological**, **EEG**, and **chemistry/hematology labs**. In contrast, C_3 had both a more varied (18 primary conditions) and more severe (higher prevalence values) EHR structure. This cluster had prevalence values exceeding 0.75 in **ophthalmologic/otologic services**, **mental disorders**, **chemistry/hematology labs**, **EEG**, and **therapeutic procedures**. Other primary categories for C_3 included **endocrine/metabolic**, **neurological**, **psychiatric evaluation/therapy**, **pathology**, **diagnostic procedures**, and **physical therapy/rehabilitation**.

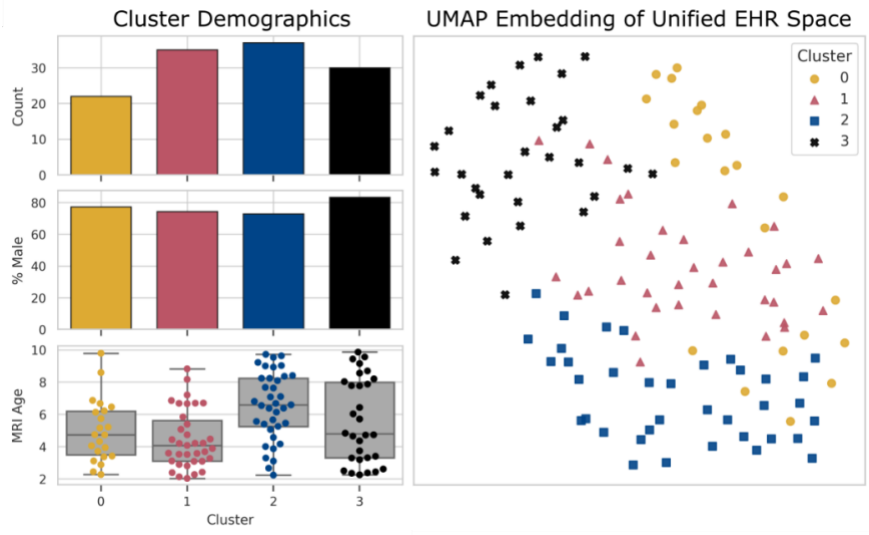


Figure IX-2 Demographics and UMAP embedding for the four EHR clusters found in the autism cohort.

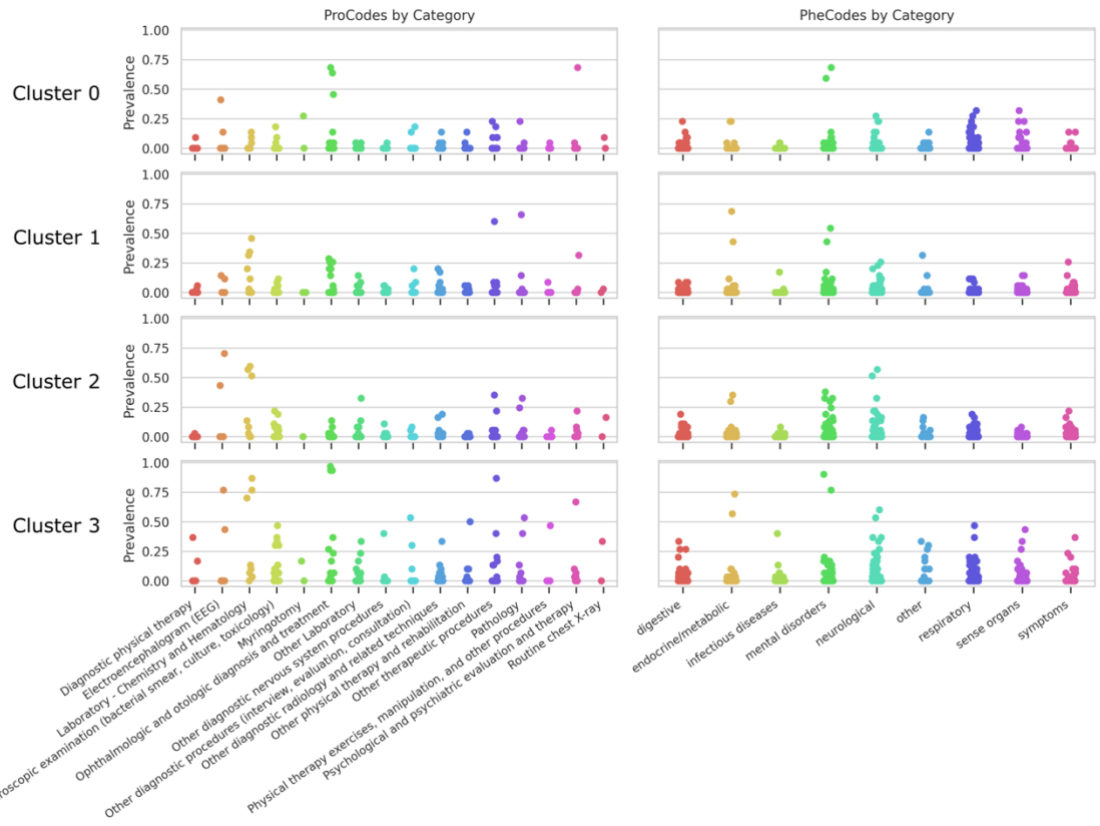


Figure IX-3 Prevalence of PheCode and ProCode categories across EHR clusters. Each point in this figure represents a unique PheCode/ProCode. Primary codes are those with prevalence ≥ 0.5. Secondary codes are those with prevalence between 0.25 and 0.5. All categories shown contain at least one primary or secondary code in any cluster.

For more detail, Table IX-2 presents all primary ProCodes and PheCodes for each cluster; a full listing of all secondary codes is included in Appendix B. All four clusters included the mental disorder PheCodes *speech/language disorder* and *developmental delays/disorders*, though at varying prevalence levels. Auditory codes (*otologic tests*, *audiometry*, *hearing loss*, etc.) were prevalent in both C_0 and C_3 as primary conditions, and several otologic ProCodes were secondary in C_1. Both C_0 and C_3 had secondary auditory codes not seen in other clusters: *tympanostomy* (C_0), *suppurative and unspecified otitis media* (C_3), and *Eustachian tube disorders* (C_3). Codes concerning physiological development were prevalent in both C_1 and C_3 (*delayed milestones, lack of normal physiological development, symptoms concerning nutrition/metabolism/development*, etc). The PheCode *convulsions* and ProCode *EEG* were joint primary codes in C_2 and C_3, and joint secondary codes in C_0. Additionally, *epilepsy/recurrent seizures/convulsions*

was a primary code for C_3 and a secondary code for C_2; C_3 included several more epilepsy-related secondary codes. Unique from other clusters, C_3 had both primary and secondary codes for speech and physical therapies, and several secondary codes related to digestive disorders.

Table IX-2 Primary conditions and procedures associated with each EHR cluster (C)

C	Prevalence	ProCode	Category	Prevalence	PheCode	Category
0	0.682	audiometry	Ophthalmologic and otologic diagnosis and treatment	0.682	Speech and language disorder	mental disorders
	0.682	Psychiatric evaluation	Psychological and psychiatric evaluation and therapy	0.591	Developmental delays and disorders	mental disorders
	0.636	otologic test	Ophthalmologic and otologic diagnosis and treatment			
1	0.657	genetic testing	Pathology	0.686	Delayed milestones	endocrine/ metabolic
	0.600	venipuncture	Other therapeutic procedures	0.543	Speech and language disorder	mental disorders
2	0.703	EEG	Electroencephalogram (EEG)	0.568	Convulsions	neurological
	0.595	hematologic tests	Laboratory - Chemistry and Hematology	0.514	Other headache syndromes	neurological
	0.568	compound specific blood test	Laboratory - Chemistry and Hematology			
	0.514	basic blood tests	Laboratory - Chemistry and Hematology			
3	0.967	otologic test	Ophthalmologic and otologic diagnosis and treatment	0.900	Speech and language disorder	mental disorders
	0.933	audiometry	Ophthalmologic and otologic diagnosis and treatment	0.767	Developmental delays and disorders	mental disorders
	0.933	acoustic test	Ophthalmologic and otologic diagnosis and treatment	0.733	Delayed milestones	endocrine/ metabolic
	0.867	compound specific blood test	Laboratory - Chemistry and Hematology	0.600	Convulsions	neurological
	0.867	venipuncture	Other therapeutic procedures	0.567	Lack of normal physiological development, unspecified	endocrine/ metabolic
	0.767	EEG	Electroencephalogram (EEG)	0.533	Epilepsy, recurrent seizures, convulsions	neurological
	0.767	hematologic tests	Laboratory - Chemistry and Hematology			
	0.700	basic blood tests	Laboratory - Chemistry and Hematology			
	0.667	Psychiatric evaluation	Psychological and psychiatric evaluation and therapy			
	0.533	Speech Therapy	Other diagnostic procedures (interview, evaluation, consultation)			
	0.533	genetic testing	Pathology			
	0.500	speech treatment	Other physical therapy and rehabilitation			

4.2. Brain volume models

As described in Section 3.3, we used a general linear model framework to estimate region volume for all 132 brain regions, followed by an F-test to gauge the usefulness of including the EHR-derived clusters in that model. This framework identified six brain regions with weakly significant ($p<0.05$, uncorrected)

associations between region volume and the EHR clusters; none of these associations survived multiple comparisons correction. Five of these six regions were in the left hemisphere. Three of the regions were in the basal ganglia (Figure IX-4): left accumbens area, left pallidum, and left putamen. The remaining three regions included the cerebellar vermal lobules (CVL) VI-VII, left temporal pole, and left transverse temporal gyrus (TTG) (Figure IX-5).

The left accumbens exhibited a diverging slope based on sex, with male region volumes increasing with age and female volumes decreasing; all other regions show similar trends of increasing or no change in volume with age, regardless of sex. Regions in the basal ganglia showed more dramatic associations with age compared to the other three regions. For all regions except the left temporal pole, C_3 had the lowest intercept and C_1 had the highest intercept. In the left temporal pole, C_1 and C_3 had nearly equivalent intercepts, lower than both C_0 and C_2.

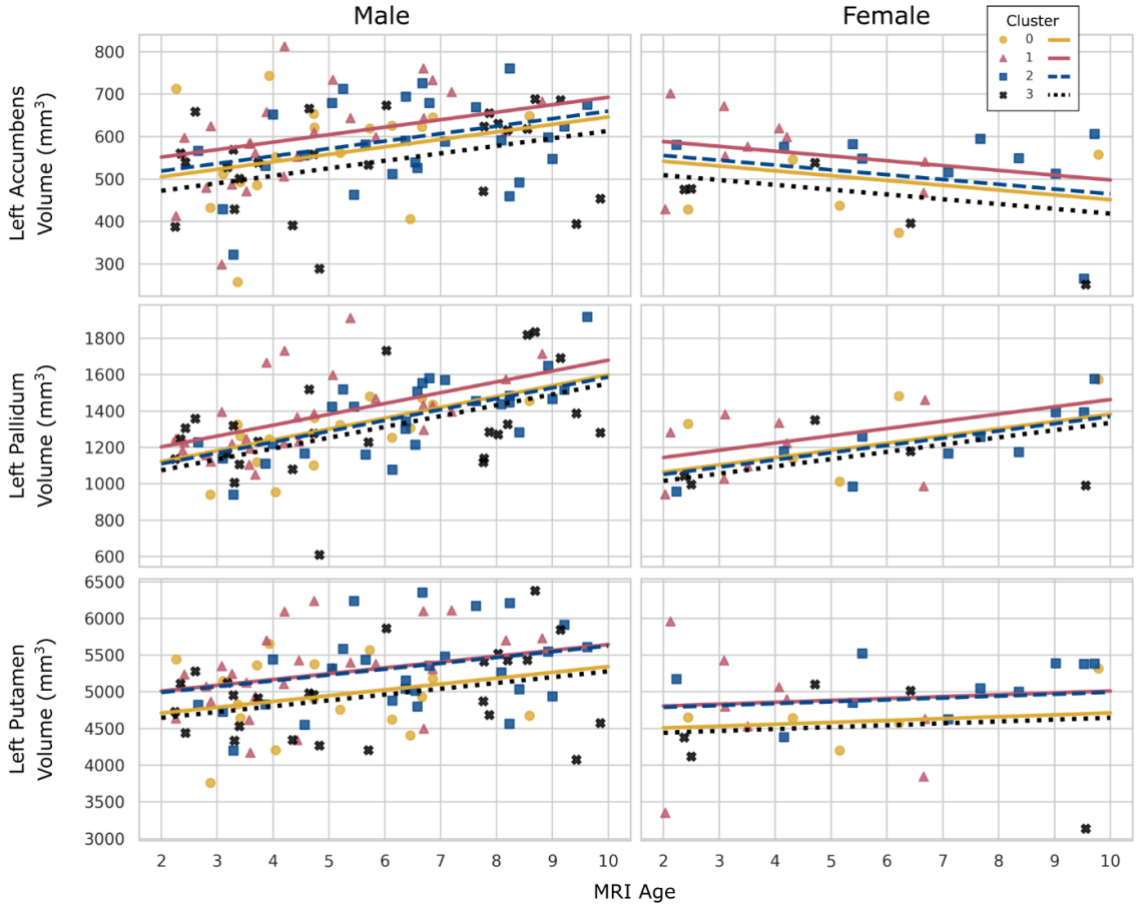


Figure IX-4 General linear models of basal ganglia region volumes as a function of age, sex, and EHR-derived clusters. All three regions shown had weakly significant ($p<0.05$) associations with EHR clusters. Models are shown split by sex (right: Male, left: Female) for clarity.

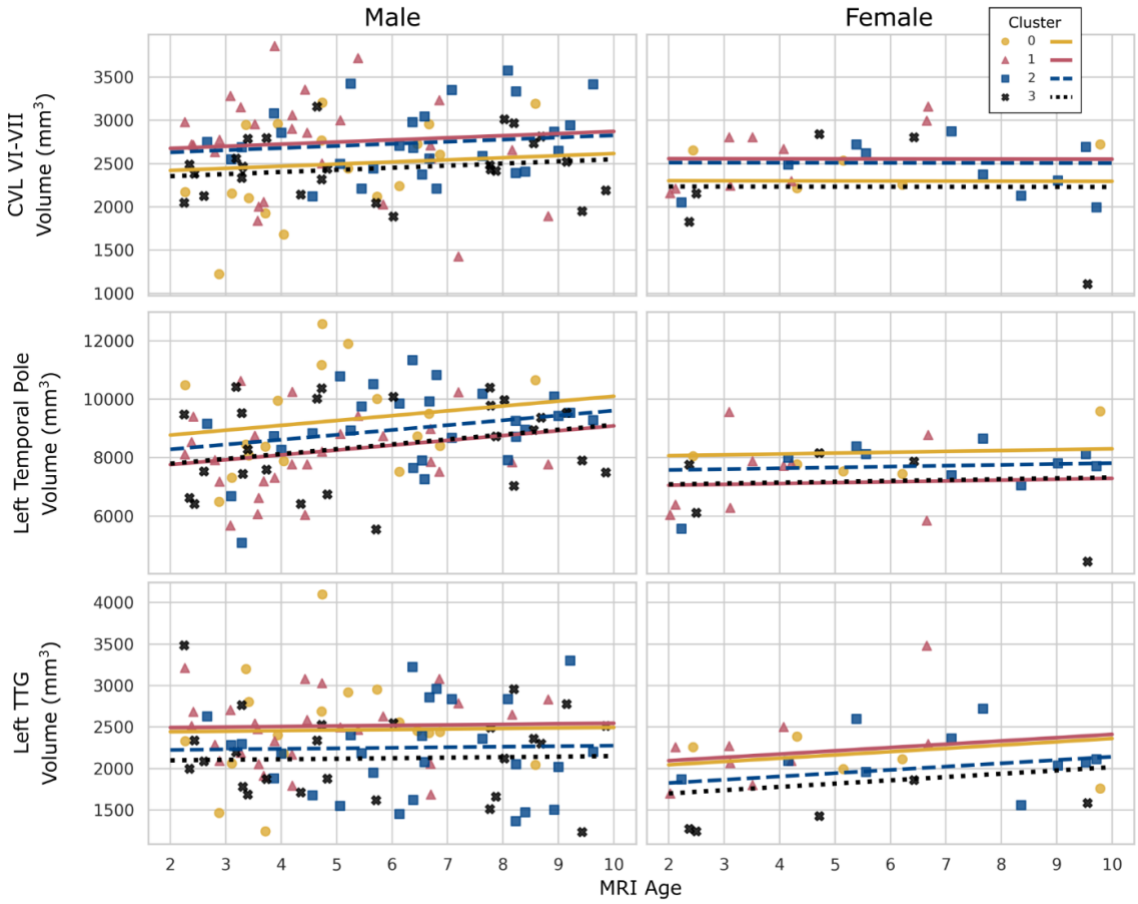


Figure IX-5 General linear models of temporal lobe and cerebellum region volumes as a function of age, sex, and EHR-derived clusters. All three regions shown had weakly significant ($p<0.05$) associations with EHR clusters. Models are shown split by sex (right: Male, left: Female) for clarity.

