# Contextualizing Medical Image Analyses with Electronic Health Histories

By

Shikha Chaganti

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Computer Science

February 28, 2019

Nashville, Tennessee

Approved:

Professor Bennett A. Landman

Professor Benoit Dawant

Professor Richard Alan Peters

Professor Thomas Lasko

Dr. Louise A. Mawn

## TABLE OF CONTENTS

Page

## ACKNOWLEDGEMENTS

I am grateful to my advisor Dr. Bennett Landman, one of the most insightful and intelligent people I have had the pleasure of knowing. Throughout the last four and a half years, you have been a great source of inspiration and support in professional and personal matters. You helped me recognize ideas of mine that were the most impactful, and often elevated them with valuable advice. I am grateful to Dr. Miga and VISE for providing me with a wonderful opportunity with the VISE training grant. The clinical interactions and the collaborative environment at VISE are truly experiences to cherish for life. Dr. Louise Mawn was pivotal in helping me understand the clinical perspective on AI in medicine. Much of my work has been inspired by watching your brilliant mind in action in the clinic and the operating room.

I would like to thank the Dr. Thomas Lasko, for always encouraging me and providing detailed and constructive feedback. I would like to thank Dr. Benoit Dawant and Dr. Alan Peters for taking the time to advise me and being a part of my thesis committee. It was a such pleasure to work with my brilliant friends and colleagues at MASI lab: Dr. Yuankai Huo, Dr. Ilwoo Lyu, Prasanna Parvathaneni, Dr. Andrew Plassard, Dr. Rob Harrigan, Katrina Nelson, Steve Damon, Dr. Shunxing Bao, Dr. Allison Hainline, Vishwesh Nath, Cam Bermudez, Justin Blaber, Roza Bayrak, Karthik Ramadass, Colin Hansen, Hyensoo Hu, Cailey Cline, Sam Remedios, Kunal Nabar, Sandra González-Villà, Gaile Lejay, Riqiang Gao, Yucheng Tang, and Anupam Kumar. Thanks for all the fun times. I would like to thank all of the wonderful members of WoV, who provided a sense of community and a nurturing environment for women like me in VISE. Special thanks to Michelle Bukowski for making it all happen.

To my parents, Ravi and Tulasi, thank you for your unconditional love and support every step of the way. Thanks for always believing in me and my ability. You never let me think that anything was out of my reach. None of this would have been possible without you. Megu, thanks for constantly being by my side through all the craziness. You get me like no one else. I'd like to thank my *ammamma*, Rajeshwari, for teaching me how to be strong and independent, and value myself. Thanks to my best friend and partner, Adam Wolsky, who never fails to make me laugh – in good times and bad. Thank you for making every day fun and for always pushing me to do better. Finally, I'd like to thank all of my friends who have been like a family to me in Nashville: Zohra, Daniela, Alexis, Ahmet, Lauren, Srijata, Emily, and Cecilia. Nashville would not have been the same without you.

## LIST OF TABLES

**Table** **Page**

## LIST OF FIGURES

**Figure** **Page**

# Chapter I. Introduction

## 1. Big Data in Medicine

In 2009, the Health Information Technology for Economic and Clinical Health (HI-TECH) Act, allocated $15.5 billion to incentivize the adoption of Electronic Health Records (EHR) in the USA. A recent survey showed that this program has been largely successful with 72% of office-based physicians using EHR systems in 2012 [1]. In 2011, it was estimated that 150 exabytes (1 billion gigabytes) of data was stored by the US health care system, and will soon reach a zettabyte ($10^{24}$ bytes) [2]. The National Institutes of Health started a program called "Big Data to Knowledge" [3] to "facilitate broad use of biomedical big data, develop and disseminate analysis methods and software, enhance training relevant for large-scale data analysis, and establish centers of excellence for biomedical big data." These programs provided a great opportunity for computer scientists to use large-scale machine learning algorithms on clinical data [4].

The abundance of medical data presents new challenges for consolidation of information for decision making in the medical community in general, and the medical imaging community specifically. Physicians are confronted with the problem of 'information overload', affecting their ability to process important information and discard irrelevant data. Wilson, in his paper about technology in health care, defines information overload as "a perception on the part of the individual that flow of information associated with work tasks is greater than can be managed effectively"[5]. The problem of information overload also leads to errors or missed test results. Ash et al conducted a study of electronic medical record systems and showed that typical use of these systems fosters errors rather than reducing their likelihood, due to cognitive overload[6]. This can happen in the form of fragmented information – switching between screens instead of

having a complete overview. Reports have the problem of being "over-complete" while decreasing in readability and information value. These systems send too many alerts, reminders and warning messages, leading to a loss of crucial information in several instances [7].

Radiologists face problems of fatigue and errors due to information overload in imaging systems [8]. The increase in the volume of imaging did not see an equal improvement in systems that automate information extraction and improve work flow for the radiologists. There is a need for technology that improves workflow in radiology departments [9], [10]. In a review article, Lee et al report that the retrospective error rate among radiologic examinations in 30%, and real-time error rate is 3-5% [11]. 75% of all malpractice claims are related to misdiagnosis against radiologists. Among the solutions proposed, Lee et al propose that improvement in computer-aided technology that can automate certain processes can vastly reduce the errors caused by radiologists and help with the visual fatigue faced by radiologists. Reiner et al suggest similar strategies to improve radiologist performance including, preliminary interpretation using CAD (computer-aided diagnosis) systems, and automated extraction of imaging/report data [12]. In a recent article, Recht and Bryan discuss the importance of artificial intelligence (AI) to radiologists[13]. They envision a radiology practice of the future, where the role of the AI system is that of an aggregator of information, and the role of a radiologist as that of an expert interpreter. The AI system will perform the mundane tasks of assembling relevant information from a patient's electronic medical record, identify anomalous regions in the patient's scans, and prepare a report of important findings. The radiologist will perform value-added functions of making sense of the integrated data, making diagnoses, and interacting with patients and clinical teams to improve overall clinical care. They predict that AI systems will contribute to reducing stress and fatigue for the radiologists, and improving job satisfaction.

## 1.1. Machine Learning in Radiology

AI and machine learning have been used extensively in the field of medical image analysis for various tasks such as segmentation of structures, image registration, computer-aided detection and diagnosis[14], [15]. Some application such as detection and segmentation of regions of interest would boost the performance of radiologists by performing mundane tasks and saving time, reducing fatigue, and improving productivity. Examples of such applications include localization of organs [16], lung nodule detection [17]–[19], and prostate lesion detection [20]. Some recent applications also provide automated annotations and text descriptions for new scans [21], [22]. A system such as this would be of great help to a radiologist by providing a template for a text report beforehand, so that less time can be spent on data entry.

Recent computer-aided diagnosis systems show classification accuracy comparable with radiologists in limited test cases[23]–[25]. Kooi et al developed a CAD system to detect lesions on a breast mammography. They used features learned from convolutional neural networks along with hand crafted features [26] and compared the performance of their system to that of human performance in a patch level study. They found that there was no statistically significant difference in the performance. Giger et al developed a system to identify malignant breast lesions from diagnostic MRI (Magnetic Resonance Imaging). They demonstrated that the performance of a radiologist improved when their AI system was used as an aid from AUC (Area Under the Curve) of 0.7055 to 0.7575 and the mean sensitivity increased from 90.4% to 94.2%. This device recently received FDA clearance[27].

However, most of these systems perform image analysis task in isolation, i.e., they do not consider the context of the disease, the patient's history, earlier test results and procedures and so on into consideration. This is one of the major drawbacks of current practice of medical image

analysis. In an article discussing the future of radiology, Blum and Zins discuss that while machine learning methods might become extremely good at detection, radiologists are still required to perform other important functions such as, integrating different sources of clinical data [28]. Similarly, Recht et al point out that while the current CAD systems are gaining proficiency in radiologic tasks, integrating clinical and imaging information is necessary for contextual reading of medical imaging [13].

### 1.2. Context-Aware Image Analysis of the Eye Orbit

In this work, we will develop machine learning methods to automatically analyze electronic medical data related to diseases of the optic nerve including imaging, past diagnoses, procedures, labs and other demographic information. Traditional medical image analysis studies focus on discovering single or multiple radiologic phenotypes associated with a disease. However, in the case of complex disease presentations, radiologic phenotypes are inadequate or incomplete in extracting relevant information about the disease as similar imaging features might be present in different diseases even though the pathological causes might be distinct. In the case of diseases of the optic nerve, visual disturbances are caused in a patient due to damage to the optic nerve. The causes for the damage range from systemic abnormalities such as endocrine problems, aging, neoplasms, or trauma. An overview of different conditions involving the optic nerve is presented in section 2 of this chapter. Our goal in this work is to learn imaging and EMR biomarkers that are associated with visual function and optic nerve disease and use these biomarkers to model prognosis and diagnosis of the diseases. The main contribution of this work is to show that context-aware models that are developed by integrating a patient's health history into image analysis improve the outcomes of the statistical models. A full list of contributions of this work is presented in Section 3 of this chapter.

Briefly, the steps involved in achieving this goal are detailed below:

1. **Image Segmentation**: We will develop models that will automatically segment the structures in the eye orbit, from clinically acquired computed tomography (CT) and magnetic resonance imaging (MRI).
2. **Discovering Biomarkers of Interest**: We will extract relevant information from automatically segmented structures and evaluate their correlation with visual function such as with visual acuity, and visual field defects.
3. **Develop Tools for Extraction of EMR features**: We will develop tools to extract EMR context signatures from past records of the patient, including past diagnoses, procedures, labs and so on.
4. **Develop Diagnostic Classifiers**: We will develop machine learning methods to classify diseases of the optic nerve from control subjects. The ultimate goal of this study will be to develop a differential diagnosis classifier that is able to separate diseases of the optic nerve based on all of the information collected on them including imaging, and health history.

## 2. Diseases of the Optic Nerve

Diseases of the optic nerve affect millions of Americans each year. These diseases include a wide range of conditions such as glaucoma, thyroid eye disease, optic neuritis, papilledema, idiopathic intracranial hypertension, and orbital infections. Diagnosis and treatment planning of orbital conditions is done based on visual testing, careful review of the patient's history and orbital imaging. In this work, we develop automated image analysis and machine learning algorithms that will aid in the process of diagnosis and prognosis prediction of the diseases of the optic nerve. In

this section, we present an overview of the diseases and diagnostic challenges associated with these conditions.

The complete list of conditions considered in this work is shown in Table I-1. These codes have been divided into five separate cohorts of related conditions as shown in the table. All of these diagnostic codes could lead to optic neuropathy, a condition where the optic nerve is damaged due to various types of pathology or trauma[29].

A patient with optic neuropathy typically exhibits symptoms of sudden mild to severe vision loss in terms of both visual acuity and visual field defects. In most optic neuropathies, the origin of the damage to the optic nerve can be ascertained from other systemic problems or trauma such as an accident. For example, patients with past thyroid imbalances or tumors have compressive optic neuropathy; patients with giant cell arteritis have non-arteric ischemic optic neuropathy, patients with high intraocular pressure have glaucomatous optic neuropathy and so on. Diseases of the optic nerve can have serious consequences such as permanent blindness when not detected and treated within the window of opportunity. Differential diagnosis of orbital conditions is often a challenging process where data from multiple sources need to be gathered to make a decision[30].

Figure I-1 shows the co-occurrence of these conditions in a group of 28,411 ophthalmology patients, whose records were collected from the Vanderbilt University Medical Center under IRB approval. ICD-9 codes within a cohort have a very high co-occurrence, however there is overlap between cohorts as well. Each node in Figure I-1 represents an ICD-9 code and is color-coded by its respective cohort. Each edge represents a co-occurrence measure between two nodes called the *Dominant Confidence (DC)* measure. The thickness of the edge represents the magnitude of *DC*, i.e. a thicker edge indicates a stronger co-occurrence. The *DC* measure between *A* and *B* is given

by $DC\ (A,B) = max(P(A|B), P(B|A))$, where *P(A|B)* is the conditional probability of having diagnosis *A* given diagnosis B, and *P(B|A)* is vice versa. The graph was created by calculating the *DC* measure for all 6,241 pairs of ICD9 codes.

**Table I-1. Cohort inclusion criteria**

| Cohort | ICD-9 codes | Description |
|---|---|---|
| Glaucoma | 365.0* | Borderline glaucoma |
| | 365.1* | Open-angle glaucoma |
| | 365.2* | Primary angle-closure glaucoma |
| | 365.3* | Corticosteroid-induced glaucoma |
| | 365.4* | Glaucoma associated with congenital anomalies, dystrophies, and systemic syndromes |
| | 365.5* | Glaucoma associated with disorders of the lens |
| | 365.6* | Glaucoma associated with other ocular disorders |
| | 365.7* | Glaucoma stage, unspecified |
| | 365.8* | Other specified forms of glaucoma |
| | 365.9* | Unspecified glaucoma |
| Intrinsic Optic Nerve Disease | 377.3* | Optic Neuritis |
| | 377.4* | Other disorders of optic nerve |
| Optic Nerve Edema | 348.2 | Idiopathic intracranial hypertension |
| | 377.0, 377.00 | Papilledema |
| | 377.01 | Papilledema, increased intracranial pressure |
| | 377.02 | Papilledema, decreased ocular pressure |
| Orbital Inflammation | 376.0, 376.00 | Acute inflammation of orbit |
| | 376.01 | Orbital cellulitis |
| | 376.02 | Orbital periostitis |
| | 376.1 | Chronic inflammation of orbit |
| | 376.11 | Orbital granuloma |
| | 376.12 | Orbital myositis |
| | 373.13 | Abscess of eyelid |

| | | |
|---|---|---|
| Thyroid Eye Disease | 242.00 | Toxic diffuse goiter without thyrotoxic crisis or storm |
| | 376.2 | Endocrine exophthalmos |
| | 376.21 | Thyrotoxic exophthalmos |
| | 376.22 | Exophthalmic ophthalmoplegia |

## 2.1. Disease Background

### *2.1.1. Glaucoma*

Glaucoma is a group of conditions that are characterized by increased intra-ocular pressure and visual field defects. Glaucoma primarily affects the optic nerve head due to various pathological causes, resulting in glaucomatous optic neuropathy [31]. All types of glaucoma can potentially lead to blindness, making it the second leading cause of blindness in the world. Globally, 8.4 million of the 60 million people affected with glaucoma are blind [32]. The ocular hypertension treatment study showed that glaucomatous damage can be delayed or prevented altogether by controlling ocular pressure at early stages[33], [34]. A study conducted by Kolker showed that medically or surgically maintaining the intraocular pressure of patients affected by glaucoma to under 18 mm Hg prevented progressive decline in visual field loss[35]. Therefore, it is essential to detect glaucomatous optic neuropathy as early as possible to prevent permanent blindness.

*2.1.2. Intrinsic Optic Nerve Disease*

The conditions included in the category of intrinsic optic nerve disease are optic neuritis and other disorders of the optic nerve. These conditions present with sudden loss in visual function due to damage to the optic nerve. Optic neuritis (ON) is a demyelinating inflammation of the optic nerve that can cause serious visual loss [36]. ON usually has good prognosis for vision outcome, since most patients fully recover their vision [37]. However, it is predictive of future multiple

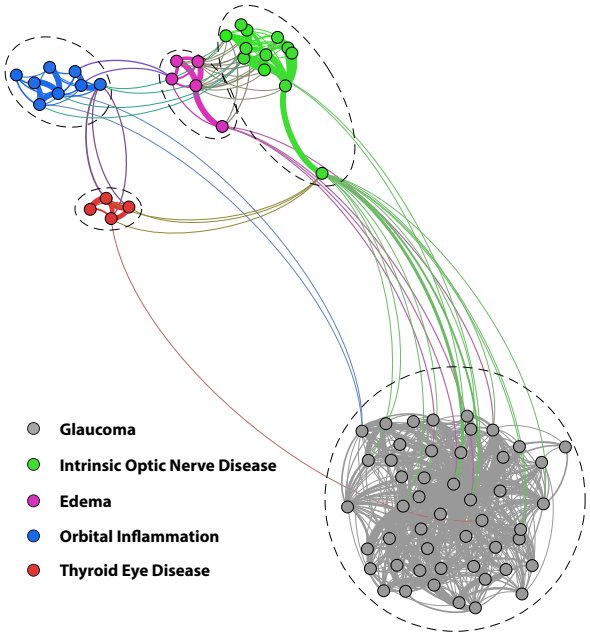



**Figure I-1. Disease co-occurrence graph. Each node represents an ICD9 code, color-coded by the disease cohort. The thickness of the edges between two nodes is proportional to the probability of co-occurrence of their ICD9 codes in the study population**

sclerosis [38], [39] and therefore early detection of ON is extremely important. Especially, since corticosteroids have been shown to improve vision and also prevent future MS lesions for at least two years for ON patients [40]. Some of the other neuropathies included in this category are: ischemic optic neuropathy which causes damage to the optic nerve due to lack of blood supply, and can result in permanent loss of vision[41]; optic nerve hypoplasia which is a non-progressive

congenital anomaly that can result in severe vision loss[42] and optic nerve sheath hemorrhage which is sometimes seen in trauma cases[43].

### *2.1.3. Optic Nerve Edema*

This category includes Idiopathic Intracranial Hypertension (IIH) and Papilledema. IIH is the presence of increased intracranial pressure (ICP) without a tumor or other lesions in the brain. IIH has devastating effects on vision, with blindness resulting in 10% of affected patients. Ney and all conducted a study to evaluate the levels of loss of visual function in IIH patients, and found that up to 83% of the patients have some loss of visual function in their visual fields [44]. The incidence of IIH is higher in women than in men, with various estimates of female to male incidence ratios ranging from 4:1 to 15:1. The average age of the people affected by IIH is 35 years, and about 95% of affected people are obese [45]. IIH is also observed in children, however in this age group, there seems to be no association with gender or obesity [46]. IIH can be present with or without papilledema; when papilledema is present this condition is also called pseudotumor cerebri syndrome (PTCS). Patients with PTCS are affected by severe visual disturbances, whereas those without papilledema are not. Since severe and permanent vision loss is imminent in many of the patients, swift detection and immediate intervention is necessary.

### *2.1.4. Thyroid Eye Disease*

Thyroid eye disease or Grave's ophthalmopathy is associated with an autoimmune condition called Grave's disease that affects the regulation of the thyroid hormone. It affects the soft tissues in the eye orbit, and can have devastating consequences without early treatment and management[47], [48]. The inflamed tissues in the orbit can lead to compressive damage to the optic nerve, which results in visual defects including visual field loss, color loss, and decreased

visual acuity. Vision loss due to thyroid disease is called dysthyroid optic neuropathy, which can lead to permanent blindness. Intervention options for thyroid eye disease include medication by immuno-suppressants as well as orbital decompression surgery. The treatment is more effective when the disease is detected in acute stages and controlled[49].

*2.1.5. Orbital Inflammation*

Orbital inflammation includes orbital cellulitis, orbital periostitis which are usually caused by infections. Acute orbital infections can have good prognosis when treated early, but any delay in intervention can cause irreversible blindness and, in some cases, death. JA Jr et al show that precipitous loss of vision can occur in patients with orbital aspergillosis without intervention in five case studies[50]. Yohai et al showed that in orbital mucormycosis the morbidity rate can be as high as 41% without early intervention[51]. This category also includes idiopathic inflammation of the orbit. The pathogenesis of idiopathic orbital inflammation is unknown and can affect multiple organs in the eye orbit including lacrimal glands, extraocular muscles and orbital fat [52]. Optic neuropathy is observed in some patients with orbital inflammation, which needs to be identified and treated with steroids to avoid permanent loss of vision.

2.2. Clinical Observation Experience: Case Studies

Detection and treatment of diseases that affect the optic nerve is a difficult process. The optic nerve is a very delicate structure and often, the window for intervention is very small. In my clinical observation experience, I observed ten different types of neuro-ophthalmological surgical procedures. In two of the cases that I observed, the patients had permanent vision loss as the window for intervention was missed. The problem with both these cases was not due to the

challenging nature of the diagnostic problem, but due to the fact that the physician missed a small detail because of information overload present in current day EMR systems.

### *2.2.1. Case 1*

A female patient was diagnosed with IIH (Idiopathic Intracranial Hypertension) about 10 years ago. Surgical intervention was not made at the time of diagnosis as the patient was on blood thinner medication from a previous DVT (deep vein thrombosis). The patient had severe vision loss due to IIH with almost complete vision loss in the left eye. There was remainder visual function in the temporal field. A new physician discovered that the patient's DVT was an acute problem from a bypass surgery and therefore the patient need not have been on Coumadin (blood thinner) for 10 years. They immediately stopped the Coumadin medication and scheduled the patient for an Optic Nerve Sheath Fenestration procedure, to reduce the intracranial pressure and preserve the remaining vision. However, this intervention would not recover the lost visual function.

The problem in this case was a communication gap between physicians in different departments regarding the patient's medical history. With electronic medical records being digitized and maintained under one umbrella system, issues such as these can be prevented by having an automated system that can analyze co-morbid conditions and identify any contraindications or the lack thereof for therapeutic intervention. This is an example of *information overload,* all the necessary information to make the diagnosis and intervention was in the patient's records, but the lack of tools that can manage this data and highlight relevant information caused a 10-year delay in intervention.

*2.2.2. Case 2*

A patient had a fungal infection that invaded the orbital space of the left eye. An intra-orbital catheter was placed for delivery of anti-fungal medicine directly to the affected region in order to treat the infection that could have potentially spread to the intracranial space. This patient went to the ER 6 months prior to the surgery complaining of a headache and blurred vision. When questioned by the ER staff about any trauma that might have occurred in the recent past, the patient mentioned a fall from 2 weeks prior to the ER visit. The ER staff immediately suspected that the fall could explain the symptoms. They performed a CT scan for indications of TBI (traumatic brain injury). The scan was clear and the patient was sent home. Five days later, the patient went blind in the affected eye. Another scan showed that there was visible infection in the orbital apex. The ER staff likely missed the infection as it is not common to check the orbital apex for infections in an ER visit. They also probably do not have the time or bandwidth to study the patient's history to see if there were any related conditions, such as a sinus infection in the patient's recent past. In this case, like the previous one, one of the problems is integration of electronic medical records to perform a quick analysis of co-morbid conditions. Another problem is that of identifying anomalies in imaging in a high-pressure, high-case-volume environment such as the ER.