5. Discussion

In this work, we identified four unique EHR sub-types in autistic children aged 2-10 and examined differences in brain region volumes across the identified sub-types. Three of these sub-types (C_0, C_1, C_2) had well-defined EHR signatures, while the fourth (C_3) had a more complex EHR structure. C_0 contained patients with cognitive delays and auditory dysfunction; C_1 contained patients primarily with physiological development delays; C_2 contained patients with convulsions and perhaps mild epilepsy; and C_3 contained patients with both epilepsy and a constellation of other conditions that overlapped with the other three clusters in specific ways.

Though statistically weak, the brain region associations found in our analysis were still intriguing given the characteristics of the identified EHR clusters. One distinguishing factor across clusters was the prevalence of convulsions and epilepsy, with clusters varying from no prevalence (C_1) to high prevalence (C_3). The accumbens, pallidum, putamen, and CBV VI-VII are all regions involved in movement regulation

[316], [317]; for all four regions, C_1 had the highest region volumes while C_3 had the lowest (Figure IX-4, Figure IX-5). This suggests that decreased volume in these regions may contribute to the presentation of convulsions/epilepsy experienced by autism patients. Similarly, another distinguishing characteristic across clusters was the prevalence of auditory dysfunction, which was elevated for clusters C_0 and C_3. The left TTG is believed to be part of the early auditory processing pathway [318]; Figure IX-5 showed that, again, C_3 tended to have lower volumes in this region compared to all other clusters, but C_0 had higher volumes than both C_3 and C_2. This suggests that C_0 and C_3 may represent varied subtypes of hearing loss, with patients in C_3 linking to decreased volume in the left TTG and patients in C_0 linking to other auditory system dysfunction outside of the brain. Other EHR patterns potentially support these diverging causes, with higher prevalence of speech and language therapies seen in C_3 compared to interventions such as tympanastomy in C_0.

Despite these points of interest, none of the identified brain regions passed multiple comparisons correction, indicating that these may have been spurious associations occurring by chance. This potential lack of brain associations may be surprising given the cognitive focus of the identified autism sub-types and the high incidence of significant associations between brain volume and autism in the literature [303], [306], [308], [309]. Perhaps these weak findings should not be surprising, however, given the small size of many autism studies which report region-specific differences [304], [319]. Larger studies of autism brain morphology tend to find fewer brain abnormalities [320]. Alternatively, it could be that region volume was not an appropriate measure for this application; many other anatomical features affect cognitive function, including gray matter thickness, gyrification, and white matter connectivity. Other studies have begun to investigate the potential associations between these alternative measures of brain structure and autism [321]–[323].

This article presented a novel analysis framework for combining information from EHR and MRI in a clinical setting. As such, it could be used in other cognitive disorders to uncover sub-types of EHR progression and corresponding brain phenotypes. The potential for this framework is not, however, limitless. Acquiring sufficiently sized joint EHR and MRI datasets is challenging and expensive. Such datasets are naturally biased towards “sicker” populations; to be included, patients must have sufficient EHR data and a clinical T1w MRI, both of which indicate acute or chronic medical problems. In addition, EHR data is typically sourced from a single medical system; any events occurring outside of that medical system are unknowable, potentially leading to noisy and unreliable EHR clusters. Many disorders, including autism, are also influenced by socio-economic and environmental factors; though the current framework does not account for such factors, it could potentially be extended to include more demographic information in the EHR clustering stage.

6. Conclusions

In this study, we examined differences in the EHR of autistic children aged 2-10. We identified four unique clusters of EHR progression in this cohort, including a cognitive delay/auditory dysfunction group, a physiological developmental delays group, a convulsions groups, and an epilepsy group with complex patterns of co-occurring conditions. We next investigated the relationship between MRI-derived brain region volumes and these clusters, accounting for age and sex effects. This analysis revealed six brain regions weakly associated with the identified clusters; these regions were primarily involved in movement regulation, a finding that corresponds with the epilepsy group having the lowest region volumes.

There is a large amount of heterogeneity in the literature regarding both co-occurring conditions and brain phenotypes in autism. Studies such as this one which combine longitudinal EHR and clinical MRI could help resolve some of these discrepancies by accounting for the varied presentation of autism across individuals. With the ability to link specific brain phenotypes to specific patterns of autism expression, this approach could enable researchers to tease apart the cognitive effects of autism and its many co-occurring conditions.

Chapter X

Conclusions & Future Work

1. Introduction

Medical image processing is the art and science of extracting clinically meaningful information from medical images. One exciting facet of this field is multi-modal modeling: combining various sources of medical data into a singular model of a disease. These different data sources, from imaging MRI to EHR, each contain a unique piece of the patient story; fusing these heterogeneous sources allows imaging models to consider the whole person when creating predictions. As discussed in this dissertation, this growing area of research has many possible clinical applications but currently faces several challenges. Limited data availability in medical imaging produces models that are biased and difficult to generalize; accumulating multiple data sources for multi-modal modeling further restricts data availability and may heighten these biases. Additionally, many modeling techniques currently being employed suffer from the "black-box" problem: though models generate highly accurate predictions, the complex decision-making process that precedes the prediction is difficult or sometimes impossible to translate for human understanding.

The work presented in this dissertation investigated model interpretability as an important component for addressing limited data settings and providing explanations for predictions. We first introduced several innovations in interpretable traditional machine learning for both neuroimaging and EHR, including adapting big data analysis methods to limited multi-modal data settings (Chapter II), translating a visually interpretable machine learning framework to multi-modal analysis (Chapter III), innovating in big EHR data performance and scalability (Chapter IV), and extending the interpretability of EHR models (Chapters V and VI). Next, we investigated deep learning models as an interpretable manifold embedding method for medical data (Chapters VII and VIII). We proposed a computationally efficient unsupervised optimization technique and demonstrated that it produces interpretable manifold embeddings of both brain MRI and EHR, which may be used for secondary classification and regression tasks. Finally, this work culminated in the development of an interpretable framework for multi-modal modeling of brain MRI and EHR (Chapter IX). The proposed pipeline identified clusters in clinical EHR and explored how those clusters are related to differences in brain structure via MRI. Together, this work expands upon the growing field of model interpretability and contributes novel methodologies for multi-modal limited-data medical inference.

2. Interpretable Machine Learning with MRI and EHR

2.1. Summary

Multi-modal modeling is an exciting and innovative area of research, but it faces several challenges, including limited data and the "black box" nature of most machine learning models. In this work, we propose that model interpretability is a key factor in tackling these challenges. Our initial investigations into interpretability focused on traditional machine learning in extremely limited MRI datasets. Through a combined model of structural and diffusion-weighted MRI in mild traumatic brain injury patients, we extended traditional machine learning methods to a small data setting (Chapter II). Even in this limited dataset, we demonstrated the ability to detect novel multi-contrast MRI biomarkers and interpret those findings through the lens of symptom information. Building on this work, we refined our multi-contrast MRI model to propose the first visually interpretable joint model of structural and diffusion MRI (Chapter III). This joint model revealed coinciding anomalies in white matter connectivity and cortical structure associated with mild traumatic brain injury in a small data setting.

We next focused on innovating in traditional machine learning models for EHR data. Due to its relatively inexpensive nature, EHR tends to be available in larger datasets than MRI, so our first efforts in this modality were dedicated to improving the scalability and accessibility of PheWAS, an interpretable EHR model (Chapter IV). Using these new tools, we then demonstrated their ability to discover potentially novel EHR phenotypes in a study of children with down syndrome (Chapter VI). While working with these models, we discovered several sources of interpretability error due primarily to model accessibility. Our final contribution to this area aimed to close this accessibility gap by developing an interactive PheWAS modeling tool (Chapter V). This tool enables clinicians and other non-technical experts to more readily interpret PheWAS models by incorporating explainable machine learning principles and making model assumptions explicit.

2.2. Technical Innovations

I. We developed first of kind methods to predict individual changes in mild traumatic brain injury on small datasets.

II. We translated advances from joint functional MRI and electroencephalogram (EEG) analysis to propose the first visually interpretable joint model of structural and diffusion MRI. Our approach provides a visually interpretable model of the factors driving group differences.

III. We dramatically improved big data performance of mass univariate regression and innovated in the human-computer interface for PheWAS interpretability.

IV. We extended the interpretability of PheWAS models and integrated explainable machine learning principles into this core big data technology.

2.3. Clinical Impacts

I. We detected first evidence of combined structural and diffusion biomarkers on small data in mild traumatic brain injury.

II. Group differences between control and mild traumatic brain injury patients were substantively improved on small datasets.

III. We enabled new studies of down syndrome, revealing several significant associations between EHR phenotypes and down syndrome generally, in addition to specific phenotypes that are associated with longitudinal surgery risk in down syndrome patients with co-morbid congenital heart disease.

IV. We increased the accessibility of PheWAS studies for clinicians and other users who may not be comfortable with typical command-line interfaces, allowing them to focus solely on designing, understanding, and interpreting PheWAS models. This streamlined approach will enable EHR analysis of under-characterized populations.

2.4. Future Directions

Though we have made measurable progress in interpretable small data machine learning for multi-contrast MRI, there are still open questions to be explored. Small data modeling is primarily useful for hypothesis generation, a preliminary step on the way to designing larger studies of a medical condition. The interpretable methods introduced in this work are advantageous because they allow researchers to form more specific hypotheses about the link between chronic symptom severity and brain structure. The generalization of hypotheses generated by these methods should be more rigorously characterized in larger cohorts. Additionally, the efficacy of the proposed methods should be tested in medical conditions outside of our testbed of mild traumatic brain injury.

Our innovations in machine learning for EHR analysis have increased both the interpretability and accessibility of PheWAS. However, these studies still suffer from limitations inherent to EHR itself. The generation of EHR includes many sources of noise, including administrative mistakes, physician burnout,

and hospital-specific practices. Additionally, EHR analysis samples are typically biases towards sicker populations, since subjects must have sufficiently sized EHR records for analysis and a larger EHR tends to mean more chronic medical conditions. Future work in these interpretable EHR models should focus on methods for teasing apart these sources of bias and noise from the main medical conditions or diseases being studied.

3. Interpretable Deep Learning with MRI and EHR

3.1. Summary

Deep neural networks having rapidly become a state-of-the-art method for feature extraction and predictive analysis in medical research. With thousands of interconnected learnable parameters, these networks may be trained to accomplish tasks such as breast cancer detection or abdominal segmentation with high accuracy. Predictions made by these networks, however, are notoriously difficult to interpret; this lack of explainability significantly restricts the potential real-world usage of these models. Autoencoders are an unsupervised deep neural network architecture often used for dimensionality reduction in medical datasets. By learning to compress and then reconstruct medical data, these models create a latent embedding: a condensed representation of the input which captures hidden patterns not obvious in the original dataspace without requiring input from clinical experts. Our next set of investigations concentrated on boosting autoencoder training to produce models with more interpretable latent embeddings of MRI and EHR data.

We first focused on batch size, a hyperparameter that controls the number of individual samples a network learns from at one time. In experiments on both MRI and EHR datasets, we determined that this parameter alone contributes significantly to the quality of the latent embedding, and that counter to conventional wisdom, smaller batch sizes produce better autoencoder models than larger batch sizes (Chapter VII). Despite these positive findings, small batch sizes require substantially longer training times, making them impractical for many applications. Building on our observations from this batch size study, we next investigated a method for achieving small batch performance while maintaining large batch training times (Chapter VIII). We proposed that large batches produced poor networks due to the high level of global similarity present in medical data; when averaged across the many individuals in a large training batch, these global similarities dominated the loss landscape. To remedy this, we developed unsupervised hard case mining, an extension of hard negative/positive mining used for unbalanced training in supervised networks. This method used large batch sizes during training, but within each batch focused the network's attention on only the hardest training examples. In experiments on both EHR and MRI autoencoders, we

demonstrated that unsupervised hard case mining may accelerate autoencoder convergence and improve the interpretability of the latent space, even at larger batch sizes.

3.2. Technical Innovations

I. We characterized the significant impact that batch size has on the interpretability of deep neural network embeddings of medical data.

II. We extended hard case mining to unsupervised neural network training and demonstrated that this simple, computationally efficient technique may improve embedding interpretability and accelerate network convergence.

3.3. Clinical Impact

I. By improving the interpretability of unsupervised deep learning models, we have enhanced their potential for novel abnormality detection and phenotype discovery in EHR and MRI datasets.

3.4. Future Directions

Despite the encouraging results seen in this work, there are still many areas of opportunity to explore in creating interpretable deep learning models for MRI and EHR. Generally, we saw that small batch sizes produced more interpretable autoencoder models than large batch sizes, but for very small batch sizes there was not much difference in model quality. Based on these experiments, we suspect that the ideal batch size may be related to the amount of inter-subject variability in a dataset, but future studies should investigate this more rigorously. Our proposed unsupervised hard case mining framework showed positive results in accelerating model convergence and improving latent embedding interpretability, but these effects were not seen uniformly across both EHR and MRI. Further study is required to fully characterize these differences and close this gap in performance.

4. Interpretable Multi-Modal Modeling for MRI and EHR

4.1. Summary

The previous two contributions focused on interpretability innovations in artificial intelligence for both MRI and EHR. We have taken strides towards more interpretable EHR models, but without incorporating anatomical context from medical imaging, these models cannot assess the underlying anatomy which might

contribute to identified health patterns. Similarly, we have demonstrated improvements in limited-data MRI analysis with interpretability via symptom scores or anatomical features, but comparisons with such limited one-time clinical assessments cannot capture the full depth of clinical phenotypes seen in EHR.

Our final contribution pulled together these innovations to propose an interpretable framework for the joint analysis of MRI and EHR in a limited-data cohort of children with autism spectrum disorder (Chapter IX). This framework involved synthesizing different EHR data sources to identify subtypes of autism spectrum disorder. We then connected these subtypes with MRI-derived structural brain measures and found associations between the EHR-derived subtypes and six specific brain regions. This framework was demonstrated to be interpretable end-to-end via longitudinal EHR phenotype characteristics and previous studies of brain function for the six identified brain regions.

4.2. Technical Innovations

I. We developed a novel framework for interpretable joint analysis of longitudinal EHR subtypes and region-specific MRI-derived brain characteristics.

4.3. Clinical Impact

I. We identified novel EHR subtypes within a cohort of autism spectrum disorder patients and detected significant associations between those subtypes and six brain regions.

4.4. Future Directions

There are still many areas of opportunity in interpretable multi-modal modeling of MRI and EHR. Our proposed framework is flexible by nature, leaving room for dataset and disease specific adjustments. Future studies could expand upon our EHR subtyping method by drawing from different sources of EHR data and incorporating important demographics and socioeconomic factors that the current design does not account for. Additional work should be done to test the generalization of this framework to other MRI-derived anatomical measurements and to other medical conditions.

Appendix

A. ICD codes for defining down syndrome and intellectual and developmental disability groups

Down Syndrome

ICD Version	ICD Code	ICD Name
9	758.0	Down's syndrome
10	Q90.0	Trisomy 21; nonmosaicism (meiotic nondisjunction)
	Q90.1	Trisomy 21, mosaicism (mitotic nondisjunction)
	Q90.2	Trisomy 21, translocation
	Q90.9	Down syndrome, unspecified

Other Intellectual and Developmental Disabilities Group

ICD Version	ICD Code	ICD Name
9	314.00	Attention deficit disorder without mention of hyperactivity
	314.01	Attention deficit disorder with hyperactivity
	314.2	Hyperkinetic conduct disorder
	317	Mild intellectual disabilities
	318	Other specified intellectual disabilities
	318.0	Moderate intellectual disabilities
	318.1	Severe intellectual disabilities
	318.2	Profound intellectual disabilities
	319	Unspecified intellectual disabilities
	315.39	Other developmental speech or language disorder
	315.31	Expressive language disorder
	315.32	Mixed receptive-expressive language disorder
	315.34	Speech and language developmental delay due to hearing loss
	315.35	Childhood onset fluency disorder
	315.02	Developmental dyslexia
	315	Specific delays in development
	315.0	Developmental reading disorder
	315.00	Developmental reading disorder; unspecified
	315.09	Other specific developmental reading disorder
	315.2	Other specific developmental learning difficulties
	315.4	Developmental coordination disorder
	315.8	Other specified delays in development
	315.9	Unspecified delay in development
	299	Pervasive developmental disorders
	299.0	Autistic disorder
	299.00	Autistic disorder; current or active state
	299.01	Autistic disorder; residual state
	299.1	Childhood disintegrative disorder
	299.10	Childhood disintegrative disorder; current or active state
	299.8	Other specified pervasive developmental disorders
	299.80	Other specified pervasive developmental disorders; current or active state
	299.81	Other specified pervasive developmental disorders; residual state
	299.9	Unspecified pervasive developmental disorder
	299.90	Unspecified pervasive developmental disorder; current or active state
	330.8	Other specified cerebral degenerations in childhood
	307.21	Transient tic disorder
	307.22	Chronic motor or vocal tic disorder
	307.23	Tourette's disorder
	307.2	Tics
	307.3	Stereotypic movement disorder
	333.71	Athetoid cerebral palsy

ICD Version	ICD Code	ICD Name
9	343.8	Other specified infantile cerebral palsy
	343.9	Infantile cerebral palsy; unspecified
	759.83	Fragile X syndrome
	759.81	Prader-Willi syndrome
	799.51	Attention or concentration deficit
	799.52	Cognitive communication deficit
	799.53	Visuospatial deficit
	799.54	Psychomotor deficit
	799.55	Frontal lobe and executive function deficit
	784.52	Fluency disorder in conditions classified elsewhere
	784.59	Other speech disturbance
	784.61	Alexia and dyslexia
	315.01	Alexia
	784.69	Other symbolic dysfunction
	784.6	Other symbolic dysfunction
	784.60	Symbolic dysfunction; unspecified
	F70	Mild intellectual disabilities
	F71	Moderate intellectual disabilities
	F72	Severe intellectual disabilities
	F73	Profound intellectual disabilities
10	F78	Other intellectual disabilities
	F79	Unspecified intellectual disabilities
	F80.0	Phonological disorder
	F80.1	Expressive language disorder
	F80.2	Mixed receptive-expressive language disorder
	F80.4	Speech and language development delay due to hearing loss
	F80.81	Childhood onset fluency disorder
	F80.82	Social pragmatic communication disorder
	F80.89	Other developmental disorders of speech and language
	F80.9	Developmental disorder of speech and language; unspecified
	F81.0	Specific reading disorder
	F81.2	Mathematics disorder
	F81.81	Disorder of written expression
	F81.89	Other developmental disorders of scholastic skills
	F82	Specific developmental disorder of motor function
	F84.0	Autistic disorder
	F84.2	Rett's syndrome
	F84.3	Other childhood disintegrative disorder
	F84.5	Asperger's syndrome
	F84.8	Other pervasive developmental disorders
	F84.9	Pervasive developmental disorder; unspecified
	F88	Other disorders of psychological development
	F89	Unspecified disorder of psychological development
	F90.0	Attention-deficit hyperactivity disorder; predominantly inattentive type
	F90.1	Attention-deficit hyperactivity disorder; predominantly hyperactive type
	F90.2	Attention-deficit hyperactivity disorder; combined type
	F90.8	Attention-deficit hyperactivity disorder; other type
	F90.9	Attention-deficit hyperactivity disorder; unspecified type
	F94.0	Selective mutism
	F94.1	Reactive attachment disorder of childhood
	F94.2	Disinhibited attachment disorder of childhood
	F94.8	Other childhood disorders of social functioning
	F94.9	Childhood disorder of social functioning; unspecified
	F95.0	Transient tic disorder
	F95.1	Chronic motor or vocal tic disorder
	F95.2	Tourette's disorder
	F95.8	Other tic disorders
	F95.9	Tic disorder; unspecified
	F98.4	Stereotyped movement disorders
	F98.8	Other specified behavioral and emotional disorders with onset usually occurring in childhood and adolescence
	G11.0	Congenital nonprogressive ataxia

ICD Version	ICD Code	ICD Name
10	F98.9	Unspecified behavioral and emotional disorders with onset usually occurring in childhood and adolescence
	G11.1	Early-onset cerebellar ataxia
	G11.2	Late-onset cerebellar ataxia
	G11.3	Cerebellar ataxia with defective DNA repair
	G11.4	Hereditary spastic paraplegia
	G11.8	Other hereditary ataxias
	G11.9	Hereditary ataxia; unspecified
	G80.0	Spastic quadriplegic cerebral palsy
	G80.1	Spastic diplegic cerebral palsy
	G80.3	Athetoid cerebral palsy
	G80.4	Ataxic cerebral palsy
	G80.8	Other cerebral palsy
	G80.9	Cerebral palsy; unspecified
	G93.0	Cerebral cysts
	Q99.2	Fragile X chromosome
	Q86.0	Fetal alcohol syndrome (dysmorphic)
	Q86.8	Other congenital malformation syndromes due to known exogenous causes
	Q87.1	Congenital malformation syndromes predominantly associated with short stature
	Q93.81	Velo-cardio-facial syndrome
	Q93.88	Other microdeletions
	Q93.89	Other deletions from the autosomes
	H53.10	Unspecified subjective visual disturbances
	H53.121	Transient visual loss; right eye
	H53.122	Transient visual loss; left eye
	H53.123	Transient visual loss; bilateral
	H53.129	Transient visual loss; unspecified eye
	H53.131	Sudden visual loss; right eye
	H53.132	Sudden visual loss; left eye
	H53.133	Sudden visual loss; bilateral
	H53.139	Sudden visual loss; unspecified eye
	H53.141	Visual discomfort; right eye
	H53.142	Visual discomfort; left eye
	H53.143	Visual discomfort; bilateral
	H53.149	Visual discomfort; unspecified
	H53.15	Visual distortions of shape and size
	H53.16	Psychophysical visual disturbances
	H53.19	Other subjective visual disturbances
	H53.30	Unspecified disorder of binocular vision
	H53.31	Abnormal retinal correspondence
	H53.32	Fusion with defective stereopsis
	H53.33	Simultaneous visual perception without fusion
	H53.34	Suppression of binocular vision
	H53.40	Unspecified visual field defects
	H53.451	Other localized visual field defect; right eye
	H53.452	Other localized visual field defect; left eye
	H53.459	Other localized visual field defect; unspecified eye
	H53.453	Other localized visual field defect; bilateral
	H53.461	Homonymous bilateral field defects; right side
	H53.462	Homonymous bilateral field defects; left side
	H53.469	Homonymous bilateral field defects; unspecified side
	H53.47	Heteronymous bilateral field defects
	H53.481	Generalized contraction of visual field; right eye
	H53.482	Generalized contraction of visual field; left eye
	H53.483	Generalized contraction of visual field; bilateral
	H53.489	Generalized contraction of visual field; unspecified eye
	H53.50	Unspecified color vision deficiencies
	H53.59	Other color vision deficiencies
	H53.8	Other visual disturbances
	H53.9	Unspecified visual disturbance

ICD Version	ICD Code	ICD Name
10	H90.0	Conductive hearing loss; bilateral
	H90.2	Conductive hearing loss; unspecified
	H90.3	Sensorineural hearing loss; bilateral
	H90.41	Sensorineural hearing loss; unilateral; right ear; with unrestricted hearing on the contralateral side
	H90.42	Sensorineural hearing loss; unilateral; left ear; with unrestricted hearing on the contralateral side
	H90.5	Unspecified sensorineural hearing loss
	H90.6	Mixed conductive and sensorineural hearing loss; bilateral
	H90.71	Mixed conductive and sensorineural hearing loss; unilateral; right ear; with unrestricted hearing on the contralateral side
	H90.72	Mixed conductive and sensorineural hearing loss; unilateral; left ear; with unrestricted hearing on the contralateral side
	H90.8	Mixed conductive and sensorineural hearing loss; unspecified
	H90.A11	Conductive hearing loss; unilateral; right ear with restricted hearing on the contralateral side
	H90.A12	Conductive hearing loss; unilateral; left ear with restricted hearing on the contralateral side
	H90.A21	Sensorineural hearing loss; unilateral; right ear; with restricted hearing on the contralateral side
	H90.A22	Sensorineural hearing loss; unilateral; left ear; with restricted hearing on the contralateral side
	H90.A31	Mixed conductive and sensorineural hearing loss; unilateral; right ear with restricted hearing on the contralateral side
	H90.A32	Mixed conductive and sensorineural hearing loss; unilateral; left ear with restricted hearing on the contralateral side
	H93.25	Central auditory processing disorder
	F99	Mental disorder; not otherwise specified
	R13.0	Aphagia
	R13.1	Dysphagia
	R13.11	Dysphagia; oral phase
	R13.12	Dysphagia; oropharyngeal phase
	R13.13	Dysphagia; pharyngeal phase
	R13.14	Dysphagia; pharyngoesophageal phase
	R13.19	Other dysphagia
	R41.9	Unspecified symptoms and signs involving cognitive functions and awareness
	R41.1	Anterograde amnesia
	R41.2	Retrograde amnesia
	R41.3	Other amnesia
	R41.81	Age-related cognitive decline
	R41.82	Altered mental status; unspecified
	R41.83	Borderline intellectual functioning
	R41.840	Attention and concentration deficit
	R41.841	Cognitive communication deficit
	R41.842	Visuospatial deficit
	R41.843	Psychomotor deficit
	R41.844	Frontal lobe and executive function deficit
	R41.89	Other symptoms and signs involving cognitive functions and awareness
	R44.0	Auditory hallucinations
	R44.1	Visual hallucinations
	R44.2	Other hallucinations
	R44.8	Other symptoms and signs involving general sensations and perceptions
	R44.9	Unspecified symptoms and signs involving general sensations and perceptions
	R47.82	Fluency disorder in conditions classified elsewhere
	R47.89	Other speech disturbances
	R47.9	Unspecified speech disturbances
	R48.0	Dyslexia and alexia
	R48.1	Agnosia
	R48.2	Apraxia
	R48.8	Other symbolic dysfunctions
	R48.9	Unspecified symbolic dysfunctions
	R62.0	Delayed milestone in childhood
	R62.50	Unspecified lack of expected normal physiological development in childhood
	R62.51	Failure to thrive (child)
	R62.52	Short stature (child)
	R62.59	Other lack of expected normal physiological development in childhood

B. Secondary conditions and procedures associated with autism subtypes

C	Prevalence	ProCode	Category	Prevalence	PheCode	Category
0	0.455	acoustic test	Ophthalmologic and otologic diagnosis and treatment	0.318	Chronic tonsillitis and adenoiditis	respiratory
	0.409	EEG	Electroencephalogram (EEG)	0.318	Hearing loss	sense organs
	0.273	tympanostomy	Myringotomy	0.273	Convulsions	neurological
				0.273	Chronic pharyngitis and nasopharyngitis	respiratory
1	0.457	compound specific blood test	Laboratory - Chemistry and Hematology	0.429	Lack of normal physiological development, unspecified	endocrine/ metabolic
	0.343	basic blood tests	Laboratory - Chemistry and Hematology	0.429	Developmental delays and disorders	mental disorders
	0.314	hematologic tests	Laboratory - Chemistry and Hematology	0.314	Symptoms concerning nutrition, metabolism, and development	other
	0.314	Psychiatric evaluation	Psychological and psychiatric evaluation and therapy	0.257	Sleep disorders	neurological
	0.286	otologic test	Ophthalmologic and otologic diagnosis and treatment	0.257	Symptoms involving nervous and musculoskeletal systems	symptoms
	0.257	Ophthalmological service	Ophthalmologic and otologic diagnosis and treatment			
	0.257	audiometry	Ophthalmologic and otologic diagnosis and treatment			
2	0.432	monitoring	Electroencephalogram (EEG)	0.378	Attention deficit hyperactivity disorder	mental disorders
	0.351	venipuncture	Other therapeutic procedures	0.351	Lack of normal physiological development, unspecified	endocrine/ metabolic
	0.324	hematologic labs	Other Laboratory	0.324	Developmental delays and disorders	mental disorders
	0.324	genetic testing	Pathology	0.324	Transient alteration of awareness	mental disorders
				0.324	Epilepsy, recurrent seizures, convulsions	neurological
				0.297	Delayed milestones	endocrine/ metabolic
				0.297	Speech and language disorder	mental disorders
3	0.467	bacterial culture	Microscopic examination (bacterial smear, culture, toxicology)	0.467	Acute upper respiratory infections of multiple or unspecified sites	respiratory
	0.467	Therapeutic procedure	Physical therapy exercises, manipulation, and other procedures	0.433	Hearing loss	sense organs
	0.433	monitoring	Electroencephalogram (EEG)	0.400	Viral infection	infectious diseases
	0.400	non-invasive	Other diagnostic nervous system procedures	0.367	Sleep disorders	neurological
	0.400	perioperative management	Other therapeutic procedures	0.367	Partial epilepsy	neurological
	0.400	histopathology	Pathology	0.367	Chronic tonsillitis and adenoiditis	respiratory
	0.367	occupational therapy	Diagnostic physical therapy	0.367	Fever of unknown origin	symptoms
	0.367	Infectious agent antigen detection by immunoassay	Microscopic examination (bacterial smear, culture, toxicology)	0.333	Constipation	digestive
	0.367	eye exam	Ophthalmologic and otologic diagnosis and treatment	0.333	Lack of coordination	neurological
	0.333	other culture	Microscopic examination (bacterial smear, culture, toxicology)	0.333	Symptoms concerning nutrition, metabolism, and development	other

3	0.333	hematologic labs	Other Laboratory	0.333	Suppurative and unspecified otitis media	sense organs
	0.333	abdomen radiologic exam	Other diagnostic radiology and related techniques	0.300	Other ill-defined and unknown causes of morbidity and mortality	other
	0.333	Radiologic exam, chest	Routine chest X-ray	0.267	GERD	digestive
	0.300	Urinalysis	Microscopic examination (bacterial smear, culture, toxicology)	0.267	Dysphagia	digestive
	0.300	other microscopic examination	Microscopic examination (bacterial smear, culture, toxicology)	0.267	Abnormal involuntary movements	neurological
	0.300	sleep study	Other diagnostic procedures (interview, evaluation, consultation)	0.267	Other tests	other
	0.267	speech test	Ophthalmologic and otologic diagnosis and treatment	0.267	Eustachian tube disorders	sense organs

REFERENCES

[1] P. Suetens, *Fundamentals of Medical Imaging*. Cambridge University Press, 2002.

[2] D. C. Castro, I. Walker, and B. Glocker, "Causality matters in medical imaging," *Nat. Commun.*, vol. 11, no. 1, pp. 1–10, Dec. 2020, doi: 10.1038/s41467-020-17478-w.

[3] W. M. Wells, "Medical Image Analysis – past, present, and future," *Medical Image Analysis*, vol. 33. Elsevier B.V., pp. 1339–1351, Oct. 01, 2016, doi: 10.1016/j.media.2016.06.013.

[4] J. S. Duncan and N. Ayache, "Medical image analysis: Progress over two decades and the challenges ahead," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 85–106, Jan. 2000, doi: 10.1109/34.824822.

[5] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science (80-.).*, vol. 349, no. 6245, pp. 255–260, Jul. 2015, doi: 10.1126/science.aaa8415.

[6] S. M. Shortreed, A. J. Cook, R. Y. Coley, J. F. Bobb, and J. C. Nelson, "Challenges and Opportunities for Using Big Health Care Data to Advance Medical Science and Public Health," *Am. J. Epidemiol.*, vol. 188, no. 5, pp. 851–861, 2019, doi: 10.1093/aje/kwy292.