This is a perfect example where there is a need for a program that automatically identifies and detects biomarkers for orbital diseases so that radiologist on duty can review them quickly to identify high risk cases. Such a system can have a huge impact on productivity and prevention of adverse outcomes.

## 2.3. The Role of Orbital Imaging in Management of Diseases of the Optic Nerve

Modern 3D imaging methods, such as computed tomography (CT) and magnetic resonance imaging (MRI), are used to establish the location of the pathology in orbital conditions. CT imaging can be used to compute measurements of the globe, extraocular muscles, and the optic nerve [53]. MRI is sensitive to optic nerve structure, inflammation, myelination, and axonal loss[54]–[56]. Imaging is used to compute volumetric changes observed in the optic nerve sheath as well changes in optic nerve diameter to diagnose IIH [56]–[59]. Optic nerve head deformity and optic nerve sheath distension can be observed on T2-weighted MRI is used in diagnosis of papilledema[60], [61].

Volumetric changes in extraocular muscles, and changes in the optic nerve are used to diagnose thyroid eye disease [62]. In addition to volumetric changes, other structural measurements such orbital apex angle, Barrett's crowding index, and ratio of soft tissue to fat are used in diagnosis of thyroid eye disease [63]–[65]. MRI or CT imaging is used in diagnosis of orbital inflammation[66], to establish anatomic location and pathology. Some of the features observed are orbital fat infiltration, bone destruction, abscess, and diffuse inflammation of the extraocular muscles. MRI and CT imaging are also used in diagnosis of orbital infections[67], and pseudotumors[68], [69]. Hickman et al showed that the optic nerve area from an MRI can be used to evaluate the progression of optic neuritis [70]. They observed swelling and inflammation consistent with visual impairment. Features of optic neuritis can also be observed using clinical computed tomography (CT) images, when MRI is not available[71]. Harrigan et al quantitatively showed that radii of the optic nerve and the surrounding cerebrospinal fluid measured from MRI can be used to distinguish optic neuritis patients [72]. In fact several imaging studies of the eye

orbit have demonstrated the value of using orbital imaging to study diseases with complicated and challenging differential diagnosis[73]–[76].

In conclusion, orbital imaging captures several structural changes that are observed in diseases of the optic nerve. These changes serve as imaging biomarkers for detection, diagnosis and management of these diseases. In this work, we develop automated methods that will segment the structures in the eye orbit in clinical CT imaging and extract relevant structural features from these segmentations. This process of automated extraction can eventually be integrated into clinical practice to automate the detection of anomalous regions for review by a radiologist or a physician. Figure I-2 shows an example output of the structures segmented in a CT image using a multi-atlas segmentation pipeline. The features extracted from the image segmentation pipeline will be used for outcome modelling and diagnostic classification using machine learning.

### 2.4. Importance of Contextual Information

Since many orbital conditions are systemic in nature, detection, diagnosis and therapeutic planning of these diseases require careful study of patient's history and likelihood of disease as well. Other data collected for differential diagnosis includes lab tests and pathology tests. In orbital conditions, searching for system wide co-morbidities is particularly important. For example, IIH is associated with several systemic conditions such as Vitamin A deficiency, polycystic ovary syndrome, diabetes, thyroid disease, anemia, stroke, migraine, systemic lupus erythematous, pregnancy, menstrual dysfunction, and side-effects to certain medications [25],[56]. Thyroid eye disease is associated with past hypothyroidism, hyperthyroidism and Grave's disease[47], [78], [79]. Optic nerve hypoplasia is associated with abnormalities in the central nervous system and endocrine system[42]. Glaucoma is associated with hypertension,

hyperlipidemia, systemic lupus, diabetes, hyperthyroidism, depression, and psychosis[80]. A multitude of systemic diseases can also cause other orbital inflammatory diseases[66], [81].

In this work, we will develop a tool that will summarize the relevant history for each condition in terms of past co-morbid conditions, and procedures to form disease EMR signatures. Image analysis studies in this work will be integrated with these EMR signatures to provide context for the biomarkers observed. We find that such context-aware studies improve outcome modelling when compared to traditional image analysis studies.

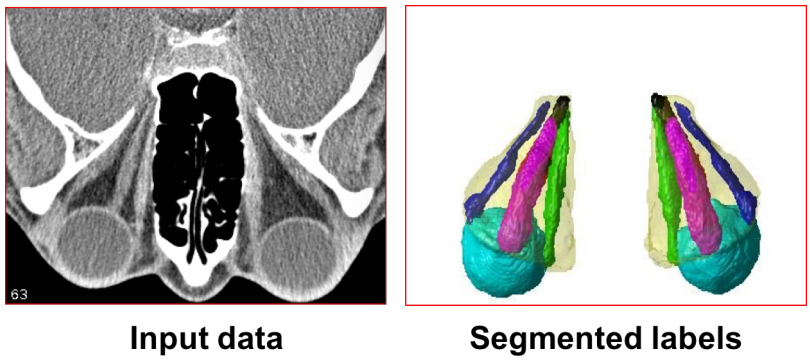



**Figure I-2. Result of segmentation. The image on the left side shows an input CT image, and the image on the right shows the segmented structures using multi-atlas segmentation. The structures segmented here are: the eye globe, the optic nerve, the superior, inferior, lateral, and medial extraocular muscles, and the orbital fat.**

## 3. Contributions

In this work, we developed several image analysis and data analytic methods to analyze diagnostic information collected in the ophthalmology department at Vanderbilt University Medical Center. The data collected includes CT and MR imaging, demographic information, past diagnoses, procedure codes and labs. In this section, we will summarize the current contributions of this work. The first step in this process is to segment the relevant structures in the eye orbit, including the optic nerve, the globe, the extraocular muscles and orbital fat, and extract structural

measurements. The image analysis methods developed in this work are explained in section 3.1. Next, we showed that the measurements extracted from orbital imaging are correlated with visual function, and the phenotypes extracted from our statistical models match with clinical findings of disease subtypes. These clinical associations are explained further in section 3.2. Finally, we show that addition of EMR data in the form of past diagnoses and/or procedures improves outcome modelling and diagnostic classification, compared to models that use imaging alone. In section 3.3, we briefly explain the methods developed to extract EMR data and the experiments designed to show that context aware image analysis models perform better.

### 3.1. Segmentation of Orbital Imaging

Multi-atlas methodology is suitable for identifying orbital structures such as the eye globe, the optic nerve, extraocular muscles and orbital fat from clinically acquired CT imaging. In this paradigm, a set of example atlases with expertly labelled structures are used to learn the segmentation in a new image. The procedure involves non-rigid registration of the atlases into the target image space, followed by label fusion of the registered atlases to learn the final label at each voxel in the target image.

However, challenges arise in the identification of the individual extraocular rectus muscles that control eye movement. This is increasingly problematic in diseased eyes, where these muscles often appear to fuse at the back of the orbit (at the resolution of clinical computed tomography imaging) due to inflammation or crowding. We show that Kalman filters can be used to track the muscles in three-dimensions to refine multi-atlas segmentation and resolve ambiguity due to imaging resolution, noise, and artifacts. We further demonstrate the utility of the approach by correlating structural metrics of the eye orbit with clinical data and visual function measures in subjects with thyroid eye disease. The pilot study demonstrates that automatically calculated

orbital metrics are strongly correlated with several clinical characteristics. Moreover, it is shown that the superior, inferior, medial and lateral rectus muscles obtained using Kalman filters are each correlated with different categories of functional deficit. These findings serve as foundation for further investigation in the use of CT imaging in the study, analysis and diagnosis of ocular diseases, specifically thyroid eye disease. The methodology and results of this work are presented in Chapter III.

One of the main steps in the multi-atlas segmentation process is non-rigid registration of example atlases to the target space. Image registration involves identification of a transformation to fit a target image to a reference image space. The success of the registration process is vital for correct interpretation of the results of many medical image-processing applications, including multi-atlas segmentation. While there are several validation metrics employed in rigid registration to examine the accuracy of the method, non-rigid registrations (NRR) are validated subjectively in most cases, validated in offline cases, or validated based on image similarity metrics, all of which have been shown to poorly correlate with true registration quality. To counter this problem, we model the error for each target scan by expanding on the idea of Assessing Quality Using Image Registration Circuits (AQUIRC)[82], which created a model for error "quality" associated with NRR. In this work, we model the Dice similarity coefficient (DSC) error in the network, for a more interpretable measure. We test four functional models using a leave-one-out strategy to evaluate the relationship between edge DSC and circuit DSC: linear, quadratic, third order, or multiplicative models. We find that the quadratic model most accurately learns the NRR-DSC, with a median correlation coefficient of 0.58 with the true NRR-DSC, we call this the QUADRATIC (QUAlity of Dice in RegistrATIon Circuits) model. The QUADRATIC model is used for multi-atlas segmentation of the orbital structures in clinically acquired CT imaging based on majority vote.

Choosing the four best atlases predicted from the QUDRATIC model resulted in a 7% increase in the DSC between segmented image and true labels. The procedure for calculating registration error and results of the error models and segmentation are presented in chapter IV.

## 3.2. Clinical Applications of Image Analysis

### *3.2.1. Thyroid Eye Disease*

The purpose of this study is to understand the phenotypes of thyroid eye disease (TED) through data derived from a multi-atlas segmentation of computed tomography (CT) imaging in patients with thyroid eye disease. Images of 170 orbits of 85 retrospectively selected TED patients were analyzed with the automated segmentation tool developed as described in the section 3.1. Twenty-five bilateral orbital structural metrics were used to perform principal component analysis (PCA). The top two principal components which accounted for 60% of the variance, aligned with the description of TED subtypes in clinical literature. The first phenotype identified was the big volume phenotype with enlarged or inflamed muscles. The second phenotype was the stretched optic nerve phenotype, which indicated proptosis of the eye. Most of the subjects in the study have either of these characteristics or a combination of both. A Kendall rank correlation between the principal components (phenotypes) and clinical data showed that, the big volume phenotype was very strongly correlated ($p$-value$<0.05$) with motility defects, and loss of visual acuity. Whereas, the stretched optic nerve phenotype was strongly correlated ($p$-value$<0.05$) with a high Hertel, relatively better visual acuity, and smoking. The methodology and detailed results of this work are presented in Chapter V.

## 3.3. Contextualizing Medical Image Analysis with Electronic Health Histories

### *3.3.1. PyPheWAS Python Package*

Natarajan et al conducted an analysis of the types of search queries in electronic health record systems, and showed that a majority of queries are informational in nature that look for past diagnoses, labs, and procedures. Billing codes for past diagnoses, procedures and labs can be easily collected from electronic medical records to be integrated into image analysis pipelines.

We developed a python package called pyPheWAS to study the associations of clinical phenotypes with a particular disease of interest[83], [84]. In this package, related ICD-9(International Classification of Disease version 9) codes or diagnostic billing codes are mapped to PheWAS codes[85] (Phenome Wide Association Study codes) or phecodes as defined by Denny et al. Each phecode gives a clinical phenotype e.g. "Visual disturbances", includes symptoms such as blurred vision, diplopia, color vision loss and so on. Next, we compare these clinical phenotypes between two populations: the disease of interest and a control population. The comparison can be performed at different granularities of time and age. For example, queries can be used to compare only patient populations within an age range or compare records *prior* to the diagnosis of a disease only. We extended this functionality to procedure codes known as CPT (current procedural terminology) codes. We established custom grouping of CPT into ProWAS codes (Procedural Wide Association Study). In the future, we will extend the functionality to lab test results as well.

Studies conducted based on phecodes will henceforth be referred to as phenome-disease association studies (PheDAS). With PheDAS, we can isolate the phenotypes most associated with a disease group. In other words, we develop an "EMR context signature" that consolidates a patient's history of diagnoses and calculates its relative probabilities. Figure I-3 shows an example

of a PheDAS graph computed on preliminary data of 1000 patients with thyroid eye disease (TED) and other patients in the ophthalmology clinic to see the differences in presentation. In this graph, records that were present 1 year before the onset of disease were used to examine predictive factors. We see that TED is associated with several endocrine problems such as thyroid imbalances and abnormalities. It is also mildly correlated with heart conditions and menopause. These conditions are therefore relevant to the diagnosis of a new patient who might have TED. Context signatures such as these can easily be integrated into imaging studies to learn context-dependent models.

*3.3.2. Association of Visual Function with EMR-Radiological Phenotypes*

Multi-modal analyses of diseases of the optic nerve, that combine radiological imaging with other electronic medical records (EMR), improve understanding of visual function. We conducted a study of 55 patients with glaucoma and 32 patients with thyroid eye disease (TED). We collected their visual assessments, orbital CT imaging, and EMR data. We developed an image-processing pipeline that segmented and extracted structural metrics from CT images. We derived EMR phenotype vectors with the help of PheWAS (from diagnostic codes) and ProWAS (from treatment codes). Next, we performed a principal component analysis and multiple-correspondence analysis to identify their association with visual function scores. We found that structural metrics derived from CT imaging are significantly associated with functional visual score for both glaucoma ($R^2$=0.32) and TED ($R^2$=0.4). Addition of EMR phenotype vectors to the model significantly improved ($p<1E-04$) the $R^2$ to 0.4 for glaucoma and 0.54 for TED. The methodology and results of this study are presented in Chapter VI.

### *3.3.3. EMR Context Signatures Improve Diagnostic Classification*

EMR signatures that capture system-wide co-morbidities for a disease population within a given time interval are computed using the pyPheWAS package. We investigate the effect of integrating these EMR signatures with radiological data to improve diagnostic classification in disease domains known to have confounding factors because of variable and complex clinical presentation. Specifically, we focus on two studies: (1) a study of four major optic nerve related conditions and (2) a study of diabetes. Addition of EMR signature vectors to radiologically-derived structural metrics improves the area under the curve (AUC) for diagnostic classification using elastic net regression, for diseases of the optic nerve. For glaucoma, the AUC improves from 0.71 to 0.83, for intrinsic optic nerve disease it increases from 0.72 to 0.91, for optic nerve edema it increases from 0.95 to 0.96, and for thyroid eye disease from 0.79 to 0.89. The EMR signatures recapitulate known comorbidities with diabetes, such as abnormal glucose but do not significantly modulate image-derived features. In summary, EMR signatures present a scalable and readily applicable method for using EMR context to increase the statistical power of image derived features. Detailed explanation of the methods and results of this work is presented in Chapter VII.

### *3.3.4. Discovering Novel Disease Comorbidities using Electronic Medical Records*

Adaptation of digitized electronic medical records (EMR) at large hospitals and university medical centers provides unique opportunities to perform population level analyses for disease etiology and progression. In this work, we present a methodology that uses ICD-9 codes from large clinical databases to learn relationships between diseases. The tool can be used to explore co-occurring conditions, pre-morbid conditions and risk factors, disease progression and prognostic factors. Two primary concerns arise with association studies: the first is the problem of multiple comparisons which could lead to erroneous inferences. The second is that there is no statistic to

establish the novelty of a discovered association which is often useful in exploratory studies. In order to tackle these issues, we describe two new statistical measures. 1) We employ *second-generation* p-values to record those associations that are clinically meaningful, and calculate the positive predictive value (PPV), which denotes the probability that an association is a true finding. 2) We introduce a *novel finding index* (NFI), which indicates the extent to which the associations discovered via this framework are already well-known or well-studied in the scientific community.

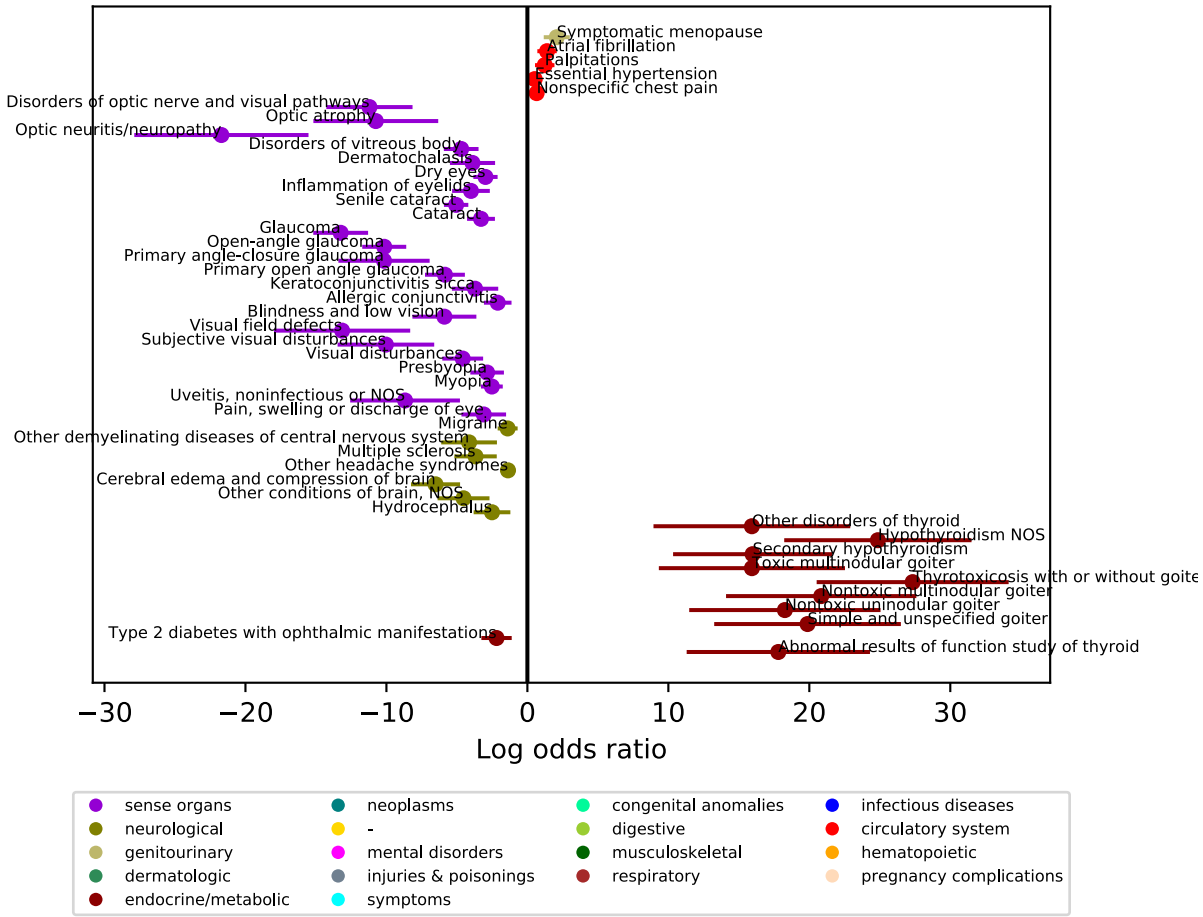



**Figure I-3. EMR signature for Thyroid Eye Disease (TED) constructed using our custom pyPheWAS package, computed on a preliminary dataset. Shown here are the log odds ratio for conditions seen in TED patients 1 year before the onset of the disease. The right side of the line shows positively correlated conditions and right side of the line shows conditions not seen in TED patients.**

This score is calculated based on calculating the frequency of hits from an automated PubMed search of a discovered relationship, and combining it with PPV. The approach described in this work is general can be used for any index disease as long as there is access to a large medical database which records ICD-9 codes. We show examples using three index diseases: Optic Neuritis, Alzheimer's and Autism Spectrum Disorder (ASD). Detailed explanation of the methods and results of this work is presented in Chapter VIII.

## Chapter II. Review of Technical Concepts

In this body of work, we present data-driven methods to identify biomarkers for diseases of the optic nerve and predict disease outcomes. In this chapter, we provide a brief overview of the concepts involved including types of clinically acquired data from electronic health records, key image analysis and machine learning methods.

### 1. Electronic Health Data

Electronic health records (EHR) store a patient's data in a digital form that can be easily accessed and processed by computers and, therefore, are conducive to big data analysis. EHR data comprises various types of records of a patient's health. It includes structured data about medications and their prescription dates, laboratory tests and their results, and administrative and billing codes. It also comprises of unstructured data such as clinical notes. EHR systems store digital formats of medical images acquired for diagnostic purposes. At Vanderbilt University, paper records have been eliminated completely since 2004 [86]. As a result, the Vanderbilt University's Research Derivative has longitudinal records of over two million patients spanning at least a decade. The rich data housed at Vanderbilt University is used for much of the research presented in this dissertation. In this section, we will elaborate on three main types of EHR data that was used in this work.

#### 1.1. Medical Images

Medical imaging describes the process of capturing visual representation of internal organs and structures in the human body for the purpose of diagnosis, intervention or evaluating function. Various techniques are used to capture images of internal structures depending on the application and the nature of the organ of interest. Some common imaging modalities used in medical

diagnosis are: magnetic resonance imaging, computed tomography, medical ultrasonography, positron emission tomography, and endoscopy. The main imaging modality studied in this work is computed tomography of the eye orbits.

Computed tomography (CT) obtains images of soft tissue structures and bones with X-ray measurements [65]. It was one of the first imaging techniques to distinguish between different tissue types owing to differential absorption of X-rays by materials of different densities. The CT equipment includes a helically rotating X-ray tube that acquires image of a volume at various angles. These images are then stitched together digitally and reconstructed to form a 3-dimensional representation of the internal structure. Practically, the CT scan can be represented as multiple slices of 2-dimensional images in any plane: sagittal, coronal or axial. These images are stored in DICOM (Digital Imaging and Communications in Medicine) format clinically. Unlike other imaging modalities, the intensity values in an image acquired through CTs can be interpreted biologically. The intensity of each voxel of a CT scan is represented on a Hounsfield scale [87]. On the Hounsfield scale, the original X-ray measurements are rescaled such that the density of water at standard temperature and pressure is 0, and that of air is -1000. These intensity values are called Hounsfield units (HU). Consequently, each type of tissue in the human body has a distinct range of HU associated with it.