[7] J. D. Van Horn and A. W. Toga, "Human neuroimaging as a 'Big Data' science," *Brain Imaging Behav.*, vol. 8, no. 2, pp. 323–331, Oct. 2014, doi: 10.1007/s11682-013-9255-y.

[8] S. Zhang and D. Metaxas, "Large-Scale medical image analytics: Recent methodologies, applications and Future directions," *Medical Image Analysis*, vol. 33. Elsevier B.V., pp. 98–101, Oct. 01, 2016, doi: 10.1016/j.media.2016.06.010.

[9] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[10] S. Wang and R. M. Summers, "Machine learning and radiology," *Med. Image Anal.*, vol. 16, no. 5, pp. 933–951, Jul. 2012, doi: 10.1016/j.media.2012.02.005.

[11] A. J. London, "Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability," John Wiley and Sons Inc., Feb. 2019. doi: 10.1002/hast.973.

[12] V. Miglani and M. Bhatia, "Skin lesion classification: a transfer learning approach using efficientnets," in *Advances in Intelligent Systems and Computing*, Feb. 2021, vol. 1141, pp. 315–324, doi: 10.1007/978-981-15-3383-9_29.

[13] Y. Tang *et al.*, "High-resolution 3D abdominal segmentation with random patch network fusion," *Med. Image Anal.*, vol. 69, p. 101894, Apr. 2021, doi: 10.1016/j.media.2020.101894.

[14] D. Rueckert, B. Glocker, and B. Kainz, "Learning clinically useful information from images: Past, present and future," *Med. Image Anal.*, vol. 33, pp. 1339–1351, Oct. 2016, doi: 10.1016/j.media.2016.06.009.

[15] M. de Bruijne, "Machine learning approaches in medical image analysis: From detection to diagnosis," *Med. Image Anal.*, vol. 33, pp. 94–97, Oct. 2016, doi: 10.1016/j.media.2016.06.032.

[16] D. S. Knopman, T. H. Mosley, D. J. Catellier, and L. H. Coker, "Fourteen-year longitudinal study of vascular risk factors, APOE genotype, and cognition: The ARIC MRI Study," *Alzheimer's Dement.*, vol. 5, no. 3, pp. 207–214, 2009, doi: 10.1016/j.jalz.2009.01.027.

[17] S. Chaganti, K. P. Nabar, K. M. Nelson, L. A. Mawn, and B. A. Landman, "Phenotype Analysis of Early Risk Factors from Electronic Medical Records Improves Image-Derived Diagnostic Classifiers for Optic Nerve Pathology," in *Proc. SPIE 10138, Medical Imaging 2017: Imaging Informatics for Healthcare, Research, and Applications*, 2017, pp. 100–106, doi: 10.1117/12.2254618.

[18] Q. Zhang *et al.*, "Discriminative Analysis of Migraine without Aura: Using Functional and Structural MRI with a Multi-Feature Classification Approach," *PLoS One*, vol. 11, no. 9, p. e0163875, Sep. 2016, doi: 10.1371/journal.pone.0163875.

[19] J. Platt and others, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Adv. large margin Classif.*, vol. 10, no. 3, pp. 61–74, 1999, Accessed: Apr. 04, 2021. [Online]. Available: https://www.researchgate.net/publication/2594015.

[20] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[21] D. Greenstein, J. D. Malley, B. Weisinger, L. Clasen, and N. Gogtay, "Using Multivariate Machine Learning Methods and Structural MRI to Classify Childhood Onset Schizophrenia and Healthy Controls," *Front. Psychiatry*, vol. 3, p. 53, Jun. 2012, doi: 10.3389/fpsyt.2012.00053.

[22] F. Y. Kuo and I. H. Sloan, "Lifting the Curse of Dimensionality," *Not. AMS*, vol. 52, no. 11, pp. 1320–1328, 2005.

[23] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.

[24] K. Suzuki, “Overview of deep learning in medical imaging,” doi: 10.1007/s12194-017-0406-5.
[25] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, “Medical Image Analysis using Convolutional Neural Networks: A Review,” *J. Med. Syst.*, vol. 42, no. 11, pp. 1–13, Nov. 2018, doi: 10.1007/s10916-018-1088-1.
[26] A. Esteva *et al.*, “Deep learning-enabled medical computer vision,” *npj Digit. Med.*, vol. 4, no. 5, pp. 1–9, 2021, doi: 10.1038/s41746-020-00376-2.
[27] N. Srivastava, G. Hinton, A. Krizhevsky, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” 2014.
[28] R. Bruffaerts, “Machine learning in neurology: what neurologists can learn from machines and vice versa,” *J. Neurol.*, vol. 265, no. 11, pp. 2745–2748, Aug. 2018, doi: 10.1007/s00415-018-8990-9.
[29] M. D. Failla, K. L. Schwartz, S. Chaganti, L. E. Cutting, B. A. Landman, and C. J. Cascio, “Using phecode analysis to characterize co-occurring medical conditions in autism spectrum disorder,” *Autism*, 2020, doi: 10.1177/1362361320934561.
[30] N. C. Fox *et al.*, “Presymptomatic hippocampal atrophy in Alzheimer’s disease A longitudinal MRI study,” 1996. Accessed: Apr. 05, 2021. [Online]. Available: https://academic.oup.com/brain/article/119/6/2001/466591.
[31] Y. Xu, “Deep learning in multimodal medical image analysis,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Oct. 2019, vol. 11837 LNCS, pp. 193–200, doi: 10.1007/978-3-030-32962-4_18.
[32] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, “Pixel-level image fusion: A survey of the state of the art,” *Inf. Fusion*, vol. 33, pp. 100–112, Jan. 2017, doi: 10.1016/j.inffus.2016.05.004.
[33] G. Bhatnagar, Q. M. J. Wu, and Z. Liu, “A new contrast based multimodal medical image fusion framework,” *Neurocomputing*, vol. 157, pp. 143–152, Jun. 2015, doi: 10.1016/j.neucom.2015.01.025.
[34] B. Huang, F. Yang, M. Yin, X. Mo, and C. Zhong, “A Review of Multimodal Medical Image Fusion Techniques,” *Computational and Mathematical Methods in Medicine*, vol. 2020. Hindawi Limited, 2020, doi: 10.1155/2020/8279342.
[35] T.-D. Vu, N.-H. Ho, H.-J. Yang, J. Kim, and H.-C. Song, “Non-white matter tissue extraction and deep convolutional neural network for Alzheimer’s disease detection,” *Soft Comput.*, vol. 22, pp. 6825–6833, 2018, doi: 10.1007/s00500-018-3421-5.
[36] L. Sun, S. Zhang, H. Chen, and L. Luo, “Brain Tumor Segmentation and Survival Prediction Using Multimodal MRI Scans With Deep Learning,” *Front. Neurosci.*, vol. 13, no. JUL, p. 810, Aug. 2019, doi: 10.3389/fnins.2019.00810.
[37] Y. Chang, T. Huang, Y. Liu, H. Chung, and C. Juan, “Classification of parotid gland tumors by using multimodal MRI and deep learning,” *NMR Biomed.*, vol. 34, no. 1, p. e4408, Jan. 2021, doi: 10.1002/nbm.4408.
[38] H.-I. Suk, S.-W. Lee, D. Shen, and ADNI, “Subclass-based multi-task learning for Alzheimer’s disease diagnosis,” *Front. Aging Neurosci.*, vol. 6, pp. 1–20, 2014, doi: 10.3389/fnagi.2014.00168.
[39] M. Liu, D. Cheng, K. Wang, and Y. Wang, “Multi-Modality Cascaded Convolutional Neural Networks for Alzheimer’s Disease Diagnosis,” *Neuroinformatics*, vol. 16, no. 3–4, pp. 295–308, Oct. 2018, doi: 10.1007/s12021-018-9370-4.
[40] H. Yang, J. Zhang, Q. Liu, and Y. Wang, “Multimodal MRI-based classification of migraine: Using deep learning convolutional neural network 08 Information and Computing Sciences 0801 Artificial Intelligence and Image Processing,” *Biomed. Eng. Online*, vol. 17, no. 1, p. 138, Oct. 2018, doi: 10.1186/s12938-018-0587-0.
[41] M. R. Karim *et al.*, “DeepKneeExplainer: Explainable Knee Osteoarthritis Diagnosis from Radiographs and Magnetic Resonance Imaging,” *IEEE Access*, 2021, doi: 10.1109/ACCESS.2021.3062493.
[42] M. H. Le *et al.*, “Automated diagnosis of prostate cancer in multi-parametric MRI based on multimodal convolutional neural networks,” *Phys. Med. Biol.*, vol. 62, no. 16, pp. 6497–6514, Jul. 2017, doi: 10.1007/978-3-319-66179-7_49.
[43] X. Yang *et al.*, “Joint Detection and Diagnosis of Prostate Cancer in Multi-parametric MRI Based on Multimodal Convolutional Neural Networks,” in *Medical Image Computing and Computer Assisted Intervention − MICCAI 2017*, 2017, pp. 426–434, doi: 10.1007/978-3-319-66179-7_49.
[44] E. Zihni *et al.*, “Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome,” *PLoS One*, vol. 15, no. 4, p. e0231166, Apr. 2020, doi: 10.1371/journal.pone.0231166.
[45] D. T. A. Luong *et al.*, “Extracting Deep Phenotypes for Chronic Kidney Disease Using Electronic Health Records,” *EGEMS*, vol. 5, no. 1, p. 9, Jun. 2017, doi: 10.5334/egems.226.

[46] C. Lee and M. Van Der Schaar, "Temporal Phenotyping using Deep Predictive Clustering of Disease Progression," in *International Conference on Machine Learning*, 2020.

[47] J. C. Denny *et al.*, "PheWAS: Demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations," *Bioinformatics*, vol. 26, no. 9, pp. 1205–1210, Mar. 2010, doi: 10.1093/bioinformatics/btq126.

[48] S. Chaganti *et al.*, "Electronic Medical Record Context Signatures Improve Diagnostic Classification Using Medical Image Computing," *IEEE J. Biomed. Heal. INFORMATICS*, vol. 23, no. 5, pp. 2052–2062, 2019, doi: 10.1017/9781316671849.008.

[49] K. J. O'Malley, K. F. Cook, M. D. Price, K. R. Wildes, J. F. Hurdle, and C. M. Ashton, "Measuring diagnoses: ICD code accuracy," *Health Serv. Res.*, vol. 40, no. 5 II, pp. 1620–1639, 2005, doi: 10.1111/j.1475-6773.2005.00444.x.

[50] S. Chaganti, J. R. Robinson, C. Bermudez, T. Lasko, L. A. Mawn, and B. A. Landman, "EMR-Radiological Phenotypes in Diseases of the Optic Nerve and their Association with Visual Function," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2017, pp. 373–381, doi: 10.1007/978-3-319-67558-9.

[51] I. Carmichael *et al.*, "Joint and individual analysis of breast cancer histologic images and genomic covariates." arXiv, Dec. 01, 2019, Accessed: Mar. 19, 2021. [Online]. Available: http://arxiv.org/abs/1912.00434.

[52] Z. Zhang, P. Chen, M. Sapkota, and L. Yang, "TandemNet: Distilling knowledge from medical images using diagnostic reports as optional semantic references," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Sep. 2017, vol. 10435 LNCS, pp. 320–328, doi: 10.1007/978-3-319-66179-7_37.

[53] Z. Zhang, P. Chen, X. Shi, and L. Yang, "Text-guided Neural Network Training for Image Recognition in Natural Scenes and Medicine," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2019, doi: 10.1109/TPAMI.2019.2955476.

[54] G. Hripcsak and D. J. Albers, "Next-generation phenotyping of electronic health records," *J. Am. Med. Informatics Assoc.*, vol. 20, no. 1, pp. 117–121, 2013, doi: 10.1136/amiajnl-2012-001145.

[55] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, Sep. 2018, doi: 10.1109/ACCESS.2018.2870052.

[56] G. Montavon, W. Samek, and K. R. Müller, "Methods for interpreting and understanding deep neural networks," *Digit. Signal Process. A Rev. J.*, vol. 73, pp. 1–15, Feb. 2018, doi: 10.1016/j.dsp.2017.10.011.

[57] D. Carvalho, E. Pereira, and J. Cardoso, "Machine Learning Interpretability: A Survey on Methods and Metrics," *Electron.*, vol. 8, no. 8, 2019, doi: 10.3390/electronics8080832.

[58] A. Barredo Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[59] N. Xie, G. Ras, M. van Gerven, and D. Doran, "Explainable Deep Learning: A Field Guide for the Uninitiated," *arXiv*, Apr. 2020, Accessed: Mar. 19, 2021. [Online]. Available: http://arxiv.org/abs/2004.14545.

[60] J. Townsend, T. Chaton, and J. M. Monteiro, "Extracting Relational Explanations from Deep Neural Networks: A Survey from a Neural-Symbolic Perspective," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 31, no. 9, pp. 3456–3470, Sep. 2020, doi: 10.1109/TNNLS.2019.2944672.

[61] F. Eitel and K. Ritter, "Testing the robustness of attribution methods for convolutional neural networks in MRI-based Alzheimer's disease classification," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Oct. 2019, vol. 11797 LNCS, pp. 3–11, doi: 10.1007/978-3-030-33850-3_1.

[62] M. Dyrba, A. H. Pallath, and E. N. Marzban, "Comparison of CNN visualization methods to aid model interpretability for detecting alzheimer's disease," in *Bildverarbeitung für die Medizin 2020*, 2020, pp. 307–312, doi: 10.1007/978-3-658-29267-6_68.

[63] J. Adebayo *et al.*, "Sanity Checks for Saliency Maps," in *NeuRIPS 2018*, 2018, Accessed: Apr. 01, 2021. [Online]. Available: https://goo.gl/hBmhDt.

[64] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for Deep Neural Networks," in *International Conference on Learning Representations (ICLR) 2018*, Feb. 2018.

[65] J. Dodge, Q. Vera Liao, Y. Zhang, R. K. E. Bellamy, and C. Dugan, "Explaining models: An empirical study of how explanations impact fairness judgment," *Int. Conf. Intell. User Interfaces, Proc. IUI*, vol. Part F1476, pp. 275–285, 2019, doi: 10.1145/3301275.3302310.