The nature of CT images provides excellent contrast of the eye orbit which includes tissues of very different densities, namely: fat, extraocular muscles, the optic nerve and bone[65]. CT imaging is also quick and cheap to acquire, and can be rendered in any plane without significant artifacts. CT imaging is routinely used in diagnosis and treatment planning of optic nerve conditions. Most of the imaging data used in this work includes thin-slice (~1mm) CT scans of the orbit. The data are acquired in DICOM format, which contains PHI (Protected Health

Information). This data is then de-identified and converted to NifTI (Neuroimaging Informatics Technology Initiative) format, a data format which is used in medical imaging research. The anonymized NifTI files are stored on XNAT, an imaging informatics platform that facilitates efficient storage and processing of medical imaging data.

### 1.2. Administrative Data

EHR systems store different types of data collected during a clinic visit or hospitalization, which can be mined for information about a patient's history. A lot of data about a patient's health status is present in unstructured form such as clinic notes, discharge summaries, pathology reports and radiology reports. EHR systems also collect data collected that are used for billing and hospital administration. This data contain information about diagnosis and procedures in a codified form. The two main types of codes used ubiquitously are International Classification of Diseases (ICD) and Current Procedural Terminology (CPT). In much of this work, we use the administrative codes for analysis of a patient's medical record.

The ICD system is an international standard that is used to code diseases, signs and symptoms, injuries, mental disorders, and pregnancy and childbirth events[88]. These codes are provided each time a patient makes a visit to the clinic or a hospital. The 9$^{th}$ version of ICD coding system (ICD-9) is widely adapted in the medical informatics field. The ICD-9 codes are four- or five-digit codes that represent disease categories at different levels of granularity. The first three digits represent the main disease category, one or two decimal points signifying sub classifications of the main disease. For example, the code 415 represents acute pulmonary heart disease. 415.1 denotes pulmonary embolism, and 415.11 and 415.12 denote iatrogenic and septic pulmonary embolism respectively. Over 15,000 ICD-9 codes are used to represent all of the possible

diagnostic categories. Recently, newer versions ($10^{th}$ and $11^{th}$) of the ICD system have been introduced. However, in this work we use the $9^{th}$ version in this work as there has been extensive work in the field of informatics to extract clinical phenotypes from this classification system.

Current Procedural Terminology (CPT) codes are defined by the American Medical Association to describe procedures and services administered to patients[89]. CPT codes are five-digit codes that represent categories of medical procedures including surgeries, pathology examinations, labs, anesthesia, radiology exams, and therapeutic services. CPT and ICD codes together provide a complete picture of a patient's health over time. Since they are present in EHR systems in categorical form, they are ideal for data analysis.

### 1.3. Visual Disability Scores

Additional data collected in this study includes scores for visual testing for patients with optic nerve conditions. Routine eye examination involves a visual acuity test using a Snellen chart[90], where the patient will be asked to read letters that become smaller at each line. The line of the Snellen chart at which the patient can comfortably see determines their uncorrected visual acuity. Another examination involves visual field testing which examines how well the patient can see in the entire field of vision. This is done by either a Goldmann perimetry test or tangent screen testing[91]. The test can show defects in central scotoma, altitudinal defect, Bjerrum's scotoma, nasal step, depression of peripheral isopter or other field defects. American Medical Association defined nine visual outcomes[92], which are scores on a scale of 0-100 based to summarize a patient's visual function based visual acuity and field testing:

- Right and left visual acuity scores are calculated as $VAS_{od}$ and $VAS_{os}$ respectively.
- The visual acuity for both eyes, $VAS_{ou}$ is calculated as the best of $VAS_{od}$ and $VAS_{os}$.

- The functional acuity score, FAS is a weighted score of $VAS_{od}$, $VAS_{os}$, and $VAS_{ou}$ with weights 1:1:3. Right and left visual acuity scores are calculated as $VAS_{od}$ and $VAS_{os}$ respectively.
- Right and left visual field scores are calculated as $VFS_{od}$ and $VFS_{os}$ respectively.
- The visual field score for both eyes, $VFS_{ou}$ is calculated as the best of $VFS_{od}$ and $VFS_{os}$.
- The functional field score, FFS is a weighted score of $VFS_{od}$, $VFS_{os}$, and $VFS_{ou}$ with weights 1:1:3.
- An overall measure called functional visual score (FVS), is calculated as the average of FAS and FFS.

## 2. Medical Data Analysis

In this section, we present a brief overview of data analysis methods used in this work. Data analysis methods can be broadly classified into two parts, the first part involves image processing and informatics methods for data pre-processing and feature extraction. The second part involves using machine learning methods to learn outcomes from medical data.

### 2.1. Medical Image Processing

Medical images can be represented as a three-dimensional matrix of voxels, where each voxel denotes a corresponding volume in space of the structures being imaged. In the case of CT scans, the value of each voxel in the three-dimensional matrix corresponds to the density of the tissue volume that it represents. In other imaging modalities such as MRI, the voxel value denotes signal intensity based on the type of acquisition sequence such as T1, T2, PD or diffusion. Medical image processing can be described as the process of analyzing these matrices of voxels for better visualization, for interpretation, and to obtain quantitative measurements. In this section, we

discuss two important aspects of medical image processing, image registration and image segmentation.

### *2.1.1. Image Registration*

Image registration, as it relates to medical imaging, describes that process of transforming images to a common coordinate space to achieve spatial correspondence, i.e., voxels at each spatial coordinate in each of the registered images correspond to the same anatomical structure[93]. Registration is useful is many data analysis scenarios, for instance when images from multiple subjects are aligned to summarize average anatomical structural properties of a specific population, or when images of the same subject are compared over time to evaluate the effect of an intervention. Image registration is also often used as a preliminary step in image segmentation. Image registration can be divided into two main categories: rigid and nonrigid. Rigid registration can be described as a linear transformation of the image. Nonrigid registration (NRR) methods allow non-linear transformation of images.

The registration process requires a deformation model, which defines geometric constraints for the image transformation and an objective function which finds criteria to match or align two images. For rigid registrations, the deformation model can only allow translations and rotations. For non-rigid registration, the deformation model can have millions of parameters. Holden et al provide a review of the types of geometric transformations that are used in image registration algorithms [94]. For matching criteria, several methods can be used to drive the registration process. Some methods use anatomical landmarks to align images [95], some methods used external landmarks called fiducials[96]. Some studies propose using intensity-based similarity metrics such as cross-correlation to find correspondence between two images [97]. Sotiras et al

provide a review of all types of objective functions and their corresponding optimization procedures used in image registration [98].

*2.1.2. Image Segmentation*

Image segmentation of medical images can be described as a process of segmenting a digital image into composite parts, where each part has an anatomical meaning[99]. A medical image is usually represented as a collection of intensities in a matrix. The segmentation process helps identify a label that corresponds to a structure or organ of interest at each voxel. Quantitative measurements can be obtained for each structure of interest based on the segmentations.

A popular method for segmentation of images is multi-atlas segmentation where, the segmentations in a new target image are learnt based on examples. In this approach, a set of scans or *atlases* are provided along with expertly marked labels which denote the class membership of each voxel. Here, class can be considered to be any distinct anatomical structure of interest in the image volume. Each of these atlases are non-rigidly registered to the target image space such that there is voxel-wise correspondence between each example atlas and the target image, which can be used to learn the labels in the new target scan. Registrations often have errors associated with them and perfect correspondence cannot always be assumed. Several *label fusion* methods were proposed to solve the problem of combining or fusing labels from example atlases. The simplest approach is to simply take a majority vote at each voxel. Let $D \in \mathcal{L}^{R\times N}$ be the set of co-registered atlases, where N is the number of voxels, and R is the number of registrations. For each target voxel *j*, the probability that the voxel has a label *s*, is given by

$$P\left(X_j = s \middle| D\right) = \frac{1}{\sum_{i=1}^{R} \epsilon_{XD_i}} \sum_{i=1}^{R} \epsilon_{XD_i}\, \delta(s, D_{i,j})$$

where $\delta$ is the Kronecker delta given by,

$$\delta(a,b) = \begin{cases} 1, if\ a = b \\ 0, otherwise \end{cases}$$

The errors associated with registered atlases could arise from different sources including errors due to marking the labels incorrectly, and errors due to registration. Errors could also be varying spatially and might be different for each label. Instead of performing a majority vote, each of these errors could be modelled explicitly. That is, instead of learning $P(X_j = s|D)$, we could learn $P(X_j = s|D, \theta)$, where $\theta$, which indicates the latent distribution of error due to raters, location or label, is also learnt simultaneously. A review of different approaches to label fusion can be found in a paper by Eugenio et al [100].

2.2. EMR Phenotype Extraction

EMR contain several types of data including clinical summaries that are text data and billing codes, labs, medications which are categorical data. Historical data present in EHR records can be used to learn context-specific clinical information from which clinical decision-making regarding treatments and diagnosis can be improves. The main challenge in performing data analysis on clinical records is extracting clinical phenotypes from EHR data. A disease phenotype can be described as a set of characteristics or symptoms that consistently define the clinical presentation disease of interest. Automated methods of phenotype extraction have been proposed based on the type of clinical data. Clinical summaries such as physicians' notes, discharge summaries, pathology reports and radiology reports are written in natural language by a care provider. Several natural language processing (NLP) algorithms have been proposed to extract features from unstructured clinical texts. cTAKES was developed by the Mayo clinic to extract phenotypes from clinical text[101]. The PheKB catalog provides algorithms to extract clinical

phenotypes from clinical text, labs, medications and multiple sources in EMR[102]. Lia et al proposed i2b2, another natural language processing (NLP) system that used EMR to extract phenotypes[103]. Health Information Text Extraction (HITEx) is another tool that was developed to extract information from EMR including principal diagnosis, co-morbidities and smoking statues. These studies have shown promising results, but validation of these methods is a challenge. There is a lot of variation in styles and cultures of individual medical centers which introduce high variability in notes. In addition, there is a high chance that clinic notes are missing altogether due to inconsistent EHR practices between hospitals. All these issues make reproducibility for NLP algorithms a challenge.

Administrative codes such as ICD-9 codes and CPT codes provide a good alternative since they are used in a standard form across all hospitals. There are a few issues with the use of ICD9 codes for scientific research[104]: 1) The number of digits or the type of code used to describe a sign or symptom is inconsistent among doctors. 2) The hierarchical system here is designed such that there are multiple categories for common diseases and sometimes, just a single code for a complex disease—since they're primarily designed for the purposes for measuring hospital utilization and billing[85]. To overcome these problems, Denny et al designed a conversion system from ICD9 codes to Phecodes such that the conceptual granularity of the codes is consistent across diseases with the help of medical experts. Phecodes are widely used in genomic research and are generally considered an acceptable standard in medical conceptual categorization. In this work, we convert the ICD9 codes to Phecodes for EMR feature extraction. In addition to Phecodes, we introduce our own coding system called ProWAS codes, which are used to group CPT codes.

## 2.3. Machine Learning

In the previous section, we discussed methods to extract useful features from medical images and EHR data. In this section, we discuss machine learning methods that are used to learn outcomes from these features. Mitchell[105] formalizes machine learning as follows, "a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." In the case of medical data, the experience E is features derived from images and EHR records. Several types of tasks T, can be learnt from medical data e.g., predicting the outcome of an intervention, diagnosis of a disease, or risk stratification of patients by clustering. The performance measure P depends on the task T, e.g. P could be used to measure the error rate of a machine learning model that is trying to predict an outcome. There are two main types of machine learning problems, unsupervised learning and supervised learning, which are discussed below.

### *2.3.1. Unsupervised Learning*

In unsupervised learning, we are just given data without any labels[106]. The goal is to learn interesting structures from the data, without any knowledge about the desired output. One example of unsupervised learning is *clustering*. For example, given a set of data points of EMR records for subjects with optic neuropathy, including age, sex and symptoms, we would like to find clusters of disease subtypes in the data. Dimensionality reduction is another important type of unsupervised learning. It is used to project data in high-dimensional spaces (such as images) into a low dimensional subspace which captures the underlying information. A popular technique for dimensionality reduction is called principal component analysis (PCA). Given an $n$ −dimensional dataset, principal component analysis finds a lower, $d$ −dimensional subspace which can explain

most of the variance in the data. The $d$ −dimensional subspace can be represented by $d$ orthogonal vectors called principal components in a new coordinate system. Ghodsi provides a brief overview of different dimensionality reduction methods in [107].

*2.3.2. Supervised Learning*

Supervised learning involves learning from data that has labels or targets[106]. For instance, as experience E, we have a set of CT scans of the eye orbit that are labelled as healthy or as having glaucoma. The task T, is to learn an algorithm which correctly predicts or *classifies* a new image of the eye orbit into one of the two categories or classes. The performance measure of this task would measure accuracy of the model, such as a 0-1 loss. The 0-1 loss of a particular example is 1 if the model predicts its diagnostic class correctly, and 0 if not. A common classification algorithm that is used in this work is logistic regression, which models the posterior probabilities of *k* classes via a linear function in *x*[106]. In the case of two classes the model can be written as,

$$log\frac{P(y=1|X=x)}{P(y=2|X=x)} = \beta_0 + \beta_1^T x,$$

where $x$ denotes the data points, $y$ denotes the output class, and $\beta$s represent parameters to be estimated to learn the relationship between $x$ and $y$. Given N data points, let $p(x;\beta) = p(y = k|X = x;\beta)$. Then, the log-likelihood for this can be written as,

$$\ell(\beta) = \sum_{i=1}^{N}\{y_i \log p(x_i;\beta) + (1-y_i)\log(1-p(x_i;\beta))\}$$

Logistic regression model is fit by maximizing the log-likelihood function $\ell(\beta)$. This does not have a closed-form solution and an optimization procedure such as Newton-Raphson algorithm

or gradient descent is used to solve it. The complete derivation for this can be found in [106]. Often, when dimensionality of $x$ is large, it could lead to overfitting. In the limit case of dimensionality of $x$ equaling N, the model can fit exactly to the training data but might not generalize very well to new data samples. The problem of overfitting can be handled by introducing regularization term to the log-likelihood function during the optimization process. Zou and Hastie describe a method called elastic net which uses a combination of L1 and L2 penalties to learn a sparse model when the number of parameters is very high[108]. Logistic regression is a very powerful method with high interpretability and can be used for diagnostic classifiers. Langer et al used logistic regression to detect prostate cancer on MRI scans. Other classification methods such as support vector machines[109], [110] have been used for detection of breast cancer; and Random forests have been used for identification of Alzheimer's [111] .

Machine learning can be used to solve other types of problems such as regression, in which the outcome is a continuous variable instead of a category. These types of algorithms can be used to learn a continuous clinical outcome. For example, to predict the visual disability of a patient from their CT imaging (Chapter VI). In fact, even the segmentation problem can be rephrased as a machine learning problem where the model is trying to learn voxel labels in a new scan based on experience from example atlases. A review of machine learning methods in radiology can be found in [112].

### *2.3.3. Deep Learning*

Recently, deep learning algorithms have shown incredible improvements in performance in medical image analysis[113]. Traditional approaches to machine learning in radiology relied on construction of manually designed features for description of images. Deep learning methods, on

the other hand, learn representations automatically. They learn multiple-levels of representations starting at individual pixel- or voxel-level and get more abstract at higher levels. The higher levels of patterns retain discriminative features, while suppressing irrelevant features. Most algorithms in medical image analysis use convolutional neural networks (CNN)[114]. A CNN has many layers. The first few are convolutional and pooling layers. A CNN has sparse connections, since there is local correlation in images. The inputs for units in convolutional layer *m,* come from a subset of units in layer *m-1*. The local patch in layer *m-1*, called a local receptive field, connects to one hidden neuron in layer *m*. The local receptive field is slid over the entire region of layer *m-1* to get the units in layer *m.* All of these local networks share the same weights, so that local image features in an image can be learnt irrespective of their position in the visual field. This also drastically reduces the number of parameters. Pooling layers are used immediately after the convolutional layers, to merge semantically similar features into one unit. Several pairs of convolutional and pooling layers are typically followed by fully connected layers. The final layer has an activation function such as logistic function or SoftMax to give the final classification.

Deep learning in radiology has been used for medical image registration, Agrawal et al used features learned from deep learning for correspondence-based shape model which is used in statistical analysis of population anatomical shape[115]. Though multi-atlas segmentation methods have been the standard practice in medical imaging for a long time, several deep learning methods have been shown to improve upon stat-of-the-art results. Guo et al use SAE (Stacked Auto Encoder), a type of a deep neural network for hippocampus segmentation from infant MR images[116]. Roth et al present a method for automated detection of lymph nodes in CT images using a deep CNN[117]. Computer-aided diagnosis systems using deep learning are very popular in breast cancer screening [26], [118], [119]. They have also been used for TB screening[120],

Alzheimer's detection[121], and lung nodule detection[122]. Litjens et al conducted a survey of deep learning methods used in various application of medical imaging in [26].

## 2.4. Other Statistical Concepts

### *2.4.1. Correlation*

Often, in addition to predicting outcomes, one is also interested in learning correlations between features extracted from EHR data and clinical outcomes. Correlations can be used to examine the relationship between two variables of interest[123]. The most common type of correlation is the Pearson correlation, which examines a linear relationship between two variables. In this work, we use Kendall-Tau, which is a rank-based correlation as it is sensitive to non-linear relationships as well. Given two variables $X$ and $Y$, a pair $(x_i, y_i)$ and $(x_j, y_j)$ are called concordant if $x_i < y_i$ and $x_j < y_j$, or $x_i > y_i$ and $x_j > y_j$. $(x_i, y_i)$ and $(x_j, y_j)$ are called discordant if $x_i < y_i$ and $x_j > y_j$, or $x_i > y_i$ and $x_j < y_j$. The Kendall-Tau coefficient $\tau$ is defined by,

$$\tau = \frac{(number\ concordant\ pairs - number\ discordant\ pairs)}{total\ number\ of\ pairs}$$

To put it generally, the Kendall-Tau coefficient examines if X and Y increase together or X and Y decrease together.

### *2.4.2. Hypothesis Testing*

In several methods described in this section, we are testing if a certain hypothesis is true[124]. For instance, in section 2.4.1. we describe a method to examine if two variables are correlated, and in section 2.3.2. we describe methods to test if certain variables can be used learn

outcomes based on empirical data. The process of statistically determining the probability that a hypothesis is indeed true, is called hypothesis testing. In order to do so, we formulate a null hypothesis $H_0$ and an alternate hypothesis $H_\alpha$. For instance, the null hypothesis could be that the volume of an optic nerve is not correlated with visual function score. The alternate hypothesis for this would be that the volume of an optic nerve is correlated with visual function. Next, we identify a statistical test that can be used to test this hypothesis. For Kendall-Tau correlation, $\tau$ would be the test statistic. Then, we compute a $p - value$, which is the probability that the correlation $\tau$ would be observed if the null-hypothesis was true. If $p \leq \alpha$, where $\alpha$ (usually set to 0.05) is the level of significance, then $H_0$ is rejected and $H_\alpha$ is said to be valid.

Problems can arise when multiple comparisons or tests are conducted at the same time. Since each test is evaluated as a probability, the chance of making a false discovery increases with the number of tests. Several techniques are proposed to tackle this problem. One such correction is the Bonferroni correction, where $H_0$ is rejected if $p \leq \alpha/m$, where $m$ is the number of comparisons. The false discovery rate or the Benjamini-Hochberg procedure controls false discoveries by rejecting $H_0$ if $p_k \leq k.\alpha/m$, where $p_k$ is the $k^{th}$ largest p – value of the $m$ tests that satisfies the condition. Recently, a second-generation p-value was introduced by Blume et al[125], in which a null interval was suggested instead of a null-point hypothesis. In this case, the significant hypotheses have an effect size that fall completely outside of the pre-defined null interval

# Chapter III. Structural Functional Associations of the Orbit in Thyroid Eye Disease: Kalman Filters to Track Extraocular Rectal Muscles

*Parts of this chapter have been published in Medical Imaging 2016: Image Processing, vol. 9784, p. 97841G. International Society for Optics and Photonics, 2016.*

## 1. Introduction

Pathologies of the optic nerve and orbit, such as glaucoma, thyroid eye disease, multiple sclerosis, and optic neuritis impact millions of Americans. Successful treatment of these pathologies is sensitive to the early diagnosis. However, current diagnostic techniques are dependent on variable clinical presentations between patients and subjective clinical testing. A quantitative assessment of the orbital structures would provide objective markers to enhance diagnostic accuracy, improve timely intervention, and eventually, preserve visual function. Modern image processing and machine learning methods allow for the development of automated pipelines for large-scale analysis of these diseases. The primary task of such a pipeline is the automated identification of anatomical structures in the visual system, such as the optic nerve, extraocular rectal muscles, eye globe, and orbital fat, and automated computation of structural metrics to correlate with clinical characteristics. We have created a large-scale image processing and data analytics database on Pathologies of the Human Eye, Orbit, and The Optic Nerve (PHOTON) to better understand early disease stages, enable timely intervention, and improve disease management. PHOTON is a collection of electronic medical records and medical imaging spanning 5 major disease cohorts of 80 individual ICD-9 codes. As a pilot study on this database, we study thyroid eye disease using statistical label fusion methods and Kalman filters to identify orbital structures of interest and investigate correlations between these structures and eye functionality.

Presently, computed tomography (CT) imaging is the modality of choice in evaluating the orbit for evidence of thyroid eye disease[64]. In CT, the intensity of a pixel depends on the density of the tissue with respect to water. Therefore, distinct structures such as globe, nerve, muscle, and fat can be identified with a high accuracy. Extraocular muscle, bone, fat, and orbital volume indices are among the metrics used in previous study[61], [63], [85]–[96] as objective findings used to aide in early diagnosis.

Our novel analysis pipeline builds off multi-atlas segmentation methods. Briefly, a human expert labels the anatomical structures of interest in a set of representative training images (i.e., the atlases). The structures are identified in each target image by registering the training atlas and assigning a label to each voxel in the target image by statistical voting. Recent studies have shown that the multi-atlas methodology is suitable for identifying orbital structures[97]–[100]. However, challenges arise in the identification of the individual extraocular muscles that control eye movement. This is increasingly problematic in diseased eyes, where the muscles often appear to fuse (at the resolution of clinical CT) at the back of the orbit due to inflammation. We propose the use of Kalman filters to track the muscles in three-dimensions and identify individual extraocular rectus muscles. The purpose of our study is to investigate a method of automatically generating orbital metrics from CT imaging and correlating these to known clinical characteristics.

## 2. Methodology

### 2.1. Data

Eighty-five subjects between the ages on 18-85 were selected based on having met clinical criteria for thyroid eye disease and undergoing CT imaging as part of their regular clinical care. Of these, sixty-three (74%) were female. Variable CT imaging protocols (head, orbital,

maxillofacial, etc.) were acquired and the highest resolution scan without severe orbital artifact and with a field of view including the full optic nerves was manually selected for each of these 85 patients. Clinical characteristics including demographic information, ocular mobility, visual acuity, color vision, and visual field testing were recorded. Institutional Review Board approval for this retrospective study was obtained at Vanderbilt University.

Visual disability was assessed with the American Medical Association Functional Vision Score (FVS) which "provides criteria for evaluating permanent impairment of the visual system as it affects an individual's ability to perform activities of daily living" as a percentage of disability relative to a healthy control[101]. The FVS is characterized by four sub-scores: an individual assessment of visual acuity in each eye, Visual Acuity Score (VAS); a composite of visual acuity over both eyes, Functional Acuity Score (FAS); individual assessment of field perception in each eye, Visual Field Score (VFS); and the composite of field perception over both eyes, Functional Field Score (FFS).

2.2. Multi-Atlas Segmentation

The selected CT image for each patient was loaded into eXtensible Neuroimaging Archive Toolkit [102], [103] and automatically segmented using a previously described multi-atlas segmentation pipeline which uses non-local STAPLE, a label fusion algorithm, to identify the optic nerves (including surrounding CSF sheaths), rectus muscles, globes, and orbital fat. Briefly, segmentation followed a multi-atlas labeling framework[104] in which a set of manually labeled example scans were non-rigidly registered to each patient's scan and statistical fusion was used to combine the labels from each of the examples to estimate the structure for each point in the target

scan. Figure III-1 (a) shows axial view of an input CT scan and 1 (b) shows the result of the multi-atlas segmentation pipeline.

2.3. Kalman Filters

We use Kalman filters to identify the Superior Rectus Muscle, Inferior Rectus Muscle, Lateral Rectus Muscle and Medial Rectus Muscle from the muscle labels obtained from the multi-atlas segmentation pipeline shown in Figure III-1(b). For each image volume, we start at a coronal slice at the center of the globe, where the muscles are well-separated, and use Kalman filters to track each muscle in the anterior to posterior direction (z-direction). The globe and the optic nerve pass through the center of the orbit and can be used as landmarks at each coronal slice to help identify the muscle positions. Five Kalman filters are defined for each of the four muscles and the landmarks. To keep the model simple, the centroids of the structures are used for tracking. Therefore, the state of system is the centroid of the two-dimensional slice in the coronal plane. The predicted state at slice $z$, given state $z-1$ is

$$[\bar{x}_z \bar{y}_z] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} [x_{z-1} y_{z-1}] + \epsilon \quad (1)$$

where $\epsilon$ is the process error defined by covariance R. That is, the process expects the muscle/nerve structure to be in the same position within a margin of error. Note that the control vector is eliminated as the tracking "moves" in the z-direction at a constant rate. At each step, the measured positions are given by a watershed calculation on the distance transform of a coronal slice, as seen in Figures III-1 (c) and (d). The predicted muscle positions of the previous step are used to impose a maxima on the distance transform. The current predicted position of each filter is then given by,

$$[x_{predicted}, y_{predicted}] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} [x_{watershed}, y_{watershed}] + \delta \quad (2)$$

where $\delta$ is the measurement error defined by covariance Q.

For each image volume, a region of interest is selected as the set of all the coronal slices containing muscle, globe, and the optic nerve. A Kalman filter is initiated for the centroids of globe and optic nerve, which are used as landmarks to identify the initial muscle positions.

Initial labels for the rectus muscles are assigned based on their relative position to the landmark filter. Once the algorithm finds the first slice containing a rectus muscle, it initiates a Kalman filter for the muscle at that slice. The Kalman filter is then used to track that specific rectus muscle until the end of the orbit. The predicted mean $\bar{\mu_z}$, and variance $\bar{\sigma_z}$ of the state at each position are given by,

$$\bar{\mu}_z = A\mu_{z-1} \quad (3)$$

$$\bar{\sigma}_z = A\sigma_{z-1}A^T + R \quad (4)$$

Here, A is the identity matrix and R is the process error covariance. The Kalman gain $K$ is given by,

$$K = \bar{\sigma}_z C^T (A\sigma_{z-1}A^T + Q^T) \quad (5)$$

This yields a predicted position $\mu_z$ and variance:

$$\mu_z = \bar{\mu}_z + K(M_z - C\bar{\mu}_z) \quad (6)$$

$$\sigma_z = (1 - KC)\bar{\sigma}_z \tag{7}$$

where, C is the identity matrix and Q is the measurement covariance.

2.4. Automated Structural Metric Calculation

Following multi-atlas segmentation and extraction of individual rectus muscles using Kalman filters, we compute descriptive features from the segmentation of the orbital anatomy for each patient to assess correlations between functional or clinical data and structural measures. These features included the (1-13) volume, maximum diameter, and average diameter for the superior, inferior, medial, and lateral rectus muscles and total rectus muscle volume[62], [92], [105], [106]; (14) Barrett index[90]; (15,16) volume and diameter of the globe[54], [61], [106], [107]; (17) orbital volume; (18) volume crowding index[93]; (19) orbital angle; (20) degree of proptosis; and (21-24) length, volume, average area, and maximum diameter of the optic nerve[108], [109]. All metrics were performed bilaterally, which resulted in 24 measures for each eye.

## 3. Results

Kalman filters provide a convenient mechanism to distinguish between individual rectus muscles in eye orbits segmented in the multi-atlas framework. Error covariance values for the process and the measurement are determined heuristically as follows:

$$R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{8}$$

$$Q = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix} \tag{9}$$

Figure III-2 shows representative final segmented images in two-dimensional and three-dimensional views. To assess the utility of this approach we compute geometrical metrics based on the identified orbital structures and see how they correlate with the subject's clinical data which is routinely used in the diagnosis of ocular disease[91], [92], such as visual disability scores, ocular motility, hertel score, color vision and neuropathy.

In this pilot study, we compute the correlation between structural metrics from the segmentations and clinical data using a Spearman correlation. The results of univariate correlations for orbital metrics (including orbital volume, volumetric crowding index[96], and proptosis) and optic nerve metrics (include length, cross sectional area, volume, and diameter) are shown in Table III-1. Similarly, the correlations for clinical characteristics with muscle metrics are shown in Table III-2. Notice from Table III-1 that Hertel index measurements demonstrated strong correlation with nearly all optic nerve and orbital metrics. Additional findings included a strong correlation

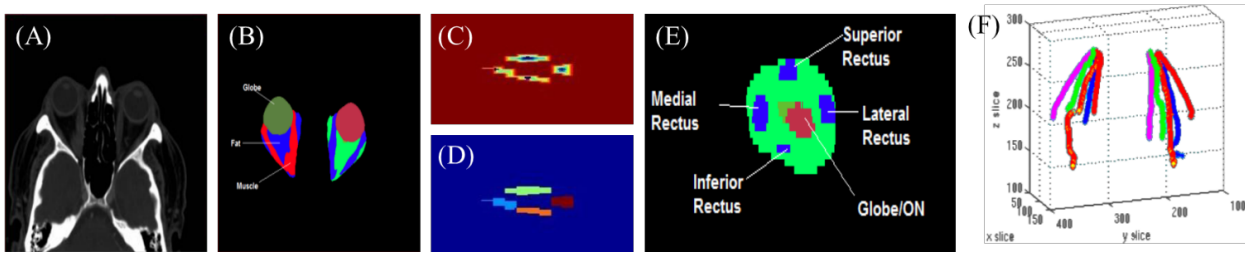



**Figure III-1. Muscle tracking using Kalman filters: (A) Input CT scan (B) output of label fusion (C) distance transform with imposed maxima (D) measured muscle pieces (E) Coronal view of muscles with respect to landmark (F) the resultant five Kalman filters, for each eye, tracking the muscles. Note that the muscles are well-separated in the front of the orbit (B), but as we approach the back of the orbit there is no longer a clean boundary between them due to inflammation and crowding. The measured positions at each slice are given by a watershed calculation as shown in (C) and segmented in (D). A distance function is calculated over slice z wherein the value of each pixel is given by the distance to its nearest non-zero pixel as shown in (C), creating a contour where there is a maxima at the center of each of the four muscles. At each slice, the algorithm examines the top, bottom, right, and left quadrants for each of the four muscles as see in (E). In total, ten Kalman filters (5 for each eye) are used to track the muscles (F) which results in the 3-D tracks shown in (F).**

between visual acuity and volumetric crowding index and between smoking and degree of proptosis.

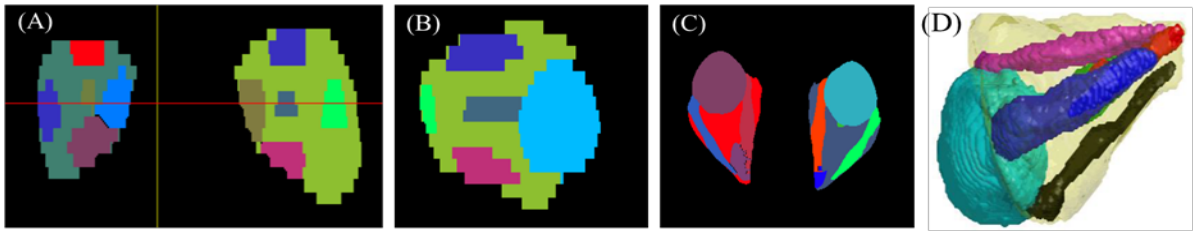



**Figure III-2. Representative segmentation in different views: (A) Coronal, (B) sagittal, (C) axial, and (D) 3-dimensional.**

**Table III-1. Spearman rho correlations of orbital metrics with clinical characteristics. Yellow highlights indicate a p value less than 0.1, and**

| Orbit metrics | Logmar visual acuity | Color vison (ratio) | Smoking (no, former, current) | Motility defect (no, yes) | Hertel 's index | Optic neuropathy (no, yes) | Best corrected visual acuity | Visual acuity score | Functional acuity score |
|---|---|---|---|---|---|---|---|---|---|
| Barrett Index | 0.04 | -0.01 | 0.06 | 0.21 | -0.03 | -0.03 | 0.13 | -0.16 | -0.16 |
| Globe Volume | 0.18 | -0.04 | 0.02 | 0.13 | -0.04 | -0.07 | 0.27 | -0.20 | -0.04 |
| Globe Diameter | 0.18 | -0.04 | 0.01 | 0.12 | -0.04 | -0.07 | 0.27 | -0.20 | -0.04 |
| ON Length | 0.01 | -0.01 | 0.06 | -0.01 | 0.13 | -0.09 | -0.14 | 0.10 | 0.11 |
| ON Length Traditional | 0.00 | -0.07 | 0.10 | -0.02 | 0.26 | -0.09 | -0.20 | 0.13 | 0.12 |
| ON Volume | 0.11 | 0.10 | 0.01 | 0.02 | -0.15 | -0.15 | 0.28 | -0.25 | -0.16 |
| ON Avg Area | 0.10 | -0.01 | -0.07 | 0.01 | -0.22 | -0.12 | 0.31 | -0.26 | -0.19 |
| ON Max Diameter | 0.07 | -0.02 | -0.05 | 0.04 | -0.21 | -0.16 | 0.18 | -0.12 | -0.12 |
| Orbital Volume | 0.07 | -0.07 | 0.08 | 0.16 | 0.23 | -0.07 | 0.02 | -0.06 | -0.04 |
| Volumetric Crowding Index | 0.15 | 0.02 | -0.14 | -0.02 | -0.23 | -0.08 | 0.24 | -0.22 | -0.14 |
| Proptosis | 0.06 | 0.08 | 0.13 | -0.05 | 0.35 | -0.07 | -0.06 | 0.00 | 0.02 |
| Angle | 0.08 | 0.02 | 0.06 | 0.00 | 0.14 | 0.03 | 0.11 | -0.10 | -0.09 |

Several extraocular muscle metrics including average diameter, maximum diameter, and muscle volume of the superior rectus, inferior rectus, and lateral rectus demonstrated strong correlation (p-value <0.05) with the presence of ocular motility deficit. On the other hand, medial rectus muscle demonstrated only a mild correlation with motility deficit. Color vision measurements demonstrated strong correlation (p-value <0.05) with inferior rectus, medial rectus,

and superior rectus muscle maximum diameters while demonstrating a mild correlation (p-value = 0.056) with the lateral rectus maximum diameter. Additional findings included a strong correlation between visual acuity and superior rectus maximum diameter.

**Table III-2. Spearman rho correlations of muscle metrics with clinical characteristics. Yellow highlights indicate a p value less than 0.1, and green**

| Muscle metrics | Logmar visual acuity | Color vison (ratio) | Smoking (no, former, current) | Motility defect (No, Yes) | Hertel 's index | Optic neuropathy (No, Yes) | Best corrected visual acuity | Visual acuity score | Functional acuity score |
|---|---|---|---|---|---|---|---|---|---|
| Superior Muscle Volume | 0.19 | -0.13 | 0.01 | 0.10 | -0.17 | -0.09 | 0.17 | -0.16 | -0.15 |
| Inferior Muscle Volume | 0.04 | -0.02 | 0.13 | 0.15 | 0.00 | -0.14 | 0.03 | -0.06 | -0.04 |
| Lateral Muscle Volume | 0.09 | -0.02 | 0.15 | 0.11 | 0.04 | -0.01 | 0.03 | -0.09 | -0.06 |
| Medial Muscle Volume | 0.11 | -0.05 | 0.05 | 0.10 | 0.04 | -0.05 | 0.13 | -0.19 | -0.20 |
| Superior Muscle Diameter Avg | 0.22 | -0.05 | 0.07 | 0.14 | -0.07 | 0.01 | 0.19 | -0.17 | -0.17 |
| Inferior Muscle Diameter Avg | 0.15 | -0.09 | 0.17 | 0.17 | 0.04 | -0.05 | 0.04 | -0.04 | 0.00 |
| Lateral Muscle Diameter Avg | 0.07 | -0.03 | 0.12 | 0.13 | 0.03 | -0.02 | 0.02 | -0.07 | -0.09 |
| Medial Muscle Diameter Avg | 0.12 | -0.05 | 0.05 | 0.09 | 0.02 | 0.00 | 0.19 | -0.25 | -0.26 |
| Superior Muscle Diameter Max | 0.22 | -0.14 | 0.02 | 0.12 | -0.03 | 0.07 | 0.14 | -0.12 | -0.11 |
| Inferior Muscle Diameter Max | 0.12 | -0.17 | 0.16 | 0.21 | 0.11 | -0.03 | 0.00 | -0.05 | -0.15 |
| Lateral Muscle Diameter Max | 0.09 | -0.07 | 0.13 | 0.12 | 0.05 | 0.00 | 0.11 | -0.16 | -0.17 |
| Medial Muscle Diameter Max | 0.16 | -0.08 | 0.03 | 0.03 | 0.02 | -0.10 | 0.13 | -0.16 | -0.12 |
| Total Muscle Volume | 0.14 | -0.06 | 0.10 | 0.14 | 0.01 | -0.04 | 0.08 | -0.12 | -0.10 |

## 4. Discussion

Identifying the individual ocular structures has significant advantages for the diagnosis, analysis and study of ocular diseases. Traditional methods of orbital segmentation are manual and tedious, or cannot capture the 3-D structure of orbital anatomy. In this study, we have shown the significant structural-functional correlations of the orbital structures such as muscles, the optic nerve, and eye globe with visual function and other clinical data. Further, we highlight the importance of studying each muscle separately by establishing that they have varying degrees of

predictive power. We demonstrate that Kalman filters provide a simple, yet fast and efficient solution to improving muscle segmentation in diseased eye. The strong correlation demonstrated by several clinical characteristics with the automatically obtained orbital metrics serves as foundation for further investigation. In future study, machine learning methods can be employed to find latent features in these structures and robust models of disease and treatment can be built based on these algorithms.

## Chapter IV. QUADRATIC: Quality of Dice in Registration Circuits

*Parts of this chapter have been published in Medical Imaging 2018: Image Processing, vol. 10574, p. 105740P. International Society for Optics and Photonics, 2018.*

### 1. Introduction

Non-rigid registration (NRR) is used in various applications in the medical image processing community, such as comparing anatomical structures between patients, tracking change in anatomical structures over time within the same patient, and performing segmentation using multi-atlas algorithms[144], [150]. The ability to estimate the accuracy of NRR is vital in the correct interpretation of these results. Specifically, in use cases such as multi-atlas segmentation, the ability to quantify the magnitude and location of error associated with NRR procedures is important as it can propagate to further steps of the segmentation process.

In the past, NRR was evaluated either subjectively by visual inspection [151] or offline evaluation of atlases based on an image similarity metric[152], [153]. The optimization process used by NRR inherently biases image similarity metrics, and often is not predictive of structural correspondence. Quantifying the error in correspondence as a result of NRR in a target image is a challenging problem, with some calling automatic registration techniques impossible to validate[153]. Datteri et al have developed the AQUIRC (Assessing Quality Using Image Registration Circuits)[154], [155] model to assess registration quality is target images using registration circuits. The authors proposed to use registration circuits to model the local error associated with NRR and have shown moderate success in applying this to multi-atlas segmentation[82]. In the AQUIRC model, the error "quality" in edges of the circuits is measured. In our model, we use the idea of image registration circuits to estimate the actual error associated with each non-rigid deformation, in this case, the Dice Similarity Coefficient (DSC) of the edges

in the network. To model DSC in the network, we test four models to determine the relationship between edge DSC and circuit DSC. The success or failure of the non-rigid registration is predicted using the error model and is used for atlas selection in majority vote multi-atlas segmentation. We find that the quadratic error model is the most successful in modeling DSC error in registration circuits, we call this model QUADRATIC (QUAlity of Dice in RegistrATIon Circuits).

## 2. Methods

### 2.1. Data

The data consists of 16 computed tomography scans of eye orbit collected at the Vanderbilt University Medical Center which were retrieved and de-identified under institutional review board approval and previously described in[156]. The following structures in the eye orbit were labeled by experts to form 16 atlases: the optic nerve, the globe, the rectus muscles, and orbital fat. A leave-one-out approach was used where one of the atlases was treated as a target scan, while the other 15 were used for NRR error modeling and segmentation.

### 2.2. Registration

Each of the 16 atlases was registered to the other 15 atlases using ANTS SyN registration[157], with cross-correlation metric. The registration process non-rigidly transforms each atlas to the target space such that a voxel-wise correspondence between the atlas and the target is estimated. The deformation field that was obtained by registering the atlases to the target space was used to propagate the expert labels from the atlas space to the target scan using nearest neighbor interpolation.

The quality of the registration was measured based on structural correspondence between the target image and registered atlas. The structural correspondence was measured using the Dice

Similarity Coefficient (DSC). It gives a value between 0 and 1 indicating similarity between the target and the registered atlas, by calculating the proportion of voxels that are the same in both images. It is given by,

$$DSC = 1 - \frac{2|X \cap Y|}{|X| + |Y|}$$

The DSC value is computed for each of the four structures bilaterally as well as for the whole segmentation, we will refer to these DSC values as NRR-DSC (Non-Rigid registration Dice similarity coefficient), in order to differentiate it from DSC values obtained from segmentation results that will be computed later in the paper.