[66] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267,

2019, doi: 10.1016/j.artint.2018.07.007.
[67] L. H. Gilpin, C. Testart, N. Fruchter, and J. Adebayo, "Explaining Explanations to Society," in *32nd Conference on Neural Information Processing Systems (NIPS 2018)*, 2018, Accessed: Mar. 19, 2021. [Online]. Available: http://arxiv.org/abs/1901.06560.
[68] M. Ribera and A. Lapedriza, "Can we do better explanations? A proposal of user-centered explainable AI," in *Joint Proceedings of the ACM IUI Workshops*, 2019.
[69] Y. Liang, S. Li, C. Yan, M. Li, and C. Jiang, "Explaining the black-box model: A survey of local interpretation methods for deep neural networks," *Neurocomputing*, vol. 419, pp. 168–182, Jan. 2021, doi: 10.1016/j.neucom.2020.08.011.
[70] D. T. Huff, A. J. Weisman, and R. Jeraj, "Interpretation and visualization techniques for deep learning models in medical imaging," *Phys. Med. Biol.*, vol. 66, no. 4, p. 04TR01, Feb. 2021, doi: 10.1088/1361-6560/abcd17.
[71] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 9, no. 4, 2019, doi: 10.1002/widm.1312.
[72] A. P. Engelbrecht, I. Cloete, and J. M. Zurada, "Determining the significance of input parameters using sensitivity analysis," in *From Natural to Artificial Neural Computation. IWANN 1995.*, J. Mira and F. Sandoval, Eds. Springer Berlin Heidelberg, 1995, pp. 382–388.
[73] F. Chollet, "Grad-CAM class activation visualization," 2021. https://keras.io/examples/vision/grad_cam/ (accessed Mar. 04, 2021).
[74] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K. R. Müller, "Explaining nonlinear classification decisions with deep Taylor decomposition," *Pattern Recognit.*, vol. 65, pp. 211–222, May 2017, doi: 10.1016/j.patcog.2016.11.008.
[75] J. Kauffmann, K. R. Müller, and G. Montavon, "Towards explaining anomalies: A deep Taylor decomposition of one-class models," *Pattern Recognit.*, vol. 101, p. 107198, May 2020, doi: 10.1016/j.patcog.2020.107198.
[76] I. Sturm, S. Lapuschkin, W. Samek, and K. R. Müller, "Interpretable deep neural networks for single-trial EEG classification," *J. Neurosci. Methods*, vol. 274, pp. 141–145, Dec. 2016, doi: 10.1016/j.jneumeth.2016.10.008.
[77] H. Lee *et al.*, "An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets," *Nat. Biomed. Eng.*, vol. 3, pp. 173–182, 2019, doi: 10.1038/s41551-018-0324-9.
[78] I. Grigorescu, L. Cordero-Grande, D. Edwards, J. Hajnal, M. Modat, and M. Deprez, "Interpretable Convolutional Neural Networks for Preterm Birth Classification," in *Medical Imaging with Deep Learning 2019*, 2015, Accessed: Apr. 01, 2021. [Online]. Available: http://www.developingconnectome.org/.
[79] T. Nakagawa, M. Ishida, J. Naito, A. Nagai, S. Yamaguchi, and K. Onoda, "Prediction of conversion to Alzheimer's disease using deep survival analysis of MRI images," *Brain Commun.*, vol. 2, no. 1, Jan. 2020, doi: 10.1093/braincomms/fcaa057.
[80] M. R. Karim, T. Dohmen, M. Cochez, O. Beyan, D. Rebholz-Schuhmann, and S. Decker, "DeepCOVIDExplainer: Explainable COVID-19 Diagnosis from Chest X-ray Images," in *Proceedings - 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020*, Dec. 2020, pp. 1034–1037, doi: 10.1109/BIBM49941.2020.9313304.
[81] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014, vol. 8689 LNCS, no. PART 1, pp. 818–833, doi: 10.1007/978-3-319-10590-1_53.
[82] M. Aminu, N. A. Ahmad, and M. H. Mohd Noor, "Covid-19 detection via deep neural network and occlusion sensitivity maps," *Alexandria Eng. J.*, Mar. 2021, doi: 10.1016/j.aej.2021.03.052.
[83] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' Explaining the predictions of any classifier," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-Augu, pp. 1135–1144, 2016, doi: 10.1145/2939672.2939778.
[84] P. R. Magesh, R. D. Myloth, and R. J. Tom, "An Explainable Machine Learning Model for Early Detection of Parkinson's Disease using LIME on DaTSCAN Imagery," *Comput. Biol. Med.*, vol. 126, Nov. 2020, doi: 10.1016/j.compbiomed.2020.104041.
[85] C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin, "This Looks Like That: Deep Learning for Interpretable Image Recognition," in *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, Jun. 2019.
[86] O. Li, H. Liu, C. Chen, and C. Rudin, "Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains Its Predictions," *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, pp. 3530–3537, 2018, Accessed: Mar. 30, 2021. [Online]. Available: www.aaai.org.

[87] W. Samek, T. Wiegand, and K.-R. Müller, "EXPLAINABLE ARTIFICIAL INTELLIGENCE: UNDERSTANDING, VISUALIZING AND INTERPRETING DEEP LEARNING MODELS," *ITU J. ICT Discov.*, vol. Special Issue No. 1, 2017.

[88] J. D. Cassidy *et al.*, "Incidence, risk factors and prevention of mild traumatic brain injury: results of the WHO collaborating centre task force on mild traumatic brain injury," *J. Rehabil. Med.*, vol. 36, pp. 28–60, Feb. 2004, doi: 10.1080/16501960410023732.

[89] L. K. Lindquist, H. C. Love, and E. B. Elbogen, "Traumatic Brain Injury in Iraq and Afghanistan Veterans: New Results From a National Random Sample Study," *J. Neuropsychiatry Clin. Neurosci.*, vol. 29, no. 3, pp. 254–259, Jul. 2017, doi: 10.1176/appi.neuropsych.16050100.

[90] C. W. Hoge, D. McGurk, J. L. Thomas, A. L. Cox, C. C. Engel, and C. A. Castro, "Mild Traumatic Brain Injury in U.S. Soldiers Returning from Iraq," *N. Engl. J. Med.*, vol. 358, no. 5, pp. 453–463, Jan. 2008, doi: 10.1056/NEJMoa072972.

[91] K. McInnes, C. L. Friesen, D. E. MacKenzie, D. A. Westwood, and S. G. Boe, "Mild Traumatic Brain Injury (mTBI) and chronic cognitive impairment: A scoping review," *PLoS One*, vol. 12, no. 4, 2017, doi: 10.1371/journal.pone.0174847.

[92] V. L. Kristman *et al.*, "Methodological issues and research recommendations for prognosis after mild traumatic brain injury: Results of the international collaboration on mild traumatic brain injury prognosis," *Arch. Phys. Med. Rehabil.*, vol. 95, no. 3 SUPPL, 2014, doi: 10.1016/j.apmr.2013.04.026.

[93] E. D. Bigler *et al.*, "Heterogeneity of brain lesions in pediatric traumatic brain injury," *Neuropsychology*, vol. 27, no. 4, pp. 438–451, 2013, doi: 10.1037/a0032837.

[94] P. Dall'Acqua *et al.*, "Prefrontal Cortical Thickening after Mild Traumatic Brain Injury: A One-Year Magnetic Resonance Imaging Study," *J. Neurotrauma*, vol. 34, no. 23, pp. 3270–3279, 2017, doi: 10.1089/neu.2017.5124.

[95] G. I. Guberman, J. C. Houde, A. Ptito, I. Gagnon, and M. Descoteaux, "Structural abnormalities in thalamo-prefrontal tracks revealed by high angular resolution diffusion imaging predict working memory scores in concussed children," *Brain Struct. Funct.*, vol. 225, no. 1, pp. 441–459, 2020, doi: 10.1007/s00429-019-02002-8.

[96] B. Zablotsky *et al.*, "Prevalence and trends of developmental disabilities among children in the United States: 2009–2017," *Pediatrics*, vol. 144, no. 4, Oct. 2019, doi: 10.1542/PEDS.2019-0811/76974.

[97] C. J. Epstein, "Down Syndrome (Trisomy 21)," in *The Metabolic and Molecular Bases of Inherited Disease*, C. R. Scriver, A. L. Beaudet, W. S. Sly, and D. Valle, Eds. New York, NY: McGraw-Hill Medical, 1989, pp. 291–326.

[98] S. E. Antonarakis *et al.*, "Down syndrome," *Nat. Rev. Dis. Prim.*, vol. 6, no. 1, p. 9, Jan. 2020, doi: 10.1038/s41572-019-0143-7.

[99] S. Chaganti *et al.*, "Discovering novel disease comorbidities using electronic medical records," *PLoS One*, vol. 14, no. 11, pp. 1–14, 2019, doi: 10.1371/journal.pone.0225495.

[100] J. B. Pereira *et al.*, "Longitudinal degeneration of the basal forebrain predicts subsequent dementia in Parkinson's disease," *Neurobiol. Dis.*, vol. 139, Jun. 2020, doi: 10.1016/j.nbd.2020.104831.

[101] A. Ortiz, J. Munilla, M. Martínez-Ibañez, J. M. Górriz, J. Ramírez, and D. Salas-Gonzalez, "Parkinson's Disease Detection Using Isosurfaces-Based Features and Convolutional Neural Networks.," *Front. Neuroinform.*, vol. 13, p. 48, May 2019, doi: 10.3389/fninf.2019.00048.

[102] L. C. Jiskoot *et al.*, "Longitudinal multimodal MRI as prognostic and diagnostic biomarker in presymptomatic familial frontotemporal dementia," *Brain*, vol. 142, no. 1, pp. 193–208, Jan. 2019, doi: 10.1093/brain/awy288.

[103] A. J. Lawrence *et al.*, "Longitudinal decline in structural networks predicts dementia in cerebral small vessel disease," *Neurology*, vol. 90, no. 21, pp. e1898–e1910, May 2018, doi: 10.1212/WNL.0000000000005551.

[104] Y. Compta *et al.*, "Combined dementia-risk biomarkers in Parkinson's disease: A prospective longitudinal study," *Park. Relat. Disord.*, vol. 19, no. 8, pp. 717–724, Aug. 2013, doi: 10.1016/j.parkreldis.2013.03.009.

[105] L. L. Chao *et al.*, "ASL perfusion MRI predicts cognitive decline and conversion from MCI to dementia," *Alzheimer Dis. Assoc. Disord.*, vol. 24, no. 1, pp. 19–27, Jan. 2010, doi: 10.1097/WAD.0b013e3181b4f736.

[106] D. G. Bruce *et al.*, "Predictors of cognitive impairment and dementia in older people with diabetes," *Diabetologia*, vol. 51, no. 2, pp. 241–248, Feb. 2008, doi: 10.1007/s00125-007-0894-7.

[107] J. H. Kramer *et al.*, "Longitudinal MRI and Cognitive Change in Healthy Elderly," *Neuropsychology*, vol. 21, no. 4, pp. 412–418, Jul. 2007, doi: 10.1037/0894-4105.21.4.412.

[108] W. Jagust, "Positron emission tomography and magnetic resonance imaging in the diagnosis and prediction of dementia," *Alzheimer's Dement.*, vol. 2, no. 1, pp. 36–42, Jan. 2006, doi: 10.1016/j.jalz.2005.11.002.

[109] L. M. Waite, D. A. Grayson, O. Piguet, H. Creasey, H. P. Bennett, and G. A. Broe, "Gait slowing as a predictor

of incident dementia: 6-Year longitudinal data from the Sydney Older Persons Study," in *Journal of the Neurological Sciences*, Mar. 2005, vol. 229–230, pp. 89–93, doi: 10.1016/j.jns.2004.11.009.
[110] D. Mungas *et al.*, "Longitudinal volumetric MRI change and rate of cognitive decline," *Neurology*, vol. 65, no. 4, pp. 565–571, Aug. 2005, doi: 10.1212/01.wnl.0000172913.88973.0d.
[111] S. M. Resnick *et al.*, "One-year age changes in MRI brain volumes in older adults," *Cereb. Cortex*, vol. 10, no. 5, pp. 464–472, May 2000, doi: 10.1093/cercor/10.5.464.
[112] F. E. K. Al-Khuzaie, O. Bayat, A. D. Duru, and M. Y. Alzahrani, "Diagnosis of Alzheimer Disease Using 2D MRI Slices by Convolutional Neural Network," *Appl. bionics Biomech.*, vol. 2021, 2021, doi: 10.1155/2021/6690539.
[113] H. W. Kim, H. E. Lee, K. Oh, S. Lee, M. Yun, and S. K. Yoo, "Multi-slice representational learning of convolutional neural network for Alzheimer's disease classification using positron emission tomography," *Biomed. Eng. Online*, vol. 19, no. 1, Sep. 2020, doi: 10.1186/s12938-020-00813-z.
[114] Z. Tang *et al.*, "Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline," *Nat. Commun.*, vol. 10, no. 1, pp. 2173–2173, 2019, doi: 10.1038/s41467-019-10212-1.
[115] K. Oh, Y. C. Chung, K. W. Kim, W. S. Kim, and I. S. Oh, "Classification and Visualization of Alzheimer's Disease using Volumetric Convolutional Neural Network and Transfer Learning," *Sci. Rep.*, vol. 9, no. 1, Dec. 2019, doi: 10.1038/s41598-019-54548-6.
[116] M. Maqsood *et al.*, "Transfer Learning Assisted Classification and Detection of Alzheimer's Disease Stages Using 3D MRI Scans," 2019, doi: 10.3390/s19112645.
[117] E. Goceri, "Diagnosis of Alzheimer's disease with Sobolev gradient-based optimization and 3D convolutional neural network," *Int. j. numer. method. biomed. eng.*, vol. 35, no. 7, Jul. 2019, doi: 10.1002/cnm.3225.
[118] J. Rieke, F. Eitel, M. Weygandt, J. D. Haynes, and K. Ritter, "Visualizing convolutional networks for MRI-based diagnosis of alzheimer's disease," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Sep. 2018, vol. 11038 LNCS, pp. 24–31, doi: 10.1007/978-3-030-02628-8_3.
[119] T. R. Stoub *et al.*, "MRI predictors of risk of incident Alzheimer disease: A longitudinal study," *Neurology*, vol. 64, no. 9, pp. 1520–1524, May 2005, doi: 10.1212/01.WNL.0000160089.43264.1A.
[120] C. R. Jack *et al.*, "Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment," *Neurology*, vol. 52, no. 7, pp. 1397–1403, Apr. 1999, doi: 10.1212/wnl.52.7.1397.
[121] J. Bin Bae *et al.*, "Identification of Alzheimer's disease using a convolutional neural network model based on T1-weighted magnetic resonance imaging," *Sci. Rep.*, vol. 10, no. 1, pp. 1–10, Dec. 2020, doi: 10.1038/s41598-020-79243-9.
[122] M. Liu *et al.*, "A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease," *Neuroimage*, vol. 208, p. 116459, Mar. 2020, doi: 10.1016/j.neuroimage.2019.116459.
[123] C. R. Jack Jr *et al.*, "The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI Methods," *J. Magn. Reson. IMAGING*, vol. 27, pp. 685–691, 2008, doi: 10.1002/jmri.21049.
[124] D. S. Marcus, A. F. Fotenos, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open access series of imaging studies: Longitudinal MRI data in nondemented and demented older adults," *J. Cogn. Neurosci.*, vol. 22, no. 12, pp. 2677–2684, Dec. 2010, doi: 10.1162/jocn.2009.21407.
[125] P. J. LaMontagne *et al.*, "OASIS-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease," *medRxiv*. medRxiv, p. 2019.12.13.19014902, Dec. 15, 2019, doi: 10.1101/2019.12.13.19014902.
[126] S. Gauthier *et al.*, "Mild cognitive impairment," *Lancet*, vol. 367, no. 9518, pp. 1262–1270, Apr. 2006, doi: 10.1016/S0140-6736(06)68542-5.
[127] R. C. Petersen and S. Negash, "Mild Cognitive Impairment: An Overview," 2008.
[128] A. J. Asman, A. S. Dagley, and B. A. Landman, "Statistical label fusion with hierarchical performance models.," *Proc. SPIE--the Int. Soc. Opt. Eng.*, vol. 9034, p. 90341E, Mar. 2014, doi: 10.1117/12.2043182.
[129] J. L. R. Andersson and S. N. Sotiropoulos, "An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging," *Neuroimage*, vol. 125, pp. 1063–1078, Jan. 2016, doi: 10.1016/j.neuroimage.2015.10.019.
[130] K. D. Cicerone and K. Kalmar, "Persistent postconcussion syndrome: The structure of subjective complaints after mild traumatic brain injury," *J. Head Trauma Rehabil.*, vol. 10, no. 3, pp. 1–17, Jun. 1995, doi: 10.1097/00001199-199510030-00002.
[131] J.-D. Tournier *et al.*, "MRtrix3: A fast, flexible and open software framework for medical image processing and visualisation," *bioRxiv*, 2019, doi: https://doi.org/10.1101/551739.

[132] R. E. Smith, J. D. Tournier, F. Calamante, and A. Connelly, "Anatomically-constrained tractography: Improved diffusion MRI streamlines tractography through effective use of anatomical information," *Neuroimage*, vol. 62, no. 3, pp. 1924–1938, 2012, doi: 10.1016/j.neuroimage.2012.06.005.
[133] R. E. Smith, J.-D. Tournier, F. Calamante, and A. Connelly, "SIFT: Spherical-deconvolution informed filtering of tractograms," *Neuroimage*, vol. 67, pp. 298–312, Feb. 2013, doi: 10.1016/J.NEUROIMAGE.2012.11.049.
[134] Y. Huo *et al.*, "Consistent cortical reconstruction and multi-atlas brain segmentation," *Neuroimage*, vol. 138, pp. 197–210, Sep. 2016, doi: 10.1016/j.neuroimage.2016.05.030.
[135] I. Lyu, H. Kang, N. D. Woodward, and B. A. Landman, "Sulcal depth-based cortical shape analysis in normal healthy control and schizophrenia groups," p. 1, 2018, doi: 10.1117/12.2293275.
[136] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," 1998. Accessed: Aug. 07, 2019. [Online]. Available: https://link.springer.com/content/pdf/10.1023%2FA%3A1009715923555.pdf.
[137] P. Dall'Acqua *et al.*, "Connectomic and surface-based morphometric correlates of acute mild traumatic brain injury," *Front. Hum. Neurosci.*, vol. 10, no. MAR2016, pp. 1–15, 2016, doi: 10.3389/fnhum.2016.00127.
[138] C. I. Kerley *et al.*, "MRI correlates of chronic symptoms in mild traumatic brain injury," in *Medical Imaging 2020: Image Processing*, Mar. 2020, vol. 11313, p. 97, doi: 10.1117/12.2549493.
[139] Y. Huo *et al.*, "3D whole brain segmentation using spatially localized atlas network tiles.," *Neuroimage*, vol. 194, pp. 105–119, Jul. 2019, doi: 10.1016/j.neuroimage.2019.03.041.
[140] I. Lyu, H. Kang, N. D. Woodward, M. A. Styner, and B. A. Landman, "Hierarchical spherical deformation for cortical surface registration," *Med. Image Anal.*, vol. 57, pp. 72–88, 2019, doi: 10.1016/j.media.2019.06.013.
[141] P. Parvathaneni *et al.*, "Cortical Surface Parcellation Using Spherical Convolutional Neural Networks," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Oct. 2019, vol. 11766 LNCS, pp. 501–509, doi: 10.1007/978-3-030-32248-9_56.
[142] J. J. Koenderink and A. J. Van Doorn, "Surface shape and curvature scales," *Image Vis. Comput.*, vol. 10, no. 8, pp. 557–564, 1992.
[143] S. H. Kim *et al.*, "Development of cortical shape in the human brain from 6 to 24months of age via a novel measure of shape complexity," *Neuroimage*, vol. 135, pp. 163–176, Jul. 2016, doi: 10.1016/j.neuroimage.2016.04.053.
[144] I. Lyu, S. H. Kim, J. B. Girault, J. H. Gilmore, and M. A. Styner, "A cortical shape-adaptive approach to local gyrification index," *Med. Image Anal.*, vol. 48, pp. 244–258, Aug. 2018, doi: 10.1016/j.media.2018.06.009.
[145] I. Lyu, S. H. Kim, N. D. Woodward, M. A. Styner, and B. A. Landman, "TRACE: A Topological Graph Representation for Automatic Sulcal Curve Extraction," *IEEE Trans. Med. Imaging*, vol. 37, no. 7, pp. 1653–1663, Jul. 2018, doi: 10.1109/TMI.2017.2787589.
[146] J. D. Tournier, F. Calamante, and A. Connelly, "Determination of the appropriate b value and number of gradient directions for high-angular-resolution diffusion-weighted imaging," *NMR Biomed.*, vol. 26, no. 12, pp. 1775–1786, Dec. 2013, doi: 10.1002/nbm.3017.
[147] J.-D. Tournier, F. Calamante, and A. Connelly, "Improved probabilistic streamlines tractography by 2nd order integration over fibre orientation distributions," in *Proceedings of the International Society for Magnetic Resonance in Medicine*, 2010, p. 1670.
[148] J.-D. Tournier *et al.*, "MRtrix3: A fast, flexible and open software framework for medical image processing and visualisation," *bioRxiv*, p. 551739, 2019, doi: 10.1101/551739.
[149] Q. V. Le, A. Karpenko, J. Ngiam, and A. Y. Ng, "ICA with reconstruction cost for efficient overcomplete feature learning," *Adv. Neural Inf. Process. Syst. 24 25th Annu. Conf. Neural Inf. Process. Syst. 2011, NIPS 2011*, pp. 1–9, 2011.
[150] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, Mar. 1964, doi: 10.1007/BF02289565.
[151] M. R. Luo, "CIELAB," in *Encyclopedia of Color Science and Technology*, no. C, Springer, 2015, pp. 1–7.
[152] S. B. Eickhoff, K. Amunts, H. Mohlberg, and K. Zilles, "The Human Parietal Operculum. II. Stereotaxic Maps and Correlation with Functional Imaging Results," *Cereb. Cortex*, vol. 16, no. 2, pp. 268–279, Feb. 2006, doi: 10.1093/cercor/bhi106.
[153] B. D. McCandliss, L. Cohen, and S. Dehaene, "The visual word form area: expertise for reading in the fusiform gyrus," *Trends Cogn. Sci.*, vol. 7, no. 7, pp. 293–299, Jul. 2003, doi: 10.1016/S1364-6613(03)00134-7.
[154] I. A. J. van Kooten *et al.*, "Neurons in the fusiform gyrus are fewer and smaller in autism," *Brain*, vol. 131, no. 4, pp. 987–999, Apr. 2008, doi: 10.1093/brain/awn033.
[155] M.-E. Meadows, "Calcarine Cortex," in *Encyclopedia of Clinical Neuropsychology*, New York, NY: Springer

New York, 2011, pp. 472–472.
[156] S. Japee, K. Holiday, M. D. Satyshur, I. Mukai, and L. G. Ungerleider, “A role of right middle frontal gyrus in reorienting of attention: a case study,” *Front. Syst. Neurosci.*, vol. 9, Mar. 2015, doi: 10.3389/fnsys.2015.00023.
[157] S. Cutini *et al.*, “Selective activation of the superior frontal gyrus in task-switching: An event-related fNIRS study,” *Neuroimage*, vol. 42, no. 2, pp. 945–955, Aug. 2008, doi: 10.1016/j.neuroimage.2008.05.013.
[158] M. L. Seghier, “The Angular Gyrus,” *Neurosci.*, vol. 19, no. 1, pp. 43–61, Feb. 2013, doi: 10.1177/1073858412440596.
[159] U. Hasson and P. Tremblay, “Neurobiology of Statistical Information Processing in the Auditory Domain,” in *Neurobiology of Language*, Elsevier, 2016, pp. 527–537.
[160] S. L. MacKenzie, M. C. Wyatt, R. Schuff, J. D. Tenenbaum, and N. Anderson, “Practices and perspectives on building integrated data repositories: Results from a 2010 CTSA survey,” *J. Am. Med. Informatics Assoc.*, vol. 19, no. E1, pp. e119–e124, Jun. 2012, doi: 10.1136/amiajnl-2011-000508.
[161] I. Danciu *et al.*, “Secondary use of clinical data: The Vanderbilt approach,” *J. Biomed. Inform.*, vol. 52, pp. 28–35, Dec. 2014, doi: 10.1016/j.jbi.2014.02.003.
[162] R. S. Evans, J. F. Lloyd, and L. A. Pierce, “Clinical use of an enterprise data warehouse.,” *AMIA Annu. Symp. Proc.*, vol. 2012, pp. 189–198, 2012, Accessed: Jun. 01, 2021. [Online]. Available: /pmc/articles/PMC3540441/.
[163] C. Safran *et al.*, “Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper,” *J. Am. Med. Informatics Assoc.*, vol. 14, no. 1, pp. 1–9, Jan. 2007, doi: 10.1197/jamia.M2273.
[164] N. A. Ahmad, M. L. Kochman, W. B. Long, E. E. Furth, and G. G. Ginsberg, “Efficacy, safety, and clinical outcomes of endoscopic mucosal resection: A study of 101 cases,” *Gastrointest. Endosc.*, vol. 55, no. 3, pp. 390–396, 2002, doi: 10.1067/mge.2002.121881.
[165] J. C. Kirby *et al.*, “PheKB: A catalog and workflow for creating electronic phenotype algorithms for transportability,” *J. Am. Med. Informatics Assoc.*, vol. 23, no. 6, pp. 1046–1052, 2016, doi: 10.1093/jamia/ocv202.
[166] L. A. Hindorff *et al.*, “Potential etiologic and functional implications of genome-wide association loci for human diseases and traits,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 23, pp. 9362–9367, 2009, doi: 10.1073/pnas.0903103106.
[167] J. C. Denny *et al.*, “Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data,” *Nat. Biotechnol.*, vol. 31, no. 12, pp. 1102–1110, 2013, doi: 10.1038/nbt.2749.
[168] W.-Q. Wei *et al.*, “Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record,” *PLoS One*, vol. 12, no. 7, pp. 1–16, Jul. 2017, doi: 10.1371/journal.pone.0175508.
[169] S. J. Hebbring, S. J. Schrodi, Z. Ye, Z. Zhou, D. Page, and M. H. Brilliant, “A PheWAS approach in studying HLA-DRB1*1501,” *Genes Immun.*, vol. 14, no. 3, pp. 187–191, 2013, doi: 10.1038/gene.2013.2.
[170] J. Liu *et al.*, “Phenome-wide association study maps new diseases to the human major histocompatibility complex region,” *J. Med. Genet.*, vol. 53, no. 10, pp. 681–689, 2016, doi: 10.1136/jmedgenet-2016-103867.
[171] X. Li *et al.*, “MR-PheWAS: Exploring the causal effect of SUA level on multiple disease outcomes by using genetic instruments in UK biobank,” *Ann. Rheum. Dis.*, vol. 77, no. 7, pp. 1039–1047, 2018, doi: 10.1136/annrheumdis-2017-212534.
[172] J. C. Denny *et al.*, “Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: Using electronic medical records for genome- and phenome-wide studies,” *Am. J. Hum. Genet.*, vol. 89, no. 4, pp. 529–542, 2011, doi: 10.1016/j.ajhg.2011.09.008.
[173] C. N. Simonti *et al.*, “The phenotypic legacy of admixture between modern humans and Neandertals,” *Science (80-.).*, vol. 351, no. 6274, pp. 737–741, 2016, doi: 10.1126/science.aad2149.
[174] M. G. Ehm *et al.*, “Phenome-wide association study using research participants’ self-reported data provides insight into the Th17 and IL-17 pathway,” *PLoS One*, vol. 12, no. 11, pp. 1–14, 2017, doi: 10.1371/journal.pone.0186405.
[175] S. J. Hebbring, “The challenges, advantages and future of phenome-wide association studies,” *Immunology*, vol. 141, no. 2, pp. 157–165, 2014, doi: 10.1111/imm.12195.
[176] M. R. Boland *et al.*, “Discovering medical conditions associated with periodontitis using linked electronic health records,” *J Clin Periodontol*, vol. 40, no. 5, pp. 1–19, 2014, doi: 10.1111/jcpe.12086.Discovering.
[177] J. L. Warner, G. Alterovitz, K. Bodio, and R. M. Joyce, “External phenome analysis enables a rational

federated query strategy to detect changing rates of treatment-related complications associated with multiple myeloma," *J. Am. Med. Informatics Assoc.*, vol. 20, no. 4, pp. 696–699, 2013, doi: 10.1136/amiajnl-2012-001355.

[178] J. L. Warner and G. Alterovitz, "Phenome based analysis as a means for discovering context dependent clinical reference ranges.," *AMIA Annu. Symp. Proc.*, vol. 2012, pp. 1441–1449, 2012.

[179] E. A. Engels *et al.*, "Comprehensive evaluation of medical conditions associated with risk of non-Hodgkin lymphoma using medicare claims ('MedWAS')," *Cancer Epidemiol. Biomarkers Prev.*, vol. 25, no. 7, pp. 1105–1113, 2016, doi: 10.1158/1055-9965.EPI-16-0212.

[180] R. J. Carroll, L. Bastarache, and J. C. Denny, "R PheWAS: Data analysis and plotting tools for phenome-wide association studies in the R environment," *Bioinformatics*, vol. 30, no. 16, pp. 2375–2376, 2014, doi: 10.1093/bioinformatics/btu197.

[181] "Healthcare Cost and Utilization Project Overview of the National (Nationwide) Inpatient Sample (NIS)," 2021. https://www.hcup-us.ahrq.gov/nisoverview.jsp.

[182] eMERGE Consortium, "Lessons learned from the eMERGE Network: balancing genomics in discovery and practice," *Hum. Genet. Genomics Adv.*, vol. 2, no. 1, p. 100018, Jan. 2021, doi: 10.1016/j.xhgg.2020.100018.

[183] "Utah Population Database," 2021. https://uofuhealth.utah.edu/huntsman/utah-population-database/.

[184] W. A. Rocca, B. P. Yawn, J. L. St. Sauver, B. R. Grossardt, and L. J. Melton, "History of the Rochester epidemiology project: Half a century of medical records linkage in a US population," *Mayo Clin. Proc.*, vol. 87, no. 12, pp. 1202–1213, 2012, doi: 10.1016/j.mayocp.2012.08.012.

[185] L. Bastarache and J. C. Denny, "The Use of ICD-9 Codes in Genetic Association Studies," in *AMIA Annu Symp Proc*, 2011, p. 1738.

[186] J. E. Hopcroft and R. M. Karp, "An n5/2 Algorithm for Maximum Matchings in Bipartite Graphs," *SIAM J. Comput.*, vol. 2, no. 4, pp. 225–231, 1973, doi: 10.1137/0202019.

[187] P. Wu *et al.*, "Mapping ICD-10 and ICD-10-CM codes to phecodes: Workflow development and initial evaluation," *J. Med. Internet Res.*, vol. 21, no. 11, pp. 1–13, 2019, doi: 10.2196/14325.

[188] S. Seabold and J. Perktold, "Statsmodels: Econometric and Statistical Modeling with Python," in *PROC. OF THE 9th PYTHON IN SCIENCE CONF*, 2010, no. January 2010, pp. 92–96, [Online]. Available: http://statsmodels.sourceforge.net/.

[189] J. D. Hunter, "Matplotlib : a 2D Graphics Environment," *Comput. Sci. Eng.*, vol. 9, pp. 90–95, 2007.

[190] "HCUP CCS-Services and Procedures," *Healthcare Cost and Utilization Project (HCUP)*. Agency for Healthcare Research and Quality, Rockville, MD, May 2018, Accessed: Jul. 11, 2020. [Online]. Available: https://www.hcup-us.ahrq.gov/toolssoftware/ccs_svcsproc/ccssvcproc.jsp.

[191] S. A. Pendergrass *et al.*, "The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery," *Genet. Epidemiol.*, vol. 35, no. 5, pp. 410–422, 2011, doi: 10.1002/gepi.20589.

[192] M. J. Bull *et al.*, "Clinical report - Health supervision for children with Down syndrome," *Pediatrics*, vol. 128, no. 2, pp. 393–406, 2011, doi: 10.1542/peds.2011-1605.

[193] M. A. Davidson, "Primary Care for Children and Adolescents with Down Syndrome," *Pediatr. Clin. North Am.*, vol. 55, no. 5, pp. 1099–1111, 2008, doi: 10.1016/j.pcl.2008.07.001.

[194] G. D. Smith and S. Ebrahim, "Data dredging, bias, or confounding," *Br. Med. J.*, vol. 325, no. 7378, pp. 1437–1438, Dec. 2002, doi: 10.1136/bmj.325.7378.1437.

[195] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011, Accessed: Feb. 11, 2021. [Online]. Available: http://scikit-learn.sourceforge.net.

[196] G. S. Birkhead, M. Klompas, and N. R. Shah, "Uses of Electronic Health Records for Public Health Surveillance to Advance Public Health," *Annu. Rev. Public Health*, vol. 36, no. 1, pp. 345–359, Mar. 2015, doi: 10.1146/annurev-publhealth-031914-122747.

[197] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis," *IEEE J. Biomed. Heal. Informatics*, vol. 22, no. 5, pp. 1589–1604, Sep. 2018, doi: 10.1109/JBHI.2017.2767063.

[198] W.-W. Yim, A. J. Wheeler, C. Curtin, T. H. Wagner, and T. Hernandez-Boussard, "Secondary use of electronic medical records for clinical research: challenges and opportunities," *Converg. Sci. Phys. Oncol.*, vol. 4, no. 1, p. 014001, Feb. 2018, doi: 10.1088/2057-1739/aaa905.

[199] C. I. Kerley *et al.*, "pyPheWAS: A Phenome-Disease Association Tool for Electronic Medical Record Analysis," *Neuroinformatics*, vol. 1, pp. 1–23, Jan. 2022, doi: 10.1007/s12021-021-09553-4.

[200] L. A. C. Millard, N. M. Davies, T. R. Gaunt, G. D. Smith, and K. Tilling, "Software application profile: PHESANT: A tool for performing automated phenome scans in UK Biobank," *Int. J. Epidemiol.*, vol. 47, no.