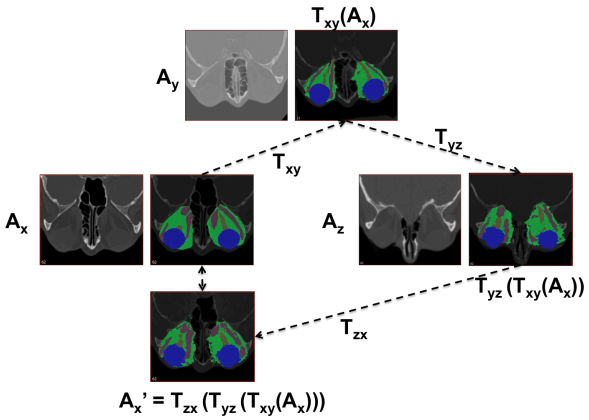



**Figure IV-1. Example of an image registration circuit. The three atlases in the circuit, $A_x, A_y, A_z$, and the transformations $T_{xy}, T_{yz}$, and $T_{yz}$. $A_x$ is transformed to $A_x' = T_{zx}\left(T_{yz}\left(T_{xy}(A_x)\right)\right)$**

## 2.3. Estimating the Error Model with Registration Circuit

This paper builds on the idea of AQUIRC to use registration circuits of size 3 to estimate registration error[82]. To model the error associated with NRR for each target scan ($A_s$), the other 15 scans are used as atlases ($A_1 - A_{15}$), with known true labels, to construct $15p3$ circuits ($n = 2730$), having $15p2$ edges ($m = 210$). Each registration circuit consists of three atlases $A_x$, $A_y$,

and $A_z$ as shown in Figure IV-1. The non-rigid transformations from $A_x$ to $A_y$, $A_y$ to $A_z$ and $A_z$ to $A_x$, are computed as $T_{xy}$,$T_{yz}$, and $T_{zx}$ respectively. The deformation fields calculated from the non-rigid transformations are used to propagate the labels associated with $A_x$ through the circuit. $A_x$ is transformed to image space $A_y$ using transformation, $T_{xy}(A_x)$. Next, it is transformed to image space $A_z$ using $T_{yz}\left(T_{xy}(A_x)\right)$, and then back to its original space to get $A'_x = T_{zx}\left(T_{yz}\left(T_{xy}(A_x)\right)\right)$. The error associated with each of the edges of the circuit is evaluated by computing the NRR-DSC value between the true labels and propagated labels from the non-rigid transformation for each edge to give $\varepsilon_{xy}$, $\varepsilon_{yz}$, and $\varepsilon_{zx}$. The NRR-DSC value between true labels of $A$ and the propagated labels of $A'$ is the error associated with the registration circuit given by $E_{xyz}$. In this paper, we evaluate the non-linear relationship between edge error, $\varepsilon$ and circuit error $E$ with a Taylor series expansion up to the third order in $f_1$ (linear), $f_2$ (quadratic), $f_3$ (third order), and also a multiplicative relationship in $f_4$, which are given by

$$f_1(\varepsilon_{xy}, \varepsilon_{yz}, \varepsilon_{zx}) = E_{xyz} = p_0 + p_1\varepsilon_{xy} + p_2\varepsilon_{yz} + p_3\varepsilon_{zx} \quad (1)$$

$$f_2(\varepsilon_{xy}, \varepsilon_{yz}, \varepsilon_{zx}) = E_{xyz} = p_0 + p_1\varepsilon_{xy} + p_2\varepsilon_{yz} + p_3\varepsilon_{zx} + p_4\varepsilon_{xy}^2 + p_5\varepsilon_{yz}^2 + p_6\varepsilon_{zx}^2 + p_7\varepsilon_{xy}\varepsilon_{yz} + p_8\varepsilon_{yz}\varepsilon_{zx} + p_9\varepsilon_{zx}\varepsilon_{xy} \quad (2)$$

$$f_3(\varepsilon_{xy}, \varepsilon_{yz}, \varepsilon_{zx}) = E_{xyz} = p_0 + p_1\varepsilon_{xy} + p_2\varepsilon_{yz} + p_3\varepsilon_{zx} + p_4\varepsilon_{xy}^2 + p_5\varepsilon_{yz}^2 + p_6\varepsilon_{zx}^2 + p_7\varepsilon_{xy}\varepsilon_{yz} + p_8\varepsilon_{yz}\varepsilon_{zx} + p_9\varepsilon_{zx}\varepsilon_{xy} + p_{10}\varepsilon_{xy}^3 + p_{11}\varepsilon_{yz}^3 + p_{12}\varepsilon_{zx}^3 + p_{13}\varepsilon_{xy}^2\varepsilon_{yz} + p_{14}\varepsilon_{yz}^2\varepsilon_{zx} + p_{15}\varepsilon_{zx}^2\varepsilon_{xy} + p_{16}\varepsilon_{xy}\varepsilon_{yz}^2 + p_{17}\varepsilon_{yz}\varepsilon_{zx}^2 + p_{18}\varepsilon_{zx}^2\varepsilon_{xy} + p_{19}\varepsilon_{xy}\varepsilon_{yz}\varepsilon_{zx} \quad (3)$$

$$f_4(\varepsilon_{xy}, \varepsilon_{yz}, \varepsilon_{zx}) = E_{xyz} = p_0 * p_1\varepsilon_{xy} * p_2\varepsilon_{yz} * p_3\varepsilon_{zx} \quad \textbf{or} \quad \log E_{xyz} = \log p_0 + \log p_1\varepsilon_{xy} + \log p_2\varepsilon_{yz} + \log p_3\varepsilon_{zx} \quad (4)$$

Note that the edge errors are known for all of the 210 edges of the circuits since the true labels are known. The parameters, $p_i$ in models 1 through 4 are evaluated using the 2730 circuits using a generalized linear model with equations (1)-(4). Here, the response variable is the circuit error $E_{xyz}$, and the independent variables are edge errors of the form $\varepsilon_{xy}$, $\varepsilon_{yz}$, and $\varepsilon_{zx}$. The estimated parameters will be indicated by $\widehat{p_i}$.

Next, the estimated parameters and known errors $\varepsilon$, are used to learn the error between the target scan $A_s$, and each of the 15 atlases, $A_{1-15}$. $15p2$ ($n'$=210) circuits are constructed with target scan $S$, where the circuit starts with an atlas. For example, given atlases $A_x$, and $A_y$, the following circuits are constructed: $A_x \rightarrow T_{xy}(A_x) \rightarrow T_{ys}\left(T_{xy}(A_x)\right) \rightarrow A'_x$ and $A_x \rightarrow T_{xs}(A_x) \rightarrow T_{sz}\left(T_{xs}(A_x)\right) \rightarrow A'_x$. This leads to a system of 210 equations with 30 unknown variables $(\varepsilon_{1s}, \varepsilon_{2s}, \dots, \varepsilon_{15s}, \varepsilon_{s1}, \varepsilon_{s2}, \dots, \varepsilon_{s15})$. For models 1 through 4, the systems of equations are:

1. $\hat{f}_1(\varepsilon_{xy}, \varepsilon_{ys}, \varepsilon_{sx}) = E_{xys} = \hat{p}_0 + \hat{p}_1\varepsilon_{xy} + \hat{p}_2\varepsilon_{ys} + \hat{p}_3\varepsilon_{sx}$

2. $\hat{f}_2(\varepsilon_{xy}, \varepsilon_{ys}, \varepsilon_{sx}) = E_{xys} = \hat{p}_0 + \hat{p}_1\varepsilon_{xy} + \hat{p}_2\varepsilon_{ys} + \hat{p}_3\varepsilon_{sx} + \hat{p}_4\varepsilon_{xy}^2 + \hat{p}_5\varepsilon_{ys}^2 + \hat{p}_6\varepsilon_{sx}^2 + \hat{p}_7\varepsilon_{xy}\varepsilon_{ys} + \hat{p}_8\varepsilon_{ys}\varepsilon_{sx} + \hat{p}_9\varepsilon_{sx}\varepsilon_{xy}$

3. $\hat{f}_3(\varepsilon_{xy}, \varepsilon_{ys}, \varepsilon_{sx}) = E_{xys} = \hat{p}_0 + \hat{p}_1\varepsilon_{xy} + \hat{p}_2\varepsilon_{ys} + \hat{p}_3\varepsilon_{sx} + \hat{p}_4\varepsilon_{xy}^2 + \hat{p}_5\varepsilon_{ys}^2 + \hat{p}_6\varepsilon_{sx}^2 + \hat{p}_7\varepsilon_{xy}\varepsilon_{ys} + \hat{p}_8\varepsilon_{ys}\varepsilon_{sx} + \hat{p}_9\varepsilon_{sx}\varepsilon_{xy} + \hat{p}_{10}\varepsilon_{xy}^3 + \hat{p}_{11}\varepsilon_{ys}^3 + \hat{p}_{12}\varepsilon_{sx}^3 + \hat{p}_{13}\varepsilon_{xy}^2\varepsilon_{ys} + \hat{p}_{14}\varepsilon_{ys}^2\varepsilon_{sx} + \hat{p}_{15}\varepsilon_{sx}^2\varepsilon_{xy} + \hat{p}_{16}\varepsilon_{xy}\varepsilon_{ys}^2 + \hat{p}_{17}\varepsilon_{ys}\varepsilon_{sx}^2 + \hat{p}_{18}\varepsilon_{sx}^2\varepsilon_{xy} + \hat{p}_{19}\varepsilon_{xy}\varepsilon_{ys}\varepsilon_{sx}$

4. $\log \widehat{f_4}(\varepsilon_{xy}, \varepsilon_{ys}, \varepsilon_{sx}) = \log E_{xys} = \log \hat{p}_0 + \log \hat{p}_1\varepsilon_{xy} + \log \hat{p}_2\varepsilon_{ys} + \log \hat{p}_3\varepsilon_{sx}$

The systems of equations are solved using the least squares non-linear (lsqnonlin) function in MATLAB for estimated edge errors:$\hat{\varepsilon}_{1s}, \hat{\varepsilon}_{2s}, \dots, \hat{\varepsilon}_{15s}, \hat{\varepsilon}_{s1}, \hat{\varepsilon}_{s2}, \dots, \hat{\varepsilon}_{s15}$.

2.4. Majority Vote with Atlas Selection

Multi-atlas segmentation based on majority vote is used to segment each target scan, with the other 15 scans used as atlases. The quality of the segmentation is dependent on the quality of the non-rigid registrations. Poor registrations to the target scan can worsen the final result of the segmentation. Predicting the NRR-DSC associated with each registration as described in section 2.3 can help determine the atlases with best registration for atlas selection. The NRR-DSC estimates are sorted in descending order from the atlas with the best predicted-NRR-DSC ($A'_1$) to the worst predicted-NRR-DSC ($A'_{15}$): $\hat{\varepsilon}_{1's}, \hat{\varepsilon}_{2's}, \dots, \hat{\varepsilon}_{15's}$. In order to determine the number of atlases to select to achieve the maximum gain in performance, we perform multi-atlas segmentation 15 times, starting with the best atlas and adding the next best one until all the atlases are used. For the $k^{th}$ segmentation, let $A^{\{k\}} = L^{k\times N}$ be the set of co-registered atlases where $N$ is the number of voxels, and $k$ is the number of registrations. For each target voxel $j$, the probability that the voxel has a label $s$, is given by

$$P\left(X_j = s \middle| D\right) = \frac{1}{k}\sum_{i=1}^{k} \delta(s, D_{i,j})$$

where, $\delta$ is the Kronecker delta given by,

$$\delta(a,b) = \begin{cases} 1, if\ a = b \\ 0, otherwise \end{cases}$$

The label decision at voxel $j$ is given by,

$$\arg_s \max\left(P\left(X_j = s \middle| D\right)\right)$$

For each of the segmentations, the DSC between the true labels and the segmented labels is calculated to assess the quality of the segmentation. We will refer to this as the seg-DSC, in order to differentiate it from the NRR-DSC.

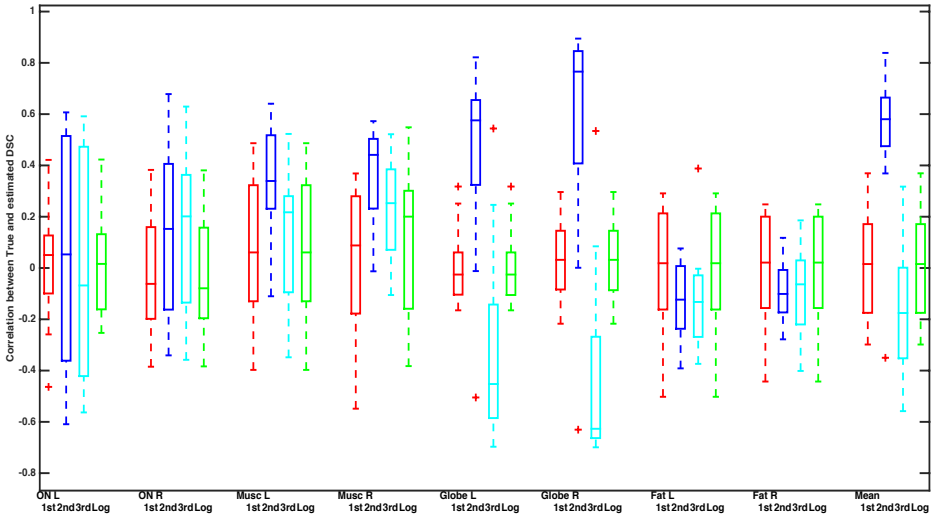



**Figure IV-2. The correlation between true and predicted NRR-DSC for each structure and the overall segmentation are shown for the four models. Red indicates the linear model, blue indicates the quadratic model, cyan indicates the third order model, and green indicates the multiplicative model.**

## 3. Results

### 3.1. Error Model

We had a set of 16 atlases of the eye orbit with expertly marked labels for bilateral orbital structures including the globe, the optic nerve, the extraocular muscles, and periorbital fat. We used a leave-one-out approach to find the error model of the non-rigid registration by considering one of the atlases to be a target scan whose labels were unknown and using the other 15 scans as atlases. The NRR-DSC of the non-rigid registration from each of the 15 atlases to the target space is calculated as described in section 2.3 for each of the structures and the overall segmentation. The predicted NRR-DSC values using each model are compared to the true NRR-DSC values for

each target atlas. The mean of the correlation values for the 16 target scans is shown in Figure IV-2. The quadratic error model has the highest correlation (p-value<0.05) with the true NRR-DSC values for all of the structures except periorbital fat. The mean correlation between true and predicted NRR-DSC values for overall segmentation using the quadratic model was 0.58 (p-value<0.001). We call this quadratic error model, QUADRATIC (QUAlity of Dice in RegistrATIon Circuits).

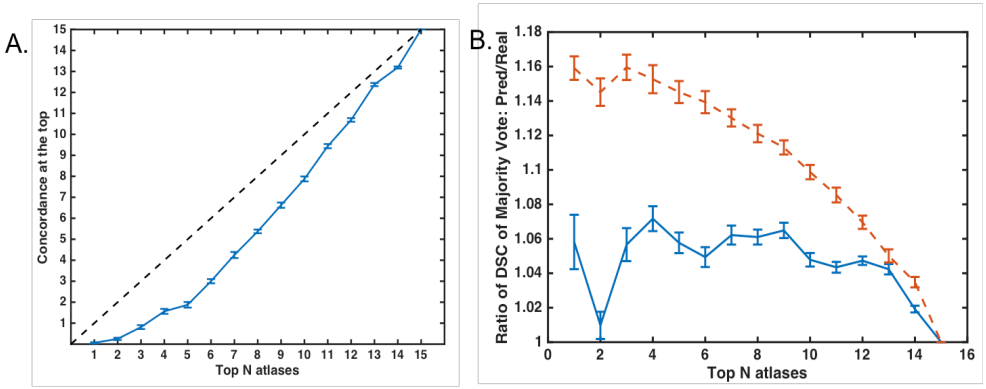



**Figure IV-3. A. Concordance at the top between true and estimated NRR-DSC, i.e., the overlap between top k atlases and the predicted top k atlases. B. Majority vote results for the 15 segmentations, where $k^{th}$ segmentation has the top k predicted atlases (blue line), and true top k atlases (red dotted line).**

## 3.2. Majority Vote Results

The mean NRR-DSC predicted using QUADRATIC is used in atlas selection. Figure IV-3A shows the concordance at the top between the true and estimated mean NRR-DSC values using the quadratic model. Concordance at the top for k atlases calculates the overlap between top k atlases based on true NRR-DSC and top k atlases based on estimated NRR-DSC. As described in section 2.4, 15 segmentations are performed for each target scan. For each of the segmentations,

the quality of the segmentation is measured by computing the seg-DSC between the true labels and the segmented labels. The ratio of seg-DSC between the top k atlases and all 15-atlases is computed to evaluate the gain in performance. Figure IV-3B shows the gain in performance by selecting the best atlases using the true NRR-DSC (red dotted line), and predicted NRR-DSC (blue line). The X-axis shows number of atlases selected, and the Y-axis shows the ratio of k:all seg-DSC. Using the best 4 predicted atlases gives an improvement of 7% on an average. Figure IV-4 shows the comparison between using the best 4 atlases to perform the majority vote and using all of the atlases. It can be seen that a gain in performance is observed in segmentation of all structures except the optic nerve using the atlas selection procedure. An example of the segmentation result can be seen in Figure IV-5.

## 4. Conclusion

Assessing the structural correspondence error in non-rigid registrations is one of the most challenging problems in medical image-processing. In this paper, we expand on the previous work in AQUIRC to model error using registration circuits. In AQUIRC a multiplicative relationship was assumed between the error quality and circuit error. However, this relationship is not predictive of DSC error. We used a Taylor series expansion of up to third order to test other non-linear relationships between edge and circuit DSC. We found that the QUADRATIC error model showed the best prediction of NRR-DSC.

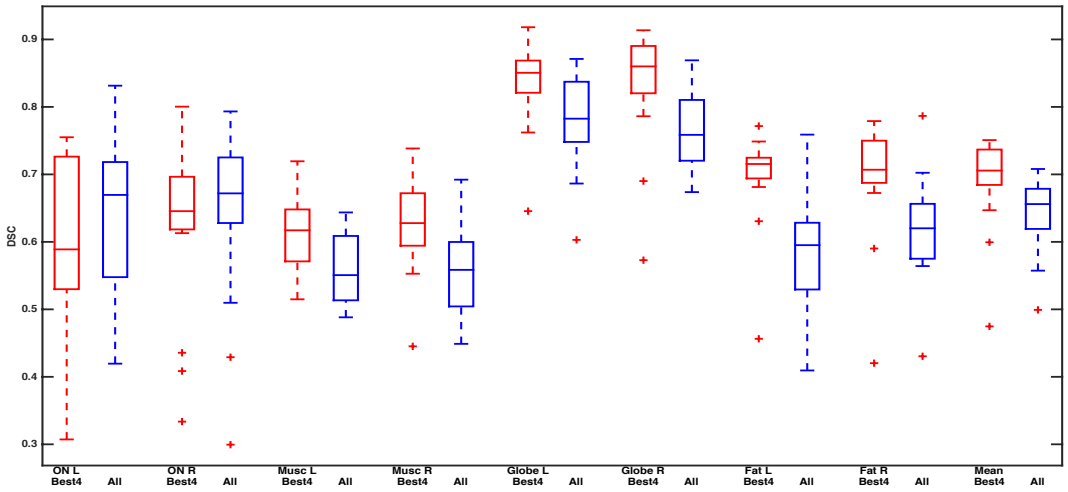



**Figure IV-4. Comparison between seg-DSC for each structure using atlas selection with k=4 (shown in red) and no atlas selection (shown in blue). The seg-DSC values are calculated for the 16 scans using a leave-one-out strategy.**

In the past, atlas-selection was performed by selecting atlases based on image similarity metrics[158], manifolds computed on image similarity[159], [160], or using meta-data such as age[161]. However, neither method is a predictor of true local structural correspondence, which affects the quality of multi-atlas segmentation. In order to show that prediction of NRR-DSC using QUADRATIC can improve multi-atlas segmentation, we performed the simplest case of a majority vote based on the best overall predicted DSC. On an average, we observe a 7% improvement in the segmentation DSC using the atlases predicted as best from the quadratic error model. From Figure IV-3C, it can be seen that when there is a perfect prediction of the NRR-DSC, the improvement in segmentation-DSC can be up to 16%. This improvement can be achieved in future work through better optimization procedures to solve the non-linear system of equations to prevent local minimas. Additionally, the NRR-DSC value is predicted for each structure of the atlas, which could lead to innovative multi-atlas segmentation techniques where atlas selection can be applied

to each structure to improve overall segmentation. The ability to have a quantitative measure of non-rigid registration can lead to several improvements in multi-atlas segmentation algorithms.

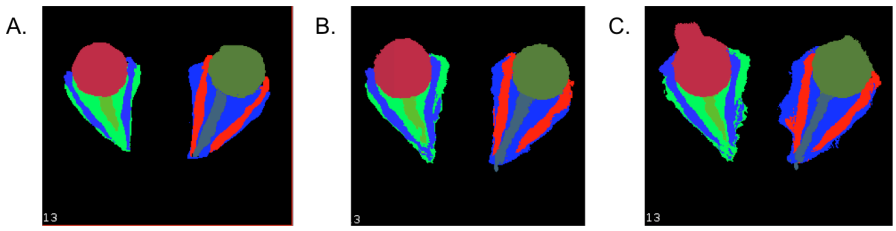



**Figure IV-5. A. Shows the true labels of a scan. B. Shows labels segmented using the atlas selection (k=4) based on the QUADRATIC error model. C. Shows the labels segmented using all 15 atlases, without atlas selection.**

## Chapter V. Imaging Biomarkers in Thyroid Eye Disease and their Clinical Associations

*Parts of this chapter have been published in Journal of Medical Imaging 5.4 (2018): 044001.*

### 1. Introduction

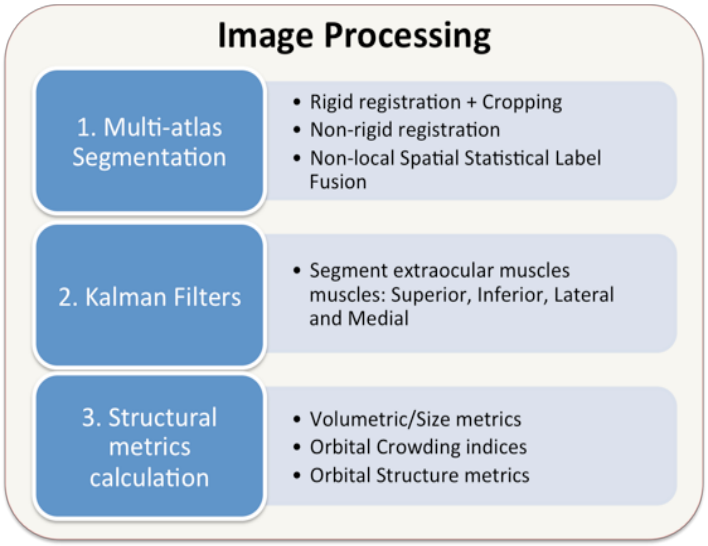



**Figure V-1. Image processing flow chart.**

Thyroid Eye Disease or Grave's ophthalmopathy is a condition that affects muscles and other soft tissues in the eye orbits. It is most often associated with Grave's disease, an autoimmune disorder causing hyperthyroidism. Computed tomography (CT) imaging of the orbit is standard

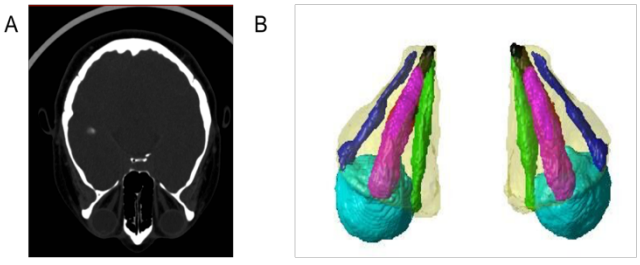



**Figure V-2. A. CT imaging of the head showing full view of the eye orbit. B. Result of segmentation showing: eye globes, optic nerve, orbital fat, superior rectus muscle, inferior rectus muscle, lateral**

clinical practice in the diagnosis and management of thyroid eye disease[65]. CT imaging shows distinct structures in the eye orbit such as globes, the optic nerve, extra ocular rectus muscles, and orbital fat. Previous studies have shown that volumetric and structural measurements of these structures are strongly associated with thyroid eye disease characteristics and risk for optic nerve pathologies[63], [64], [126], [127], [130], [132], [135], [136]. Studies have also described subtypes of thyroid eye disease based on clinical and orbital structural characteristics[162], [163]. However, these studies rely on manual measurements on orbital structures and subjective clinical criteria for description on subtypes[164], [165]. In this study, we develop an image-processing pipeline to automatically segment these structures, calculate structural metrics, perform PCA to identify dominant phenotypes or subtypes, and evaluate their associations with clinical characteristics.

**Table V-1. Clinical Characteristics.**

| Clinical measure | Description (assessment metrics) |
|---|---|
| Logmar visual acuity | Visual acuity using logmar chart (-0.3 – 1) |
| Color Vision | Ability to distinguish color. (0-20) |
| Best corrected visual acuity | Visual acuity with best possible lens correction denominator (20/10 – 20/200) |
| Visual Acuity Score | AMA defined visual acuity score (0-100) |
| Functional Acuity Score | Weighted visual acuity score (0-100) |
| Visual Field Score | AMA defined visual field perception score (0-100) |
| Functional Field Score | Weighted visual field score (0-100) |
| Visual Function Score | Percentage of ability when compared to healthy control (0-100) |
| Hertel Measurement | Measurement of proptosis using a Hertel exophthalmometer (in mm) |
| Ocular Motility Defects | Presence of difficulties in eye movement (0-absent/1-present) |

| | |
|---|---|
| Optic Neuropathy | Presence of optic neuropathy (0-absent/1-present) |
| Smoking History | History of smoking (0-non-smoker ,1-former smoker, 2-current smoker) |

## 2. Methods

Eighty-five patients were retrospectively selected from Vanderbilt University Medical Center who had orbital CT imaging performed before any surgical intervention for thyroid eye disease, after exclusion criteria for imaging issues such as severe artifact and low resolution. Institutional Review Board approval was obtained at Vanderbilt University. CT Imaging and clinical information including visual disability testing and demographic information was recorded.

### 2.1. CT Data

The selected patients underwent CT imaging of the eye orbits as part of the regular clinical care. For each of the eighty-five selected patients the highest resolution scan pre-decompression surgery was manually selected. All of the selected scans had the complete view of orbital structures of interest bilaterally: eye globe, optic nerve, extraocular rectus muscles, and orbital fat.

### 2.2. Clinical Data

Clinically pertinent information was recorded for each patient such as age, sex and visual disability testing. Eight measures of visual disability were recorded: logmar vision, color vision, best corrected visual acuity, AMA defined visual acuity score, visual field score, functional acuity score, functional field score, and functional vision score[92]. Clinical measures indicating the severity of the disease were collected such as Hertel measurements which describe the amount of proptosis[166], [167], ocular motility defects indicating difficulty in eye movement due to enlarged muscles[63], and presence of optic neuropathy, which is indicative of optic nerve

involvement[168], [169]. In addition, the smoking history of each patient was recorded as it is known to be highly associated with thyroid eye disease[163], [170]. Table V-1 presents the list of clinical characteristics used for the analysis. The clinical data was available at the time of the CT scan for all but 3 patients. The closest available clinical notes within 6 months were referred for these three patients.

2.3. Image Processing

Figure V-1 shows the image-processing pipeline. First, a three-step multi-atlas segmentation algorithm was employed to identify the orbital structures of interest. In the multi-atlas paradigm a set of expertly labeled example CT atlases are used as training examples to obtain the segmentation from a new target CT atlas[138]. The first step involves rigidly registering each of the example atlases to the target image, and propagating the corresponding labels to the target space. The sum of the labels is used a probability map which gives the approximate location of the eye orbit. A padded region of voxels having greater than 0.5 probabilities is selected for cropping. The second step involves using non-rigid registration to register the cropped example atlases to the cropped target image obtained from step 1. The corresponding labels of the example atlases are propagated to the target space using the non-rigid deformations. Step 3 involves fusing the transformed labels in step 2, using non-local statistical label fusion[144], [171] to obtain the final segmentation of the eye globe, the optic nerve, the extraocular muscles, and the orbital fat. Kalman filters were used to segment muscle labels obtained from the multi-atlas algorithm into individual extraocular muscles: the superior rectus, the inferior rectus, the medial rectus, and the lateral rectus muscle[156].

Once the final segmentation is obtained, the image-processing pipeline computes twenty-five structural metrics bilaterally, describing the sizes and structural arrangements of the eye orbit. The complete list of metrics is shown in Table V-2. For each structure, the volume, cross-sectional area, and diameter/length are measured[53], [55], [62], [64], [146]–[149]. Indices of orbital crowding, i.e., Barrett's muscle index[135] and volumetric crowding index[126] are computed. In addition, degree of proptosis[172] and orbital angle are computed[136].

### 2.4. Statistical Analysis

To identify structural phenotypes, a principal component analysis was performed over the twenty-five z-scored structural metrics for each eye. The top two components corresponding to the two major TED subtypes were crossed with the twelve clinical characteristics using a Kendall rank correlation test.

A rank correlation was used because several associations can be non-linear, however the relative change in clinical measures with changes in structural metrics is important to note. The significance of the correlation is computed using a permutation distribution of the correlation coefficient, which has the expected value of 0 under null hypothesis.

## 3. Results

The image processing pipeline produced segmentations of orbital structures in the eye as shown in Figure V-2: the eye globe, the optic nerve, the superior rectus muscle, the inferior rectus muscle, the lateral rectus muscle, and the medial rectus muscle, and the orbital fat. Twenty-five structural metrics are computed from these structures as seen in Table V-2.

3.1. PCA of Structural Metrics: TED Subtypes

The twenty-five structural metrics obtained from CT imaging are used to perform a principal component analysis. Table V-3 shows the factor loadings of the metrics for top twelve components that explain 97% of the variance. The top two components, which explain over 60% of the total variance, show two distinct clusters of metrics. Component 1 shows characteristics

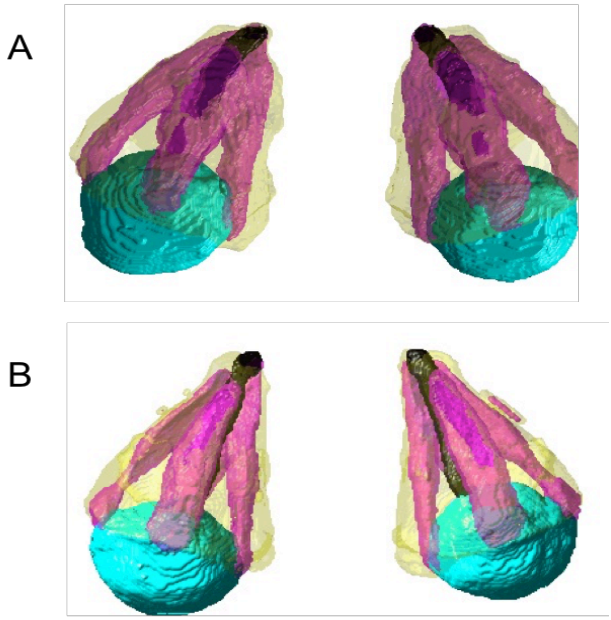



**Figure V-3. Subtypes of thyroid eye disease. A. Example of subtype 1 showing big muscles - with high values along component 1. B. Example of subtype 2 showing longer and thinner optic nerve, and greater orbital fat – with high values along component 2.**

with large size and volume measurements of the extraocular muscles. Component 2 shows longer and thinner optic nerves with increased orbital fat (volumetric crowding index is lower). These two clusters of structural metrics align with Nunery's classification of TED subtypes[162], [163]. Nunery describes a subtype of TED patients with motility defects and enlarged extraocular muscles (component 1 or Type 1) and a subtype without motility defects and increased orbital fat

(component 2 or Type 2). Examples of the two subtypes of segmented labels are shown in Figure V-3.

3.2. Correlation Subtypes of TED and Clinical Characteristics

The top two components corresponding to the two major subtypes of TED are crossed with the twelve clinical characteristics, as shown in Table V-4. Component 1, with enlarged extraocular muscles, is highly correlated with ocular motility ($p<0.05$) as postulated by Nunery. It is also associated it vision loss (logmar visual acuity and AMA visual acuity score). Component 2, with longer and thinner optic nerve and higher orbital fat, does not correlate with ocular motility issues also aligning with Nunery's description. Component 2 is correlated ($p<0.05$) with Hertel measurements indicating that this group has more proptosis. Component 2 is negatively correlated with vision loss, indicating relatively lower effects on vision compared to component 1. A positive smoking history is associated with component 2.

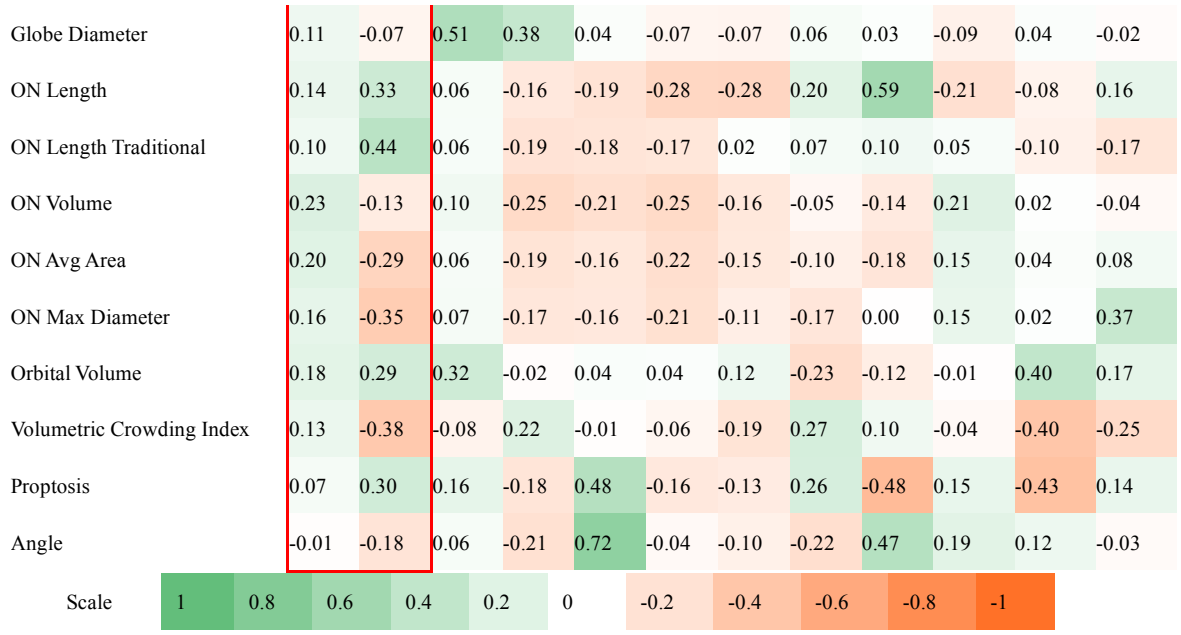
**Table V-2. Structural Metrics.**

| Metric # | Structural Metric |
|---|---|
| *Size/Volumetric measurements* | |
| 1-5 | Superior, Inferior, Lateral, Medial rectus, and Total muscle volumes |
| 6-9 | Superior, Inferior, Lateral, Medial rectus muscle maximum diameter |
| 10-13 | Superior, Inferior, Lateral, Medial rectus muscle average diameter |
| 14 | Globe volume |
| 15 | Globe diameter |
| 16 | Optic nerve volume |
| 17-18 | Optic nerve length and traditional length |
| 19 | Optic nerve average area |

| | |
|---|---|
| 20 | Optic nerve maximum diameter |
| 21 | Orbital volume |
| Crowding indices | |
| 22 | Barrett's index |
| 23 | Volumetric crowding index |
| Orbital structure metrics | |
| 24 | Degree of proptosis |
| 25 | Orbital Angle |

**Table V-3. Factor loadings of 12 principal components of the structural metrics explaining 97.46% of the variance. The top two principal components are enclosed in a red box. Positive loadings are green, and negative loadings are red.**

| | *Comp 1* | *Comp 2* | *Comp 3* | *Comp 4* | *Comp 5* | *Comp 6* | *Comp 7* | *Comp 8* | *Comp 9* | *Comp 10* | *Comp 11* | *Comp 12* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Explained Variance* | 46.3% | 13.0% | 10.1% | 6.5% | 4.7% | 4.0% | 3.7% | 3.0% | 2.5% | 1.4% | 1.3% | 0.9% |
| Superior Muscle Volume | 0.24 | -0.02 | 0.13 | -0.24 | -0.08 | 0.39 | 0.02 | 0.11 | -0.01 | 0.06 | 0.14 | -0.11 |
| Inferior Muscle Volume | 0.25 | 0.14 | -0.02 | 0.13 | -0.04 | 0.03 | 0.05 | -0.16 | 0.11 | 0.34 | -0.09 | -0.24 |
| Lateral Muscle Volume | 0.26 | 0.08 | -0.20 | 0.07 | 0.13 | -0.04 | -0.23 | -0.15 | -0.14 | -0.29 | 0.07 | -0.15 |
| Medial Muscle Volume | 0.26 | -0.05 | -0.04 | -0.08 | 0.08 | -0.18 | 0.35 | 0.14 | -0.03 | -0.03 | 0.11 | -0.23 |
| Superior Muscle DiameterAvg | 0.23 | -0.07 | 0.08 | -0.22 | -0.03 | 0.42 | -0.09 | 0.09 | -0.05 | -0.08 | -0.05 | -0.15 |
| Inferior Muscle DiameterAvg | 0.24 | 0.08 | -0.07 | 0.22 | -0.07 | 0.09 | 0.11 | -0.27 | 0.12 | 0.31 | -0.22 | -0.08 |
| Lateral Muscle DiameterAvg | 0.24 | 0.01 | -0.23 | 0.09 | 0.12 | -0.05 | -0.25 | -0.13 | -0.11 | -0.26 | 0.09 | -0.15 |
| Medial Muscle DiameterAvg | 0.23 | -0.10 | -0.04 | -0.09 | 0.12 | -0.24 | 0.38 | 0.11 | 0.09 | -0.11 | 0.04 | -0.17 |
| Superior Muscle DiameterMax | 0.21 | -0.10 | 0.14 | -0.20 | 0.01 | 0.47 | 0.02 | 0.04 | 0.15 | -0.20 | -0.27 | 0.28 |
| Inferior Muscle DiameterMax | 0.20 | 0.13 | -0.15 | 0.29 | -0.01 | 0.07 | 0.16 | -0.19 | 0.08 | 0.22 | -0.29 | 0.36 |
| Lateral Muscle DiameterMax | 0.22 | 0.06 | -0.25 | 0.14 | 0.09 | -0.02 | -0.19 | -0.20 | -0.10 | -0.31 | 0.05 | 0.26 |
| Medial Muscle DiameterMax | 0.22 | -0.13 | -0.07 | 0.02 | 0.07 | -0.13 | 0.50 | 0.22 | 0.01 | -0.23 | -0.02 | 0.32 |
| Total Muscle Volume | 0.29 | 0.05 | -0.05 | -0.02 | 0.03 | 0.06 | 0.02 | -0.03 | -0.03 | 0.00 | 0.07 | -0.20 |
| Barrett Index | 0.17 | 0.03 | -0.25 | 0.23 | 0.07 | 0.12 | -0.21 | 0.59 | 0.04 | 0.39 | 0.44 | 0.21 |
| Globe Volume | 0.11 | -0.06 | 0.52 | 0.39 | 0.04 | -0.05 | -0.07 | 0.05 | 0.01 | -0.09 | 0.04 | -0.05 |




## 4. Discussion

Thyroid eye disease is a serious condition that can lead to temporary or permanent loss of vision in some patients[133], [173]. Often, the clinical manifestations or characteristics of the disease can be confounding and variable in presentation. CT imaging is often used clinically, to make diagnostic and prognostic evaluations. An objective and quantitative assessment of the CT imaging and their related clinical characteristics can improve diagnosis and intervention.

In this study, we developed a novel method to automatically segment CT images of the eye orbit, extract structural descriptive metrics, and structural subtypes. The image-processing pipeline establishes a new quantitative method to identify characteristics of thyroid eye disease through clinically acquired CT imaging.

**Table V-4. Kendall rank correlation between clinical characteristics and the top two principal components representing the two phenotypes of thyroid of disease. Component 1 represents bigger muscle phenotype. Component 2 represents longer and thinner optic nerve.**

| Clinical Characteristics | Comp 1 | Comp 2 |
|---|---|---|
| Logmar visual acuity | 0.14 | -0.08 |
| Color vision | -0.07 | 0.05 |
| Best corrected visual acuity | 0.13 | -0.23 |
| Visual acuity score | -0.16 | 0.15 |
| Functional acuity score | -0.15 | 0.16 |
| Visual field score | -0.04 | 0.16 |
| Functional field score | -0.04 | 0.14 |
| Functional vision score | -0.10 | 0.18 |
| Hertel measurement | -0.01 | 0.35 |
| Motility Defect | 0.17 | 0.07 |
| Optic neuropathy | -0.03 | -0.07 |
| Smoking | 0.08 | 0.16 |

Moreover, we develop a method to identify the two important subtypes of thyroid eye disease from CT imaging. Though these subtypes have been described in terms of their clinical manifestations and through manual measurements of CT imaging, our study is the first to show the presence of these subtypes through a quantitative procedure. These subtypes were crossed with clinical characteristics to objectively establish associations observed in clinical literature such as Type 1 is associated with ocular motility, while Type 2 is not[63], [131], [137]; Type 1 is associated with more severe vision loss than Type 2[174]; Type 2 is associated with increased proptosis[163]; and Type 2 is associated with smoking[175].

## Chapter VI. EMR-Radiological Phenotypes in Diseases of the Optic Nerve and their Association with Visual Function

*Parts of this chapter have been published in Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 373-381. Springer, Cham, 2017.*

### 1. Introduction

Pathologies of the optic nerve affect millions of Americans each year and can severely affect an individual's quality of life due to loss of visual function[176]. Accurate characterization of these diseases and timely intervention can preserve visual function. 3D computed tomography (CT) imaging of the eye orbit can captures structural changes in the eye orbit, which indicate the extent of disease progression and characterizes pathology. In prior studies[156], [177], a quantitative relationship between 3D structural metrics of the eye orbit was shown to be associated with visual outcomes such as visual acuity and field vision in patients with optic nerve disorders. However, the percentage of explained variance due to structural data was low ($R^2 \sim 0.1$-$0.2$). Several factors influence a model's ability to explain outcomes, particularly the selection of predictive features. Also, while information is available in radiological imaging, evaluation of radiology within the context of an individual's health history is important in determining functional changes, progression of disease, and prognosis. With the rise in adoption of digital electronic medical record (EMR) systems in the US health care system[178], [179], these records are available to medical research scientists with increasing ease.

In this study we develop an automated pipeline for segmentation and metric calculation of CT eye orbits for glaucoma and thyroid eye disease (TED). Further, we show that integrating EMR data, such as ICD-9 (International Classification of Diseases – 9) codes, and CPT (Current

Procedural Terminology) codes, with imaging biomarkers improves the explained variance of disease outcomes.

## 2. Methods

### 2.1. Data

The study was conducted on a retrospective cohort of patients at Vanderbilt University Medical Center. Subjects were retrieved under Institutional Review Board (IRB) approval based on both having met clinical criteria for eye disease and undergoing CT imaging as part of their regular clinical care. The data collected include imaging records, visual testing, demographic data, complete ICD-9 codes and CPT codes. The disease groups included in this study are glaucoma (n=55) and TED (n=32).

### 2.2. Outcomes: Visual Function Scores

The outcomes in this study were calculated based on clinical visual acuity and visual field testing. Nine different outcome measures are calculated for a complete visual function evaluation as defined by the American Medical Association[92]. Right and left visual acuity scores are calculated as $VAS_{od}$ and $VAS_{os}$ respectively. The visual acuity for both eyes, $VAS_{ou}$ is calculated as the best of $VAS_{od}$ and $VAS_{os}$ . The functional acuity score, FAS is a weighted score of $VAS_{od}$, $VAS_{os}$, and $VAS_{ou}$ with weights 1:1:3. The scores from visual field testing, $VFS_{od}$, $VFS_{os}$, $VFS_{ou}$, and FFS are calculated similarly. A final score of visual function called functional visual score (FVS), is calculated as the average of FAS and FFS.

2.3. Image processing

Figure VI-1 shows the image segmentation pipeline. First, multi-atlas segmentation was employed to identify four labels: the globe, the optic nerve, the extraocular muscles and the periorbital fat. A set of twenty-five expertly labeled example 3D CT atlases is used as training examples to obtain the segmentation from a new input scan. Each of the example atlases is non-rigidly registered to the cropped input image space[157]. The corresponding labels of the example atlases are propagated to the input image space using the non-rigid deformations. Next, non-local statistical label fusion is used to obtain a segmented result with the four labels [144]. Segmenting the individual extraocular rectus muscles is challenging in diseased eyes, since obtaining true labels is difficult at the back of the orbit due to inflammation. So, we employ Kalman filters to segment muscle labels obtained from the multi-atlas algorithm[156] to identify the superior rectus muscle, the inferior rectus muscle, the lateral rectus muscle, the medial rectus muscle. Once the final segmentation is obtained twenty-five structural metrics are computed bilaterally[177]. For each structure, the volume, cross-sectional area, and diameter/length are measured. Indices of orbital crowding, i.e., Barrett's muscle index and volumetric crowding index are computed. In addition, degree of proptosis and orbital angle are computed. For each patient, $i$, a vector with 50 elements is constructed for 25 structural metrics computed bilaterally,

$$\boldsymbol{x}_{CT}^{\{i\}} = [sm_{1\_\mathrm{os}}\ sm_{2\_\mathrm{os}} \dots sm_{25\_\mathrm{os}}\ sm_{1\_\mathrm{od}}\ sm_{2\_\mathrm{od}} \dots sm_{25\_\mathrm{od}}]$$

where, $sm_{k\_os}$ indicates $k^{th}$ structural metric of the left eye and $sm_{k\_od}$ indicates $k^{th}$ structural metric of the left eye.