1, pp. 29–35, Feb. 2018, doi: 10.1093/ije/dyx204.
[201] D. Dingen *et al.*, "RegressionExplorer: Interactive Exploration of Logistic Regression Models with Subgroup Analysis," *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 1, pp. 246–255, 2019, doi: 10.1109/TVCG.2018.2865043.
[202] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*. .
[203] D. E. Farrar and R. R. Glauber, "Multicollinearity in Regression Analysis: The Problem Revisited," 1967. Accessed: Feb. 01, 2021. [Online]. Available: https://about.jstor.org/terms.
[204] D. G. Altman, "Multiple comparisons," in *Practical Statistics For Medical Research*, Chapman and Hall, 1990, pp. 210–212.
[205] S. Rouam, "False Discovery Rate (FDR)," *Encycl. Syst. Biol.*, pp. 731–732, 2013, doi: 10.1007/978-1-4419-9863-7_223.
[206] C. Brokamp, A. F. Beck, N. K. Goyal, P. Ryan, J. M. Greenberg, and E. S. Hall, "Material community deprivation and hospital utilization during the first year of life: an urban populationebased cohort study," *Ann. Epidemiol.*, vol. 30, pp. 37–43, 2019, doi: 10.1016/j.annepidem.2018.11.008.
[207] P. S. Jensen, D. Martin, and D. P. Cantwell, "Comorbidity in ADHD: Implications for research, practice, and DSM-V," *J. Am. Acad. Child Adolesc. Psychiatry*, vol. 36, no. 8, pp. 1065–1079, Aug. 1997, doi: 10.1097/00004583-199708000-00014.
[208] P. S. Jensen *et al.*, "ADHD Comorbidity Findings From the MTA Study: Comparing Comorbid Subgroups," 2001. doi: 10.1097/00004583-200102000-00009.
[209] T. J. Spencer, "ADHD and Comorbidity in Childhood," 2006.
[210] N. J. Roizen and D. Patterson, "Down's syndrome," in *Lancet*, Apr. 2003, vol. 361, no. 9365, pp. 1281–1289, doi: 10.1016/S0140-6736(03)12987-X.
[211] M. Alexander, H. Petri, Y. Ding, C. Wandel, O. Khwaja, and N. Foskett, "Morbidity and medication in a large population of individuals with Down syndrome compared to the general population," *Dev. Med. Child Neurol.*, vol. 58, no. 3, pp. 246–254, Mar. 2016, doi: 10.1111/dmcn.12868.
[212] V. Prasher, S. Ninan, and S. Haque, "Fifteen-year follow-up of thyroid status in adults with Down syndrome," *J. Intellect. Disabil. Res.*, vol. 55, no. 4, pp. 392–396, Apr. 2011, doi: 10.1111/j.1365-2788.2011.01384.x.
[213] G. T. Capone *et al.*, "Co-occurring medical conditions in adults with Down syndrome: A systematic review toward the development of health care guidelines," *Am. J. Med. Genet. Part A*, vol. 176, no. 1, pp. 116–133, Jan. 2018, doi: 10.1002/ajmg.a.38512.
[214] G. Laws and D. Gunn, "Phonological memory as a predictor of language comprehension in Down syndrome: A five-year follow-up study," *J. Child Psychol. Psychiatry Allied Discip.*, vol. 45, no. 2, pp. 326–337, Feb. 2004, doi: 10.1111/j.1469-7610.2004.00224.x.
[215] G. E. Martin, J. Klusek, B. Estigarribia, and J. E. Roberts, "Language characteristics of individuals with down syndrome," *Top. Lang. Disord.*, vol. 29, no. 2, pp. 112–132, Apr. 2009, doi: 10.1097/TLD.0b013e3181a71fe1.
[216] P. Landete *et al.*, "Obstructive sleep apnea in adults with Down syndrome," *Am. J. Med. Genet. Part A*, vol. 182, no. 12, pp. 2832–2840, Dec. 2020, doi: 10.1002/ajmg.a.61853.
[217] D. Valentini *et al.*, "Medical conditions of children and young people with Down syndrome," *J. Intellect. Disabil. Res.*, vol. 65, no. 2, pp. 199–209, Feb. 2021, doi: 10.1111/jir.12804.
[218] W. S. Bush, M. T. Oetjens, and D. C. Crawford, "Unravelling the human genome-phenome relationship using phenome-wide association studies," *Nature Reviews Genetics*, vol. 17, no. 3. Nature Publishing Group, pp. 129–145, Mar. 01, 2016, doi: 10.1038/nrg.2015.36.
[219] J. C. Denny, L. Bastarache, and D. M. Roden, "Phenome-Wide Association Studies as a Tool to Advance Precision Medicine," *Annual Review of Genomics and Human Genetics*, vol. 17. Annual Reviews Inc., pp. 353–373, 2016, doi: 10.1146/annurev-genom-090314-024956.
[220] A. Baban *et al.*, "Differences in morbidity and mortality in Down syndrome are related to the type of congenital heart defect," *Am. J. Med. Genet. Part A*, vol. 182, no. 6, pp. 1342–1350, Jun. 2020, doi: 10.1002/ajmg.a.61586.
[221] M. D'Alto and V. S. Mahadevan, "Pulmonary arterial hypertension associated with congenital heart disease," *Eur. Respir. Rev.*, vol. 21, no. 126, pp. 328–337, Dec. 2012, doi: 10.1183/09059180.00004712.
[222] J. C. Fudge *et al.*, "Congenital Heart Surgery Outcomes in Down Syndrome: Analysis of a National Clinical Database," *Pediatrics*, vol. 126, no. 2, pp. 315–322, Aug. 2010, doi: 10.1542/peds.2009-3245.
[223] R. Lange, T. Guenther, R. Busch, J. Hess, and C. Schreiber, "The presence of Down syndrome is not a risk factor in complete atrioventricular septal defect repair," *J. Thorac. Cardiovasc. Surg.*, vol. 134, no. 2, pp. 304–310, Aug. 2007, doi: 10.1016/j.jtcvs.2007.01.026.
[224] "pyPheWAS 3.1.1 documentation." https://pyphewas.readthedocs.io/en/latest/ (accessed Feb. 01, 2021).

[225] J. D. Blume, L. D'Agostino McGowan, W. D. Dupont, and R. A. Greevy, "Second-generation p-values: Improved rigor, reproducibility, & transparency in statistical analyses," *PLoS One*, vol. 13, no. 3, p. e0188299, Mar. 2018, doi: 10.1371/journal.pone.0188299.
[226] P. Khairy, R. Ionescu-Ittu, A. S. Mackie, M. Abrahamowicz, L. Pilote, and A. J. Marelli, "Changing Mortality in Congenital Heart Disease," *J. Am. Coll. Cardiol.*, vol. 56, no. 14, pp. 1149–1157, Sep. 2010, doi: 10.1016/j.jacc.2010.03.085.
[227] F. Healy, B. D. Hanna, and R. Zinman, "Pulmonary Complications of Congenital Heart Disease," *Paediatr. Respir. Rev.*, vol. 13, no. 1, pp. 10–15, Mar. 2012, doi: 10.1016/j.prrv.2011.01.007.
[228] C. D. Kemp and J. V. Conte, "The pathophysiology of heart failure," *Cardiovasc. Pathol.*, vol. 21, no. 5, pp. 365–371, Sep. 2012, doi: 10.1016/j.carpath.2011.11.007.
[229] R. W. Light *et al.*, "Prevalence and Clinical Course of Pleural Effusions at 30 Days after Coronary Artery and Cardiac Surgery," *Am. J. Respir. Crit. Care Med.*, vol. 166, no. 12, pp. 1567–1571, Dec. 2002, doi: 10.1164/rccm.200203-184OC.
[230] D. Bush, C. Galambos, D. D. Ivy, S. H. Abman, K. Wolter-Warmerdam, and F. Hickey, "Clinical Characteristics and Risk Factors for Developing Pulmonary Hypertension in Children with Down Syndrome," *J. Pediatr.*, vol. 202, pp. 212-219.e2, Nov. 2018, doi: 10.1016/j.jpeds.2018.06.031.
[231] G. M. Loughlin, J. W. Wynne, and B. E. Victorica, "Sleep apnea as a possible cause of pulmonary hypertension in Down syndrome," *J. Pediatr.*, vol. 98, no. 3, pp. 435–437, Mar. 1981, doi: 10.1016/S0022-3476(81)80716-0.
[232] E. Head *et al.*, "Cerebrovascular pathology in Down syndrome and Alzheimer disease," *Acta Neuropathol. Commun.*, vol. 5, no. 1, p. 93, Dec. 2017, doi: 10.1186/s40478-017-0499-4.
[233] J. C. Vis *et al.*, "Down syndrome: a cardiovascular perspective," *J. Intellect. Disabil. Res.*, vol. 53, no. 5, pp. 419–425, May 2009, doi: 10.1111/j.1365-2788.2009.01158.x.
[234] H. Hasle, I. Haunstrup Clemmensen, and M. Mikkelsen, "Risks of leukaemia and solid tumours in individuals with Down's syndrome," *Lancet*, vol. 355, no. 9199, pp. 165–169, Jan. 2000, doi: 10.1016/S0140-6736(99)05264-2.
[235] E. M. Tucker, L. A. Pyles, J. L. Bass, and J. H. Moller, "Permanent Pacemaker for Atrioventricular Conduction Block After Operative Repair of Perimembranous Ventricular Septal Defect," *J. Am. Coll. Cardiol.*, vol. 50, no. 12, pp. 1196–1200, Sep. 2007, doi: 10.1016/j.jacc.2007.06.014.
[236] M. A. Banks, J. Jenson, and J. D. Kugler, "Late development of atrioventricular block after congenital heart surgery in down syndrome," *Am. J. Cardiol.*, vol. 88, no. 1, pp. 86–89, Jul. 2001, doi: 10.1016/S0002-9149(01)01596-X.
[237] P. Versacci, D. Di Carlo, M. C. Digilio, and B. Marino, "Cardiovascular disease in Down syndrome," *Curr. Opin. Pediatr.*, vol. 30, no. 5, pp. 616–622, Oct. 2018, doi: 10.1097/MOP.0000000000000661.
[238] F. C. Golding, "THE ASSOCIATION OF ATROPHIC GASTRITIS WITH HYPOTHYROIDISM; A PRELIMINARY REPORT OF 11 CASES," *Ann. Intern. Med.*, vol. 17, no. 5, p. 828, Nov. 1942, doi: 10.7326/0003-4819-17-5-828.
[239] S. M. Pueschel, "Clinical aspects of down syndrome from infancy to adulthood," *Am. J. Med. Genet.*, vol. 37, no. S7, pp. 52–56, Jun. 2005, doi: 10.1002/ajmg.1320370708.
[240] C. Gong *et al.*, "A Meta-Analysis of C-Reactive Protein in Patients With Alzheimer's Disease," *Am. J. Alzheimer's Dis. Other Dementiasr*, vol. 31, no. 3, pp. 194–200, May 2016, doi: 10.1177/1533317515602087.
[241] M. E. Petersen *et al.*, "Proteomic profiles for Alzheimer's disease and mild cognitive impairment among adults with Down syndrome spanning serum and plasma: An Alzheimer's Biomarker Consortium–Down Syndrome (ABC–DS) study," *Alzheimer's Dement. Diagnosis, Assess. Dis. Monit.*, vol. 12, no. 1, Jan. 2020, doi: 10.1002/dad2.12039.
[242] K. Hirono *et al.*, "Left Ventricular Noncompaction and Congenital Heart Disease Increases the Risk of Congestive Heart Failure," *J. Clin. Med.*, vol. 9, no. 3, p. 785, Mar. 2020, doi: 10.3390/jcm9030785.
[243] A. Wieczorek, J. Hernandez-Robles, L. Ewing, J. Leshko, S. Luther, and J. Huhta, "Prediction of outcome of fetal congenital heart disease using a cardiovascular profile score," *Ultrasound Obstet. Gynecol.*, vol. 31, no. 3, pp. 284–288, Mar. 2008, doi: 10.1002/uog.5177.
[244] C. Stoll, B. Dott, Y. Alembik, and M. P. Roth, "Associated congenital anomalies among cases with Down syndrome," *Eur. J. Med. Genet.*, vol. 58, no. 12, pp. 674–680, 2015, doi: 10.1016/j.ejmg.2015.11.003.
[245] M. R. Tumanyan, O. V. Filaretova, V. V. Chechneva, R. S. Gulasaryan, I. V. Butrim, and L. A. Bockeria, "Repair of Complete Atrioventricular Septal Defect in Infants with Down Syndrome: Outcomes and Long-Term Results," *Pediatr. Cardiol.*, vol. 36, no. 1, pp. 71–75, Jan. 2015, doi: 10.1007/s00246-014-0966-7.
[246] R. J. McGrath, M. L. Stransky, W. C. Cooley, and J. B. Moeschler, "National Profile of Children with Down

Syndrome: Disease Burden, Access to Care, and Family Impact," *J. Pediatr.*, vol. 159, no. 4, pp. 535-540.e2, Oct. 2011, doi: 10.1016/j.jpeds.2011.04.019.
[247] J. M. Bachmann *et al.*, "Association of Neighborhood Socioeconomic Context With Participation in Cardiac Rehabilitation," *J. Am. Heart Assoc.*, vol. 6, no. 10, Oct. 2017, doi: 10.1161/JAHA.117.006260.
[248] R. W. Grant *et al.*, "Use of Latent Class Analysis and k-Means Clustering to Identify Complex Patient Profiles," *JAMA Netw. Open*, vol. 3, no. 12, p. e2029068, Dec. 2020, doi: 10.1001/jamanetworkopen.2020.29068.
[249] L. C. Messer *et al.*, "The Development of a Standardized Neighborhood Deprivation Index," *J. Urban Heal.*, vol. 83, no. 6, pp. 1041–1062, Dec. 2006, doi: 10.1007/s11524-006-9094-x.
[250] D. W. Frost, S. Vembu, J. Wang, K. Tu, Q. Morris, and H. B. Abrams, "Using the Electronic Medical Record to Identify Patients at High Risk for Frequent Emergency Department Visits and High System Costs," *Am. J. Med.*, vol. 130, no. 5, pp. 601.e17-601.e22, May 2017, doi: 10.1016/j.amjmed.2016.12.008.
[251] R. B. Nes *et al.*, "Adaptation to the birth of a child with a congenital anomaly: A prospective longitudinal study of maternal well-being and psychological distress.," *Dev. Psychol.*, vol. 50, no. 6, pp. 1827–1839, 2014, doi: 10.1037/a0035996.
[252] G. H. S. Singer, "Meta-analysis of comparative studies of depression in mothers of children with and without developmental disabilities," *Am. J. Ment. Retard.*, vol. 111, no. 3, May 2006, doi: 10.1352/0895-8017(2006)111[155:MOCSOD]2.0.CO;2.
[253] K. M. Hart and N. Neil, "Down syndrome caregivers' support needs: a mixed-method participatory approach," *J. Intellect. Disabil. Res.*, vol. 65, no. 1, pp. 60–76, Jan. 2021, doi: 10.1111/jir.12791.
[254] G. George *et al.*, "PheGWAS: a new dimension to visualize GWAS across multiple phenotypes," *Bioinformatics*, vol. 36, no. 8, pp. 2500–2505, Apr. 2020, doi: 10.1093/bioinformatics/btz944.
[255] Y. Wang, H. Yao, and S. Zhao, "Auto-encoder based dimensionality reduction," *Neurocomputing*, vol. 184, pp. 232–242, 2016, doi: 10.1016/j.neucom.2015.08.104.
[256] C. Song, F. Liu, Y. Huang, L. Wang, and T. Tan, "Auto-encoder based data clustering," in *Iberoamerican congress on pattern recognition*, 2013, pp. 117–124.
[257] M. Sakurada and T. Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction," *ACM Int. Conf. Proceeding Ser.*, vol. 02-Decembe, pp. 4–11, 2014, doi: 10.1145/2689746.2689747.
[258] L. Gondara, "Medical Image Denoising Using Convolutional Denoising Autoencoders," *IEEE Int. Conf. Data Min. Work. ICDMW*, vol. 0, pp. 241–246, 2016, doi: 10.1109/ICDMW.2016.0041.
[259] R. Wei and A. Mahmood, "Recent Advances in Variational Autoencoders with Representation Learning for Biomedical Informatics: A Survey," *IEEE Access*, vol. 9, pp. 4939–4956, 2021, doi: 10.1109/ACCESS.2020.3048309.
[260] M. Chen, X. Shi, Y. Zhang, D. Wu, and M. Guizani, "Deep Feature Learning for Medical Image Analysis with Convolutional Autoencoder Neural Network," *IEEE Trans. Big Data*, vol. 7, no. 4, pp. 750–758, 2017, doi: 10.1109/tbdata.2017.2717439.
[261] S. L. Smith, P.-J. Kindermans, C. Ying, and Q. V Le, "Don't Decay the Learning Rate, Increase the Batch Size.," in *International Conference on Learning Representations - ICLR*, 2018.
[262] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7700 LECTU, pp. 437–478, 2012, doi: 10.1007/978-3-642-35289-8_26.
[263] D. R. Wilson and T. R. Martinez, "The general inefficiency of batch training for gradient descent learning," *Neural Networks*, vol. 16, no. 10, pp. 1429–1451, 2003, doi: 10.1016/S0893-6080(03)00138-2.
[264] D. Masters and C. Luschi, "Revisiting Small Batch Training for Deep Neural Networks," pp. 1–18, 2018.
[265] N. S. Keskar, J. Nocedal, P. T. P. Tang, D. Mudigere, and M. Smelyanskiy, "On large-batch training for deep learning: Generalization gap and sharp minima," *5th Int. Conf. Learn. Represent. ICLR 2017 - Conf. Track Proc.*, pp. 1–16, 2017.
[266] N. W. Shock, *Normal human aging: the Baltimore longitudinal study of aging*. Baltimore, Md: U.S. Dept. of Health and Human Services, Public Health Service, National Institutes of Health, National Institute on Aging, Gerontology Research Center, 1984.
[267] S. Bakas *et al.*, "Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge," 2018.
[268] B. H. Menze *et al.*, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Trans. Med. Imaging*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015, doi: 10.1109/TMI.2014.2377694.
[269] G. Grabner, A. L. Janke, M. M. Budge, D. Smith, J. Pruessner, and D. L. Collins, "Symmetric atlasing and model based segmentation: an application to the hippocampus in older adults," in *International Conference*