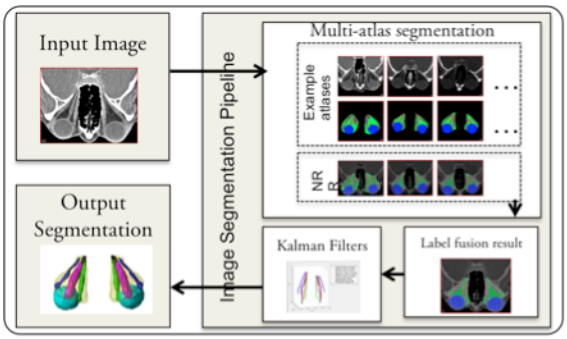



**Figure VI-1. Overview of image segmentation. Multi-atlas label fusion is used to segment the optic nerve, globe, muscle, and orbital fat. Kalman filters are used to segment the four individual extraocular muscles based on the result to achieve the final 3D segmentation result.**

From the EMR, complete ICD-9 codes and CPT codes were extracted for diagnostic and treatment information for each patient. However, only the ICD-9 and CPT codes available one month or more before the diagnosis are considered, since we are interested in understanding how a patient's history provides a context for imaging information.

*PheWAS codes*. There are over 14,000 ICD-9 codes defined. A hierarchical system was defined that maps each ICD-9 code to a smaller group of 1865 phenotype codes originally used in phenome-wide association studies (PheWAS) [180]. Each phenotype, called a PheWAS code, indicates a related group of medical diagnoses and conditions.

*ProWAS codes*. We introduce a similar hierarchical grouping to map each CPT code to a group of related procedures, which we indicate by a procedure wide association study (ProWAS) code. We define 1682 ProWAS codes, which are finer granularity subgroups of the Clinical Classification Systems coding provided by the Healthcare Cost and Utilization Project (HCUP) Agency for Healthcare Research and Quality [181].

For each patient, $i$, a binary vector with 1865 elements, $\boldsymbol{x}_{PheWAS}^{\{i\}}$ is defined,

$$\boldsymbol{x}_{PheWAS}^{\{i\}} = [d_1\ d_2\ \dots\ d_{1865}\ ]$$

where, $d_k$ is 1 if the patient $i$ has had the diagnosis phenotype $d_k$ in the past and 0 otherwise. Similarly, a binary vector, $\boldsymbol{x}_{ProWAS}^{\{i\}}$ is defined with 1682 elements,

$$\boldsymbol{x}_{ProWAS}^{\{i\}} = [t_1\ t_2\ \dots\ t_{1682}\ ]$$

where, $t_k$ is 1 if the patient $i$ has had the treatment phenotype $t_k$ in the past and 0 otherwise.

2.4. Dimensionality Reduction: PCA and MCA

A large amount of data is available for each patient; the final data vector for a patient $i$ has 3597 elements in it. However, the data are correlated with each other, and it is possible to find underlying principal variables in the data. For the structural metrics, a principal component analysis (PCA) [182] is performed to reduce the dimensionality of the dataset. The first five principal components explaining about three fourths of the variance are extracted to give, for subject $i$,

$$\boldsymbol{x}_{CT\_pca}^{\{i\}} = [\ sm_1'\ \ sm_2'\ \ \dots\ \ sm_5'] \dots\dots (1)$$

For the PheWAS and ProWAS binary vectors, multiple correspondence analysis (MCA) [183] is used to extract orthogonal components that are decomposed using the $\chi^2$-statistic. The first five components are considered for both PheWAS and ProWAS vectors. As a result of MCA, we get two vectors of smaller dimensionality for each patient,

$$\boldsymbol{x}_{PheWAS\_mca}^{\{i\}} = [d_1'\ \ d_2'\ \dots\ \ \mathrm{d}_5'] \dots\dots\dots\dots (2)$$

$$\boldsymbol{x}_{ProWAS\_mca}^{\{i\}} = [t_1'\ \ t_2' \dots\ \mathrm{t}_5'] \dots\dots\dots\dots\dots (3)$$

2.5. Stepwise Generalized Linear Model

The visual acuity scores are between 0 and 100 with most patients having scores close to 100 and values closer to 0 being extremely rare. This makes the distribution of the visual outcomes left skewed. Therefore, a generalized regression model (GLM) with a Poisson distribution [184] is used to find the explanatory value of each set of datasets, given by equations (1), (2) and (3), and all the data together. These datasets are regressed over the visual outcome scores $\boldsymbol{s}_v$, where $s_v \in \{VAS_{ou}, VAS_{od}, VAS_{os}, VAS, FAS, VFS_{ou}, VFS_{od}, VFS_{os}, FFS, FVS\}$. Four models are defined for each $v$,

$$M1\text{: } s_v = \beta_0 + \beta_1 sm'_1 + \cdots + \beta_k sm'_k + \beta_{k+1} d'_k + \cdots + \beta_{k+l} d'_l + \beta_{k+l+1} t'_1 + \cdots + \beta_{k+l+m} t'_m + \beta_{k+l+m+1} age + \beta_{k+l+m+2} sex + \epsilon$$

$$M2\text{: } s_v = \beta_0 + \beta_1 sm'_1 + \cdots + \beta_k sm'_k + \beta_{k+1} age + \beta_{k+2} sex + \epsilon$$

$$M3\text{: } s_v = \beta_0 + \beta_1 d'_1 + \cdots + \beta_l d'_l + \beta_{l+1} age + \beta_{l+2} sex + \epsilon$$

$$M4\text{: } s_v = \beta_0 + \beta_1 t'_1 + \cdots + \beta_m t'_m + \beta_{m+1} age + \beta_{m+2} sex + \epsilon$$

The four models are built using stepwise regression[185], with forward selection of variables. At each step, the variable that most significantly improves the model deviance is added until there is no more improvement. The explained variance of each model, $R^2$ is noted.

2.6. Test of Deviance

The deviance of a model $M$, with fitted parameters $\hat{\theta}$ is given by,

$$D(M) = -2\left(\log\left(p(y|\hat{\theta})\right) - \log(p(y|\theta_s))\right)$$

where, $\theta_s$ are the parameters of the saturated model, i.e., a model with parameters for each data point such that it is fitted exactly. The deviance can be used to test significance between two nested models $M_p(\hat{\theta}_p|\mathrm{X})$ and $M_q(\hat{\theta}_q|\mathrm{X})$, where $\hat{\theta}_p \subset \hat{\theta}_q$ and the difference in the parameters between the two models is given by $\delta$. The difference of the deviance between the two models

follows a $\chi^2$ −squared distribution with degree of freedom $\delta$. The null hypothesis, $H_0$ for the test of deviance is that adding $\delta$ parameters to model $M_p$ to get $M_q$ does not improve the model. This test is used to compare models M2-4 with M1.

## 3. Results

The average age group for glaucoma cohort is 65.4±19.5 years and 72% of the subjects were female. 91% of TED subjects were female, and the average age for this group is 57.8±16.2 years. On an average, each patient had 410 ICD-9 codes, and 660 CPT codes recorded. Figure VI-2 shows the individual distribution by sex along the first two components of the three datasets in models M2, M3, and M4. For glaucoma, the first component of the PCA on structural metrics corresponded to muscle and optic nerve measurements, and the second component corresponded to orbital and globe measurements. For TED, the first component corresponded to mostly muscle measurements, and the second component corresponded to measurements of the optic nerve. Some of the conditions associated with the first MCA component of the ICD-9 vector for Glaucoma are malaise, osteoarthrosis, and hypovolemia, and conditions associated with the second component included female genitourinary symptoms and symptoms associated with the eye such as pain and swelling. The first MCA component for TED's ICD-9 vector was associated with conditions including hyperlipidemia, diabetes, and circulatory problems, and some of the conditions most associated with the second MCA component were myalgia and abnormal blood chemistry. For the CPT vector for glaucoma, the first dimension was associated with a wide range of procedures such as CT scans, and pathology labs, and the second component was associated with cardiac testing. For TED, the first component was associated with procedures such as urinalysis and blood work, the second component was associated with physical therapy related procedures.

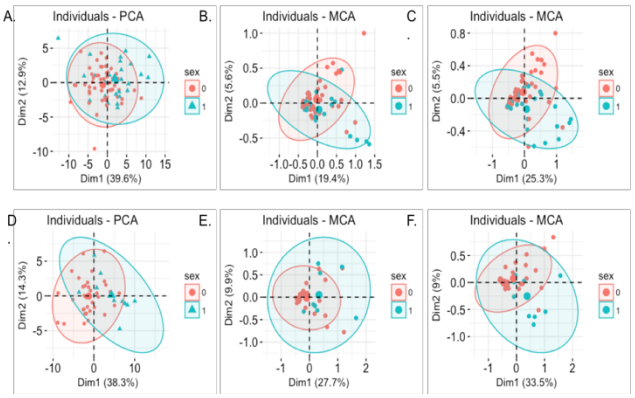



**Figure VI-2. Distribution of individuals by sex along the first two components from equations (1), (2), and (3). Red and blue indicate 95% confidence ellipses for females and males respectively. (A) $x_{CT\_pca}$ for glaucoma. (B) $x_{PheWAS\_mca}$ for glaucoma. (C) $x_{ProWAS\_mca}$ for glaucoma. (D) $x_{CT\_pca}$ for TED. (E) $x_{PheWAS\_mca}$ for TED. (F) $x_{ProWAS\_mca}$ for TED.**

Tables VI-1 and VI-3 show the $R^2$ values of models M1, M2, M3, and M4 regressed over functional visual scores for glaucoma and TED respectively. The behavior of the models is the same for both the diseases. Addition of treatment and diagnostic phenotypes to model M2 to get model M1 results in significant improvement of explained variance in most of the visual outcomes: FVS, FFS, $VFS_{ou}$, $VFS_{od}$, $VFS_{os}$, $VAS_{od}$ and $VAS_{os}$. The $R^2$ values that improve between model M2 to M1 are indicated by ** in Tables VI-1 and VI-3. The statistical significance of this improvement is tested using the test of deviance as described in section 2.7. Tables VI-2 and VI-4 show the p-values of the tests of deviance performed between M1 and its nested models M2, M3, and M4.

However, it is interesting to note that composite visual acuity scores $VAS_{ou}$ and FAS do not show an improvement between models M2 and M1, even though the right and left acuity scores $VAS_{od}$ and $VAS_{os}$ do. Note from the definition of these scores that they weight the best performing eye higher. This might indicate that changes in visual acuity might not be bilateral in these conditions. Whereas, for visual field scores the behavior of the individual eye scores is reflected

in the composite scores, indicating that visual field changes might be bilateral in glaucoma and TED.

**Table VI-1. Explained variance in Glaucoma. ** indicates that model M1 is significantly better than model M2, i.e. using structural metrics alone, based on the p-values in Table VI-2.**

| $R^2$ | $VAS_{od}$ | $VAS_{os}$ | $VFS_{od}$ | $VFS_{od}$ | $VAS_{ou}$ | $VFS_{ou}$ | FFS | FVS | FAS |
|---|---|---|---|---|---|---|---|---|---|
| M1 | 0.48** | 0.33** | 0.47** | 0.39** | 0.24 | 0.27** | 0.33** | 0.40** | 0.35 |
| M2 | 0.45 | 0.13 | 0.37 | 0.16 | 0.24 | 0.18 | 0.23 | 0.33 | 0.35 |
| M3 | 0.07 | 0.07 | 0.16 | 0.08 | 0.08 | 0.05 | 0.08 | 0.07 | 0.00 |
| M4 | 0.04 | 0.00 | 0.20 | 0.06 | 0.00 | 0.05 | 0.09 | 0.06 | 0.00 |

**Table VI-2. Test of deviance. Na indicates that the model is the same in both M1 and M2, as the same features were selected.**

| p-value | $VAS_{od}$ | $VAS_{os}$ | $VFS_{od}$ | $VFS_{od}$ | $VAS_{ou}$ | $VFS_{ou}$ | FFS | FVS | FAS |
|---|---|---|---|---|---|---|---|---|---|
| M1 vs M2 | 4.30E-05 | 5.89E-09 | 2.92E-26 | 3.31E-38 | na | 5.38E-07 | 1.34E-07 | 1.18E-07 | na |
| M1 vs M3 | 1.66E-28 | 1.55E-10 | 3.17E-62 | 9.70E-46 | 0.001 | 2.65E-11 | 8.46E-15 | 2.39E-26 | 2.40E-06 |
| M1 vs M4 | 1.42E-29 | 3.59E-12 | 3.22E-58 | 8.32E-47 | 0.0005 | 8.48E-12 | 2.35E-14 | 3.15E-26 | 2.40E-06 |

**Table VI-3. Explained variance in TED. ** indicates that model M1 is significantly better than model M2, i.e. using structural metrics alone, based on the p-values in Table VI-4.**

| $R^2$ | $VAS_{od}$ | $VAS_{os}$ | $VFS_{od}$ | $VFS_{od}$ | $VAS_{ou}$ | $VFS_{ou}$ | FFS | FVS | FAS |
|---|---|---|---|---|---|---|---|---|---|
| M1 | 0.61** | 0.30** | 0.59** | 0.37** | 0.28 | 0.36** | 0.42** | 0.54** | 0.44 |
| M2 | 0.49 | 0.23 | 0.45 | 0.26 | 0.28 | 0.23 | 0.28 | 0.40 | 0.44 |
| M3 | 0.30 | 0.22 | 0.33 | 0.24 | 0.16 | 0.20 | 0.24 | 0.28 | 0.20 |
| M4 | 0.28 | 0.18 | 0.29 | 0.18 | 0.16 | 0.17 | 0.19 | 0.29 | 0.20 |

**Table VI-4. Test of deviance. Na indicates that the model is the same in both M1 and M2, as the same features were selected.**

| p-value | $VAS_{od}$ | $VAS_{os}$ | $VFS_{od}$ | $VFS_{od}$ | $VAS_{ou}$ | $VFS_{ou}$ | FFS | FVS | FAS |
|---|---|---|---|---|---|---|---|---|---|
| M1 vs M2 | 4.56E-06 | 0.00027 | 5.15E-24 | 6.75E-07 | na | 4.38E-08 | 7.50E-09 | 9.70E-10 | na |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| M1 vs M3 | 1.09E-17 | 0.00049 | 7.27E-45 | 1.23E-08 | na | 7.02E-10 | 1.88E-12 | 3.53E-18 | 1.12E-05 |
| M1 vs M4 | 1.60E-18 | 6.84E-05 | 4.61E-52 | 1.30E-10 | na | 9.51E-12 | 6.77E-15 | 2.36E-19 | 1.12E-05 |

## 4. Discussion

To identify imaging biomarkers associated with diseases of the optic nerve such as glaucoma and thyroid eye disease, their relationship with visual function scores must be established. This study shows that addition of treatment and diagnostic phenotypes derived through MCA on ProWAS and PheWAS data can improve traditional imaging biomarker studies by providing the context of an individual's health history from clinical data. This is the first known study with the application of ProWAS mapping to identify treatment phenotypes for eye disease. We show that structural metrics of the eye orbit derived from CT imaging, treatment, and diagnostic phenotypes show a significant association with visual function scores and explain about 40% - 60% of the variance for visual outcomes in glaucoma and thyroid eye disease.

# Chapter VII. Electronic Medical Record Context Signatures Improve Diagnostic Classification using Medical Image Computing

*Parts of this chapter have been published in in IEEE Journal of Biomedical and Health Informatics. doi: 10.1109/JBHI.2018.2890084*

## 1. Introduction

With the advent of digitization of medical records, extensive medical data are available to perform large-scale studies that were previously difficult to implement[1]. However, so-called "big data" analyses also present new challenges for consolidation of information and integration of newer methods with established practices in the medical image processing community.

For example, automatic detection and diagnosis systems have been developed for medical image processing [2], [3], brain tumor classification[4], detection of prostate lesions[5], risk assessment in traumatic brain injury[6], and detection of breast cancer[7], demonstrating that medical images contain quantitative diagnostic information. However, when physicians make diagnoses in practice, they integrate multi-modal patient information, including medical history, laboratory tests, medication exposures, etc. Because much of this information can be extracted computationally from the electronic medical record (EMR)[8]–[10], we propose using EMR-derived patient signatures alongside imaging data to improve the accuracy of image analysis.

Specifically, we build on the Phenome Wide Association Study (PheWAS) [11], [12] paradigm that identifies relationships between targeted genotypes and clinical phenotypes. The phenotypes are clinical diseases represented by codes derived from EMR billing codes such as the International Classification of Diseases, Ninth Revision (ICD-9) codes. ICD9 codes are hierarchical, with the three digits codes representing a major disease category, followed by a decimal or two to describe subtypes and specific symptoms. For example, Code 377 represents

**Table VII-1. Overview of the data used and analyses performed from the two studies.**

| Center | Study 1: VUMC | Study 2: BLSA |
|---|---|---|
| **Data** | | |
| Study population | • Individuals with disease codes for glaucoma, intrinsic optic nerve disease, thyroid eye disease, optic nerve edema.<br>• Individuals with hearing loss for controls. | • Individuals from the Baltimore Longitudinal Study of Aging (BLSA). |
| Dataset E | ICD-9 codes and demographic data for 29,214 individuals from the study population collected over all visits. | ICD-9 codes and demographic data for 1,715 individuals from the study population collected over all visits. |
| Dataset RE | Computed tomography (CT) imaging of the eye orbit for 1,451 subjects along with ICD-9 codes for all visits, and demographic data. | Magnetic resonance imaging (MRI) of the brain of 124 subjects along with ICD-9 codes for all visits, and demographic data. |
| **Analyses** | | |
| Case-control experiments | 1. Glaucoma vs. healthy control<br>2. Intrinsic optic nerve disease vs. healthy control<br>3. Thyroid eye disease vs. healthy control<br>4. Optic nerve edema vs. healthy control<br>5. Glaucoma vs. other<br>6. Intrinsic optic nerve disease vs. other<br>7. Thyroid eye disease vs. other | 1. Diabetes vs. control |

"Disorders of the optic nerve and visual pathway"; 377.0 represents "Papilledema", a specific disorder of the optic nerve pathway; and 377.01 represents "Papilledema with decreased ocular pressure". However, there are a few issues with the use of ICD9 codes directly for scientific research [13]. 1. The number of digits or the type of code used to describe a sign or symptom is inconsistent among doctors. 2. The hierarchical system here is designed such that there are multiple categories for common diseases and sometimes, just a single code for a complex disease—since

they're primarily designed for the purposes for measuring hospital utilization and billing[12]. To overcome these problems, Denny et al designed a conversion system from ICD9 codes to PheWAS codes such that the conceptual granularity of the codes is consistent across diseases with the help of medical experts[11]. PheWAS codes are widely used in genomic research and are generally considered an acceptable standard in medical conceptual categorization. In this work, we convert the ICD9 codes to PheWAS codes for EMR feature extraction. We developed a Python package that uses the same PheWAS-ICD-9 mappings to study the associations of these clinical phenotypes with a particular disease of interest [14]. We refer to this approach as a *phenome-disease association study* (PheDAS). With PheDAS, we isolate the phenotypes most associated with a disease group prior to the diagnosis of the disease. In other words, we develop an "EMR context signature" that consolidates a patient's history of diagnoses and symptoms into a binary vector. We add this signature to radiological image features and evaluate the improvement it provides in disease classification.

We illustrate the PheDAS approach using two studies. In the first study, we evaluate the role of EMR context signatures in improving classification accuracy of four disease groups affecting the optic nerve: glaucoma, intrinsic optic nerve disease (optic neuritis and other optic nerve disorders), thyroid eye disease (TED), and optic nerve edema (papilledema and idiopathic intracranial hypertension). These groups were chosen because the conditions can present with similar initial symptoms and co-morbidities, making detection challenging. However, early classification and intervention are often needed to preserve visual function. Xiuya et al and Chaganti et al have also shown that visual function for these subjects is correlated with radiological features derived from computed tomography (CT) and magnetic resonance (MR) imaging [15], [16]. Herein, we evaluate the role of EMR context signatures in improving classification accuracy

for each of the four disease groups. In the second study, we use EMR to extract the context signature of a well-defined disease, and assess the classification accuracy gained by integrating this with neuroimaging information. For this study, we examine diabetes in an aging population[17]. Symptomatic risk factors for diabetes are well understood, for example impaired glucose levels are highly predictive of future diabetes [18]. Recent studies have also shown that diabetes is mildly correlated with cortical and white matter signal changes and white matter hyper intensities in the brain[17], [19]. However, classification of diabetes based on imaging remains a harder problem, as it depends on various other issues such as variance in disease population, disease stage etc. We use EMR signatures and brain MRI data acquired in the aging study to investigate the classification accuracy for Diabetes.

## 2. Methods

### 2.1. Glossary of Terms

We use the following terms to disambiguate between the multiple levels of our study design:

*Study*: An observational study of a group of individuals sharing similar type of data collected at the same center. Each study has one or more case-control experiments designed to compare subjects who have a disease and those who do not.

*Disease group:* A clinical classification of individuals having a health condition.

*Control group:* Individuals in a study that are not classified as belonging to a disease group are considered as the control group.

*Case-control experiment:* An experiment within a study to compare a disease group with a control group.

*Dataset:* A type of data collected within a study. Dataset E has a set of individuals with only EMR data, while dataset RE has a different set of individuals with both EMR and radiological imaging data.

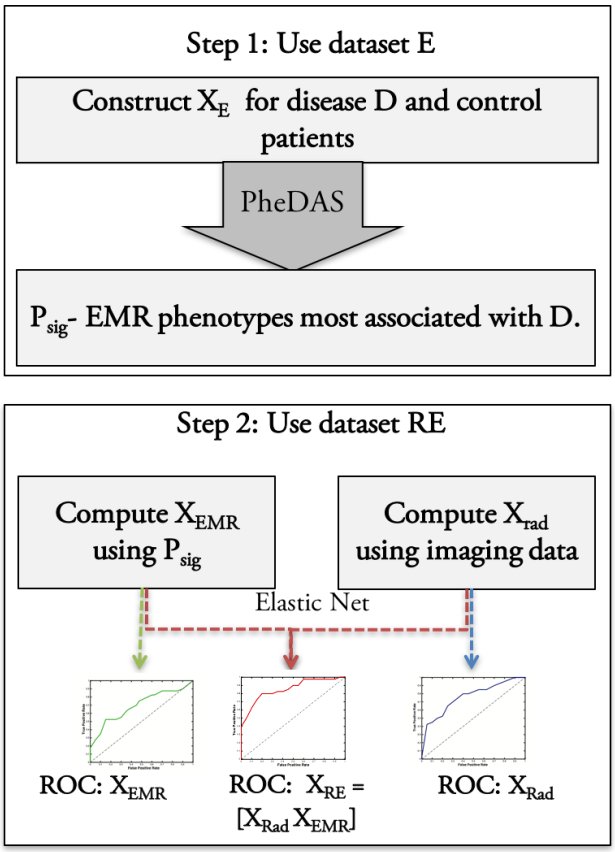



**Figure VII-1. Overview for each case control experiment. In step 1, the EMR phenotypes associated with disease D, $P_{sig}$ are learnt from dataset E. In Step 2, EMR context signature vectors, $X_{EMR}$ are calculated from $P_{sig}$ and $X_{Rad}$ is calculated from imaging. $X_{RE}$, $X_{Rad}$, and $X_{EMR}$ are used to train an elastic net classifier and ROC curves are computed.**

*Class*: An engineering term for group (disease or control). An individual's class label is a categorical indicator of which class the patient belongs to. The algorithms presented fall under *classification approaches* in which a procedure is created to infer a class label for each subject.