on Medical Image Computing and Computer-Assisted Intervention, 2006, pp. 58–66.
[270] L. Van Der Maaten and G. Hinton, "Visualizing Data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.
[271] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241.
[272] F. He, T. Liu, and D. Tao, "Control batch size and learning rate to generalize well: Theoretical and empirical evidence," *Adv. Neural Inf. Process. Syst.*, vol. 32, no. NeurIPS, 2019.
[273] I. Kandel and M. Castelli, "The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset," *ICT Express*, vol. 6, no. 4, pp. 312–315, 2020, doi: 10.1016/j.icte.2020.04.010.
[274] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *2nd Int. Conf. Learn. Represent. ICLR 2014 - Conf. Track Proc.*, no. Ml, pp. 1–14, 2014.
[275] M. Pesteie, P. Abolmaesumi, and R. N. Rohling, "Adaptive Augmentation of Medical Data Using Independently Conditional Variational Auto-Encoders," *IEEE Trans. Med. Imaging*, vol. 38, no. 12, pp. 2807–2820, Dec. 2019, doi: 10.1109/TMI.2019.2914656.
[276] M. Chamberland *et al.*, "Tractometry-based Anomaly Detection for Single-subject White Matter Analysis," pp. 1–5, 2020, [Online]. Available: http://arxiv.org/abs/2005.11082.
[277] T. A. Lasko, J. C. Denny, and M. A. Levy, "Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data," *PLoS One*, vol. 8, no. 6, p. e66341, Jun. 2013, doi: 10.1371/JOURNAL.PONE.0066341.
[278] I. Landi *et al.*, "Deep representation learning of electronic health records to unlock patient stratification at scale," *npj Digit. Med. 2020 31*, vol. 3, no. 1, pp. 1–11, Jul. 2020, doi: 10.1038/s41746-020-0301-z.
[279] C. I. Kerley, L. Y. Cai, Y. Tang, L. L. Beason-held, and S. M. Resnick, "Batch size : go big or go home ? Counterintuitive improvement in medical autoencoders with smaller batch size," in *SPIE Medical Imaging: Image Processing [under review]*, 2023.
[280] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.
[281] Q. Dong, S. Gong, and X. Zhu, "Class Rectification Hard Mining for Imbalanced Deep Learning," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1869–1878, Accessed: Sep. 24, 2022. [Online]. Available: doi: 10.1109/ICCV.2017.205.
[282] D. Li *et al.*, "Self-Guided Hard Negative Generation for Unsupervised Person Re-Identification," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 2022.
[283] J. Souyoung *et al.*, "Unsupervised Hard Example Mining from Videos for Improved Object Detection," in *Proceedings of the European Conference on Computer Vision*, 2018.
[284] H. Lin, H. Chen, Q. Dou, L. Wang, J. Qin, and P. A. Heng, "ScanNet: A fast and dense scanning framework for metastastic breast cancer detection from whole-slide image," *Proc. - 2018 IEEE Winter Conf. Appl. Comput. Vision, WACV 2018*, vol. 2018-January, pp. 539–546, May 2018, doi: 10.1109/WACV.2018.00065.
[285] D. Wang, A. Khosla, R. Gargeya, H. Irshad, A. H. Beck, and B. Israel, "Deep Learning for Identifying Metastatic Breast Cancer," *arXiv Prepr.* , 2016, Accessed: Sep. 25, 2022. [Online]. Available: http://camelyon16.grand-challenge.org/.
[286] Y. B. Tang, K. Yan, Y. X. Tang, J. Liu, J. Xiao, and R. M. Summers, "Uldor: A universal lesion detector for ct scans with pseudo masks and hard negative example mining," *Proc. - Int. Symp. Biomed. Imaging*, vol. 2019-April, pp. 833–836, Apr. 2019, doi: 10.1109/ISBI.2019.8759478.
[287] P. Kumar and M. Mayank Srivastava, "Example Mining for Incremental Learning in Medical Imaging," in *2018 IEEE symposium series on computational intelligence*, 2018.
[288] S. Zhang, J. Sun, Y. Huang, X. Ding, and Y. Zheng, "Medical Symptom Detection in Intelligent Pre-Consultation using Bi-directional Hard-Negative Noise Contrastive Estimation," *Proc. 28th ACM SIGKDD Conf. Knowl. Discov. Data Min.*, vol. 1, 2022, doi: 10.1145/3534678.3539124.
[289] A. Paszke *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems 32*, 2019, pp. 8024–8035, Accessed: Sep. 24, 2022. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.
[290] Y. LeCun, C. Cortes, and C. Burges, "MNIST handwritten digit database." http://yann.lecun.com/exdb/mnist/ (accessed Sep. 24, 2022).
[291] L. Ferrucci, "The Baltimore Longitudinal Study of Aging (BLSA): A 50-Year-Long Journey and Plans for the Future," *J. Gerontol. A. Biol. Sci. Med. Sci.*, vol. 63, no. 12, p. 1416, 2008, doi:

10.1093/GERONA/63.12.1416.
[292] A. C. Evans, D. L. Collins, S. R. Millst, E. D. Brown, R. L. Kelly, and T. M. Peters, “3D statistical neuroanatomical models from 305 MRI volumes,” in *Nuclear Science Symposium and Medical Imaging Conference*, 1993, pp. 1813–1817.
[293] Autism and Developmental Disabilities Monitoring Network, “Community Report on Autism ,” 2021. Accessed: Sep. 20, 2022. [Online]. Available: www.cdc.gov/autism.
[294] C. Lord, M. Elsabbagh, G. Baird, and J. Veenstra-Vanderweele, “Autism spectrum disorder,” *Lancet* , vol. 392, no. 10146, p. 508, Aug. 2018, doi: 10.1016/S0140-6736(18)31129-2.
[295] J. W. Smoller, “The use of electronic health records for psychiatric phenotyping and genomics,” *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.*, vol. 177, no. 7, pp. 601–612, Oct. 2018, doi: 10.1002/AJMG.B.32548.
[296] R. Al-jawahiri and E. Milne, “Resources available for autism research in the big data era: A systematic review,” *PeerJ*, vol. 2017, no. 1, p. e2880, Jan. 2017, doi: 10.7717/PEERJ.2880/SUPP-3.
[297] S. Lyalina, B. Percha, P. Lependu, S. V. Iyer, R. B. Altman, and N. H. Shah, “Identifying phenotypic signatures of neuropsychiatric disorders from electronic medical records,” *J. Am. Med. Informatics Assoc.*, vol. 20, no. E2, pp. e297–e305, Dec. 2013, doi: 10.1136/AMIAJNL-2013-001933/17376103/20-E2-E297.PDF.
[298] T. Wolfers *et al.*, “From pattern classification to stratification: towards conceptualizing the heterogeneity of Autism Spectrum Disorder,” *Neurosci. Biobehav. Rev.*, vol. 104, pp. 240–254, Sep. 2019, doi: 10.1016/J.NEUBIOREV.2019.07.010.
[299] I. S. Kohane *et al.*, “The Co-Morbidity Burden of Children and Young Adults with Autism Spectrum Disorders,” *PLoS One*, vol. 7, no. 4, p. e33224, 2012, doi: 10.1371/JOURNAL.PONE.0033224.
[300] D. Cawthorpe, “A 16-Year Cohort Analysis of Autism Spectrum Disorder-Associated Morbidity in a Pediatric Population,” *Front. Psychiatry*, vol. 9, p. 635, Nov. 2018, doi: 10.3389/FPSYT.2018.00635/BIBTEX.
[301] F. Doshi-Velez, Y. Ge, and I. Kohane, “Comorbidity clusters in Autism Spectrum Disorders: An Electronic Health Record Time-Series Analysis,” *Pediatrics*, vol. 133, no. 1, pp. e54–e63, Jan. 2014, doi: 10.1542/peds.2013-0819.
[302] S. Georgiades *et al.*, “Investigating phenotypic heterogeneity in children with autism spectrum disorder: a factor mixture modeling approach,” *J. Child Psychol. Psychiatry*, vol. 54, no. 2, pp. 206–215, Feb. 2013, doi: 10.1111/J.1469-7610.2012.02588.X.
[303] F. Rafiee, R. Rezvani Habibabadi, M. Motaghi, D. M. Yousem, and I. J. Yousem, “Brain MRI in Autism Spectrum Disorder: Narrative Review and Recent Advances,” *J. Magn. Reson. Imaging*, vol. 55, no. 6, pp. 1613–1624, Jun. 2022, doi: 10.1002/JMRI.27949.
[304] S. Brieber *et al.*, “Structural brain abnormalities in adolescents with autism spectrum disorder and patients with attention deficit/hyperactivity disorder,” *J. Child Psychol. Psychiatry*, vol. 48, no. 12, pp. 1251–1258, Dec. 2007, doi: 10.1111/J.1469-7610.2007.01799.X.
[305] D. Van Rooij *et al.*, “Cortical and subcortical brain morphometry differences between patients with autism spectrum disorder and healthy individuals across the lifespan: Results from the ENIGMA ASD working group,” *Am. J. Psychiatry*, vol. 175, no. 4, pp. 359–369, Apr. 2018, doi: 10.1176/APPI.AJP.2017.17010100/ASSET/IMAGES/LARGE/APPI.AJP.2017.17010100F3.JPEG.
[306] K. Nickel *et al.*, “Inferior Frontal Gyrus Volume Loss Distinguishes Between Autism and (Comorbid) Attention-Deficit/Hyperactivity Disorder—A FreeSurfer Analysis in Children,” *Front. Psychiatry*, vol. 9, p. 521, Oct. 2018, doi: 10.3389/FPSYT.2018.00521/XML/NLM.
[307] L. O’Dwyer *et al.*, “Brain Volumetric Correlates of Autism Spectrum Disorder Symptoms in Attention Deficit/Hyperactivity Disorder,” *PLoS One*, vol. 9, no. 6, p. e101130, Jun. 2014, doi: 10.1371/JOURNAL.PONE.0101130.
[308] L. D. Yankowitz, B. E. Yerys, J. D. Herrington, J. Pandey, and R. T. Schultz, “Dissociating regional gray matter density and gray matter volume in autism spectrum condition,” *NeuroImage Clin.*, vol. 32, p. 102888, Jan. 2021, doi: 10.1016/J.NICL.2021.102888.
[309] J. Cai, X. Hu, K. Guo, P. Yang, M. Situ, and Y. Huang, “Increased Left Inferior Temporal Gyrus Was Found in Both Low Function Autism and High Function Autism,” *Front. Psychiatry*, vol. 9, p. 542, Oct. 2018, doi: 10.3389/FPSYT.2018.00542/BIBTEX.
[310] S. D. Lukito *et al.*, “Neural Correlates of Duration Discrimination in Young Adults with Autism Spectrum Disorder, Attention-Deficit/Hyperactivity Disorder and Their Comorbid Presentation,” *Front. Psychiatry*, vol. 9, p. 569, Nov. 2018, doi: 10.3389/FPSYT.2018.00569/XML/NLM.
[311] O. Dekhil *et al.*, “A Personalized Autism Diagnosis CAD System Using a Fusion of Structural MRI and Resting-State Functional MRI Data,” *Front. Psychiatry*, vol. 10, p. 392, Jul. 2019, doi:

10.3389/FPSYT.2019.00392/XML/NLM.
[312] A. Klein and J. Tourville, “101 labeled brain images and a consistent human cortical labeling protocol,” *Front. Neurosci.*, vol. 0, no. DEC, p. 171, 2012, doi: 10.3389/FNINS.2012.00171/ABSTRACT.
[313] V. Bitsika, C. F. Sharpley, and S. Orapeleng, “An exploratory analysis of the use of cognitive, adaptive and behavioural indices for cluster analysis of ASD subgroups,” *J. Intellect. Disabil. Res.*, vol. 52, no. 11, pp. 973–985, Nov. 2008, doi: 10.1111/J.1365-2788.2008.01123.X.
[314] S. X. Yu and J. Shi, “Multiclass spectral clustering,” *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 1, pp. 313–319, 2003, doi: 10.1109/ICCV.2003.1238361.
[315] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” Feb. 2018, doi: 10.48550/arxiv.1802.03426.
[316] A. Errante and L. Fogassi, “Activation of cerebellum and basal ganglia during the observation and execution of manipulative actions,” *Sci. Rep.*, vol. 10, no. 1, pp. 1–15, Jul. 2020, doi: 10.1038/s41598-020-68928-w.
[317] A. M. Graybiel, “The basal ganglia,” *Curr. Biol.*, vol. 10, no. 14, pp. R509–R511, 2000.
[318] C. L. Runge and D. R. Friedland, “127 - Neuroanatomy of the Auditory System,” in *Cummings Otolaryngology*, 2021, pp. 1938–1944.
[319] G. Pasco *et al.*, “Comparison of neural substrates of temporal discounting between youth with autism spectrum disorder and with obsessive-compulsive disorder,” *Psychol. Med.*, vol. 47, pp. 2513–2527, 2017, doi: 10.1017/S0033291717001088.
[320] S. Haar, S. Berman, M. Behrmann, and I. Dinstein, “Anatomical Abnormalities in Autism?,” *Cereb. Cortex*, vol. 26, no. 4, pp. 1440–1452, Apr. 2016, doi: 10.1093/CERCOR/BHU242.
[321] L. Squarcina *et al.*, “Automatic classification of autism spectrum disorder in children using cortical thickness and support vector machine,” *Brain Behav.*, vol. 11, pp. 1–9, 2021, doi: 10.1002/brb3.2238.
[322] Y. C. Lo, Y. J. Chen, Y. C. Hsu, W. Y. I. Tseng, and S. S. F. Gau, “Reduced tract integrity of the model for social communication is a neural substrate of social communication deficits in autism spectrum disorder,” *J. Child Psychol. Psychiatry*, vol. 58, no. 5, pp. 576–585, May 2017, doi: 10.1111/JCPP.12641.
[323] M. A. D’Albis *et al.*, “Local structural connectivity is associated with social cognition in autism spectrum disorder,” *Brain*, vol. 141, no. 12, pp. 3472–3481, Dec. 2018, doi: 10.1093/BRAIN/AWY275.