*ICD-9 code*: A code used to describe a diagnosis or a health problem as defined by the International Classification of Diseases version 9 (ICD-9). There are roughly ~15,000 ICD-9 codes defined.

*PheWAS code*: A code based on hierarchical categorization of ICD-9 codes that describes a diagnostic "phenotype" by grouping a set of related ICD-9 codes. The ~15,000 ICD-9 codes are mapped to 1865 PheWAS codes [20].

*PheDAS:* A phenome-disease association study, which is used to identify all the diagnostic phenotypes associated with a disease group.

*EMR context signature:* A binary vector describing an individual's history that is calculated using PheDAS.

2.2. Overview

We use data from two studies to evaluate the PheDAS framework (Table I). The first study, collected at Vanderbilt University Medical Center (VUMC), characterizes diseases of the optic nerve within four major disease groups: glaucoma, intrinsic optic nerve disease, thyroid eye disease and optic nerve edema. These diseases have a complex presentation, similar symptoms, and frequently co-occur. The control group for this study is a group of individuals with hearing loss who were chosen for having similar imaging data available. In the first study, we build classifiers to classify disease groups from control group. Additionally, we build a one v. all to classify each optic nerve disease group from the other optic nerve diseases.

The second study uses data from the Baltimore Longitudinal Study of Aging (BLSA)[21], a study that collects longitudinal data of an aging population in order to examine changes in the brain as a person ages. In this study, our "treatment" group are individuals who were diagnosed

with diabetes. The control group for this study is individuals in BLSA who had no diagnosis of diabetes. Since risk factors for diabetes are well understood, it is used as a validation for our method for deriving EMR phenotypes.

For each study, we examine two separate datasets: dataset E, which is a large EMR-only dataset and dataset RE, which is a smaller EMR and radiological dataset. Dataset E is used to identify EMR phenotypes that are most associated with a given disease group using PheDAS. These phenotypes are used to construct EMR context signatures in the dataset RE to evaluate additional predictive value that EMR data provides to imaging studies. $X_E$, an $N_1 \times P_1$ matrix, is constructed using dataset E with demographic data and ICD-9 data, where $N_1$ is the number of subjects and $P_1$ is the number of significant EMR phenotypes identified. $X_{RE}$, an $N_2 \times P_2$ matrix, is constructed using dataset RE which comprises of radiological imaging data and EMR data for both control and treatment groups, where $N_2$ is the number of subjects and $P_2$ is the number of parameters. $X_{RE}$ can be written as,

$$X_{RE} = [X_{Rad} \;\; X_{EMR}] \tag{1}$$

where, $X_{Rad}$ contains the radiological imaging features for all the subjects and $X_{EMR}$ contains binary vectors describing the EMR context signature for all the subjects. $N_1$ equals 29,214 in study 1 and 1,715 in study 2. $N_2$ equals 1,451 in study 1, and 124 in study 2.

We conduct seven case-control experiments in study 1. The first four experiments, where each of the four disease groups is compared with controls. And an additional three experiments, where each disease group is compared with other disease groups in a one v. all fashion. We conduct

one case-control experiment in study 2, comparing individuals with diabetes to those without the disease. The data for each case-control experiment is described in section 2C. The procedure for a case-control experiment is shown in Fig. 1. In the first step, we construct $X_E$ from dataset E and use PheDAS to find the EMR phenotypes associated with the disease in the experiment as described in section 2D. In step two, we use dataset RE to construct radiological imaging features $X_{Rad}$, as described in section 2E and EMR context signature vectors $X_{EMR}$, as described in the end of section 2D. Finally, we evaluate the EMR context signatures $X_{EMR}$, the imaging features $X_{Rad}$, and the combination of the two, using an elastic net paradigm [22] as described in section 2F and construct their Receiver Operating Characteristic (ROC) curves [23] for comparison.

## 2.3. Data

### *2.3.1. Study 1: Diseases of the Optic Nerve*

We collect ICD-9 codes, and demographic data for datasets E for each of the four disease groups and controls. For datasets RE, we collect computed tomography (CT) imaging of the eye orbit for each of the four disease groups and controls. The anonymized CT images were acquired clinically at Vanderbilt University Medical Center (VUMC) under varied settings and a wide range of scanners such as Phillips, Marconi, and GE (detailed acquisition parameters were not available). The CT imaging for disease population was acquired for subjects who have had an ICD-9 code belonging to one of the four main disease groups: glaucoma, intrinsic optic nerve disease, thyroid eye disease, and optic nerve edema. The detailed description of ICD-9 codes to identify a disease group is shown in Table II. The CT imaging for control subjects was acquired from subjects with hearing loss who had no other known vision problems as the clinical imaging protocols for these

subjects was similar to the imaging acquired for those with optic nerve disorders in terms of acquisition parameters and field of view.

For each disease group *d,* the counts and ages for datasets $X_E^{\{d\}}$, and $X_{RE}^{\{d\}}$, are shown in Table III, along with age-matched controls. The controls for $X_E^{\{d\}}$, are all individuals in study 1 who do not belong to disease group *d*, i.e. individuals belonging to other disease groups and hearing loss patients not included in $X_{RE}^{\{d\}}$. $X_E^{\{d\}}$ is used to learn EMR context signatures for each $d$.

**Table VII-2. Disease groups in Study 1 and the ICD-9 codes used to identify them**

| Disease group | ICD-9 codes |
|---|---|
| Glaucoma | 365.0* (Borderline glaucoma), 365.1* (Open-angle glaucoma), 365.2* (Primary angle-closure glaucoma), 365.3* (Corticosteroid-induced glaucoma), 365.4* (Glaucoma associated with congenital anomalies, dystrophies, and systemic syndromes), 365.5* (Glaucoma associated with disorders of the lens), 365.6* (Glaucoma associated with other ocular disorders), 365.7* (Glaucoma stage, unspecified), 365.8* (Other specified forms of glaucoma), and 365.9*(Unspecified glaucoma) |
| Intrinsic Optic Nerve Disease | 377.3* (Optic Neuritis), and 377.4* (Other disorders of optic nerve) |
| Optic Nerve Edema | 348.2 (Idiopathic intracranial hypertension), 377.0 and 377.00 (Papilledema), 377.01 (Papilledema, increased intracranial pressure), and 377.02 (Papilledema, decreased ocular pressure) |
| Thyroid Disease | 242.00 (Toxic diffuse goiter without thyrotoxic crisis or storm), 376.2 (Endocrine exophthalmos), 376.21 (Thyrotoxic exophthalmos), and 376.22 (Exophthalmic ophthalmoplegia) |

$X_{RE}^{\{d\}}$ is used to conduct two experiments to evaluate the predictive value of EMR and imaging data for each $d$:

**Disease vs. healthy control subjects**: The predictive value of EMR and imaging data is evaluated when compared to healthy control subjects. The controls for in this case are age-matched hearing loss patients. There are four case-control studies, one for each disease.

**Disease vs. other disease groups**: The predictive value of EMR and imaging data is evaluated when compared to subjects with other optic nerve diseases. The controls for this case are subjects who have never had disease $d$, but had one of the other three eye diseases. For example, subjects with glaucoma are compared with subjects who have never had glaucoma but had one of the following: intrinsic optic nerve disease, optic nerve edema or thyroid eye disease. In this category, subjects with optic nerve edema did not have enough age-matched controls, resulting in only three case-control studies.

Only ICD-9 codes recorded one year prior to of the diagnosis are included for both $X_E^{\{d\}}$, and $X_{RE}^{\{d\}}$.

*2.3.2. Study 2: Diabetes*

We collect ICD-9 codes, and demographic data for dataset E. For dataset RE, we collect magnetic resonance (MR) imaging of the brain for the disease group and controls along with ICD-9 codes.

The EMR dataset $X_E^{\{diab\}}$, has 245 patients with diabetes and 1470 age-matched controls, comprised of other subjects in the study without diabetes. The dataset with radiology data and EMR $X_{RE}^{\{diab\}}$, has 62 participants with diabetes and 62 age-matched controls. Their average ages are shown in Table III. Only ICD-9 codes recorded one year prior to of diagnosis are included for both $X_E^{\{diab\}}$ and $X_{RE}^{\{diab\}}$.

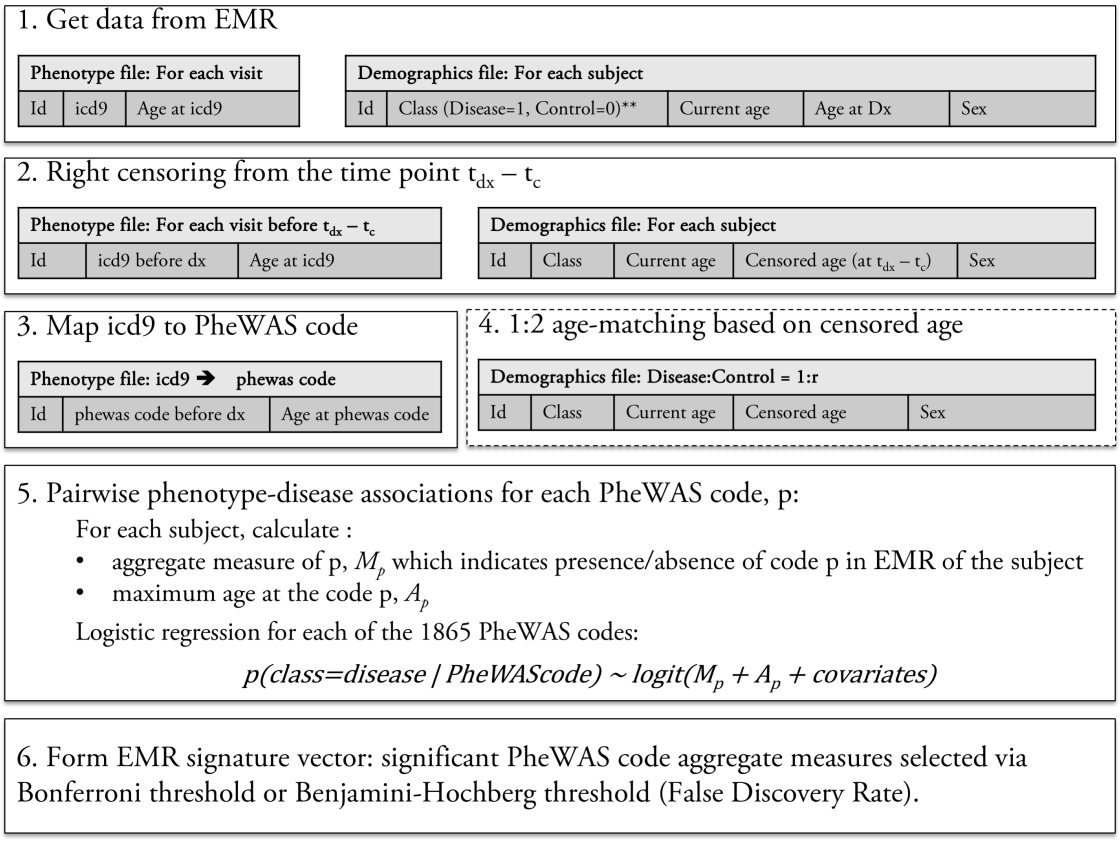



**Figure VII-2. Process flow of PheDAS. ICD-9 data for each visit and class information is extracted from EMR in step 1. Next, the data is censored by eliminating visits up to $t_c$ years before the time of diagnosis, $t_{dx}$. In step 3, ICD-9 codes are mapped to the PheWAS codes. In step 4, the data is age matched in 1:2 ratio with controls. In step 5, a logistic regression model is trained for each PheWAS code to determine if it is associated with the disease. In step 6, the EMR signature vector is constructed with conditions that are positively associated with the disease.**

## 2.4. Phenome-Disease Association Study (PheDAS)

We use our custom package pyPheWAS to extract EMR context signatures as described in this section[14]. First, phenome-disease associations are identified from EMR dataset E as shown in Fig. 2, for each case-control experiment. Two tables are extracted from dataset E for each disease group and their respective controls. The first table contains the subject identifier and each ICD-9 recorded for that subject along with the date. The second table consists of demographic information for each subject including current age, diagnosis class (1=disease group, 0=controls), age of the subject at the diagnosis, and sex. Next, the data are right censored such that all the time points that occur up to $t_c$ years before the diagnosis are eliminated, so that conditions that are temporally predictive of, or precursors to the diagnosis can be identified. In this paper, $t_c = 1$. The first table now contains only ICD-9 codes for visits before the diagnosis. The ages of the subjects are recalculated to reflect the censoring. Next, the ICD-9 codes are mapped to 1865 diagnostic

**Table VII-3. Average ages (in years) for control and disease population for study 1 and study 2.**

| Disease Group | Dataset E | | Dataset RE (Disease vs. Healthy) | | Dataset RE (Disease vs. Other) | |
|---|---|---|---|---|---|---|
| | Control | Disease | Control | Disease | Control | Disease |
| Glaucoma | 55.31±18.59 (n=11,474) | 55.31±18.76 (n=11, 499) | 56.12±17.6 (n=75) | 56.2±17.67 (n=75) | 58.10±9.40 (n=25) | 58.12±9.76 (n=25) |
| Intrinsic Optic Nerve Disease | 43.6±23.2 (n=2198) | 43.55±23.24 (n=1, 099) | 43.29±21.02 (n=135) | 43.3±21.0 (n=135) | 47.41±16.97 (n=27) | 47.51±16.91 (n=27) |
| Optic Nerve Edema | 28.32±19.09 (n=1604) | 28.25±19.06 (n=802) | 28.25±19.06 (n=115) | 30.47±16.58 (n=115) | - | - |
| Thyroid Eye Disease | 43.1±19.06 (n=3114) | 43.32±18 (n=1557) | 50.73±15.20 (n=40) | 50.65±15.27 (n=40) | 49.98±17.86 (n=15) | 50.26±17.30 (n=15) |

phenotype codes or PheWAS codes as defined in Denny et al[11]. These codes are denoted by *C*, where $C = \{c_k | k = 1 \ldots 1865\}$. Next, the diseases and controls are age-matched in a 1:2 ratio, when available, i.e., for each subject in the disease group, two subjects in the control group are selected whose ages are within 2 years. For each subject, an aggregate measure, $M_{c_k}$, of the PheWAS code $c_k$, is calculated. The PheDAS software supports computing $M_{c_k}$ by one of the following:

1. A binary variable, indicating presence or absence of the code in the subject's history.
2. A count, indicating the number of times the code was recorded in the subject's history.
3. Duration, indicating the time between the first time the code was recorded and the last time it was recorded.

In this paper, the binary measure is computed. Finally, the last recorded age of the subject for the given code, $A_{c_k}$, is calculated. Sex is the only other covariate that is considered. From this data, $X_E$ is given by,

$$X_E = \left[ sex^{N_1 \times 1} \quad M_{c_1}^{N_1 \times 1} \quad A_{c_1}^{N_1 \times 1} \quad \ldots \quad M_{c_{1865}}^{N_1 \times 1} \quad A_{c_{1865}}^{N_1 \times 1} \right]^{N_1 \times P_1}$$

Each subject belongs to class 1 (disease) or 0 (control). Each diagnostic phenotype $c_k$ is associated either with 1 or with 0 or is not associated with either. Logistic regression is used to determine the association between each $c_k$ and the disease of interest:

$$p(class = disease \mid c_k) \sim \text{logit}(\beta_{0_k} + \beta_{1_k} M_{c_k} + \beta_{2_k} A_{c_k} + \beta_{3_k} sex)$$

The p-value of $\beta_{1_k}$, the coefficient of the binary aggregate measure, is used to determine the significance of the association between the PheWAS code $c_k$ and the diagnosis of the disease. The sign of the co-efficient determines the direction of the association. The threshold for

significance, $p_{bh}$, is given by the false discovery rate[24]. The set of significant PheWAS codes that are positively correlated with the disease is given by $C_{sig} = \{c_{k^*} | \, p - value \; of \; \beta_{1_{k^*}} < p_{bh} \; and \; \beta_{1_{k^*}} > 0\}$. To illustrate this process, Fig. 3 presents $C_{sig}$ for Glaucoma dataset, $X_E^{\{glauc\}}$.

Next, the ICD-9 codes are mapped to PheWAS codes in dataset RE of the as described above, for each disease group. The set of phenotypes positively associated with the disease D, $C_{sig}$, that was identified using dataset E is used to construct EMR context signature vectors $X_{EMR}$, from dataset RE (refer to equation (1)),

$$X_{EMR} = \left[sex \; age \;\; M_{c_{1^*}}^{N_2 \times 1} \;\; M_{c_{2^*}}^{N_2 \times 1} \ldots\right]^{N_2 \times P_2} \tag{2}$$

where $c_{k^*} \in C_{sig}$ and $M_{c_{k^*}}^{N_2 \times 1}$ is the binary vector of aggregate measures for $c_{k^*}$ for each of the $N_2$ subjects.

## 2.5. Image processing

### *2.5.1. Study 1: Diseases of the Optic Nerve*

CT imaging from dataset RE is used in study 1 to extract radiological imaging features, $X_{Rad}$. Multi-atlas segmentation [25] is used to segment the optic nerve, globes, extraocular rectus muscles, and orbital fat from CT imaging, as illustrated in Fig. 4. A set of 25 example atlases with expertly marked labels for the orbital structures are used to segment the structures in a new target scan. First, the target image is registered to the example atlases to localize the eye orbit and crop it using a bounding box[26]. This is done since target atlases can have varying fields of view from a whole head scan to a scan limited to the eye orbit. A down-sampled version of the target image is used for this step since it is faster, and a rough registration is sufficient to identify a bounding box. Once the eye orbit is localized, the full-resolution images of the cropped target space are used.

Next, the example scans are non-rigidly registered [27] to the cropped target space and the expertly labeled structures are propagated to the target space. The registered labels in the target space are combined using non-local statistical label fusion [28] to identify the globe, optic nerve, muscles, and fat.

However, challenges arise in the identification of the individual extraocular rectus muscles that control eye movement. This is increasingly problematic in diseased eyes, where these muscles often appear to fuse at the back of the orbit (at the resolution of clinical computed tomography

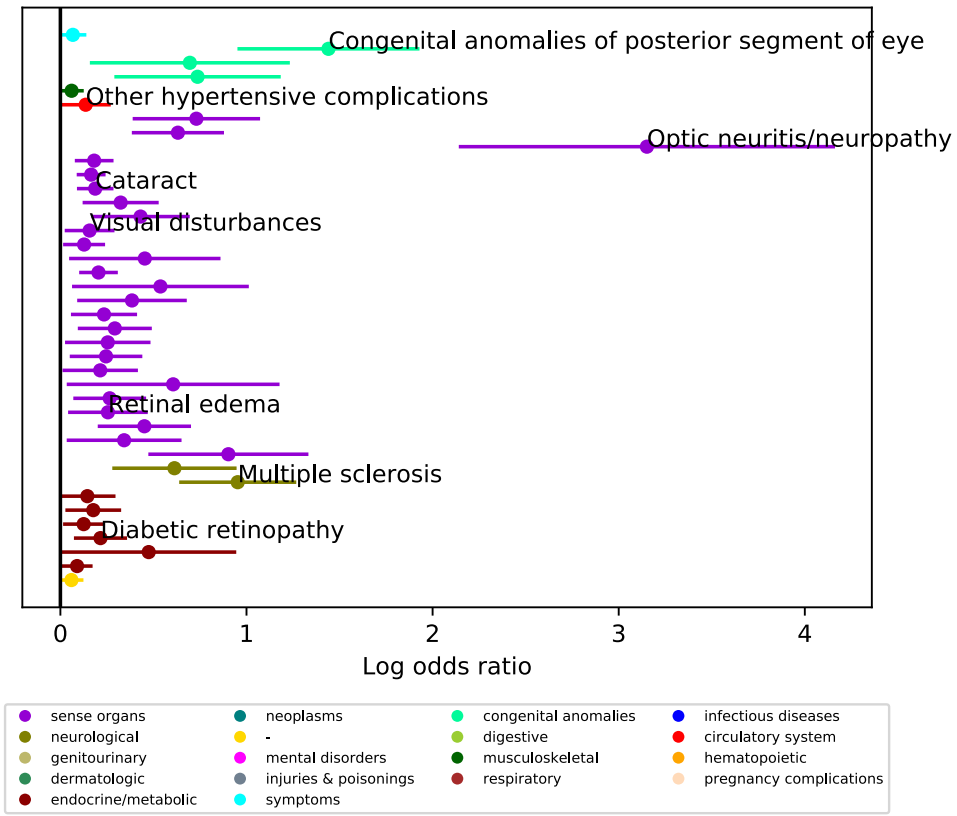



**Figure VII-3. Example result of phenotype extraction. Shown here are all the positively correlated phenotypes with Glaucoma. A few key associations are labelled. The complete list of associated conditions is shown in Table B1 of the supplementary material.**

imaging) due to inflammation or crowding. Kalman filters are used to isolate the individual extraocular muscles from the muscle labels obtained from the multi-atlas segmentation pipeline as described by Chaganti et al[15]. We start at a coronal slice at the center of the globe, where the muscles are well-separated, and use Kalman filters to track each muscle in z-direction. After the segmentation, thirty-six volume, size and intensity metrics are calculated as described in Chaganti et al[15]. These *orbital structural metrics* included the (#1-#20) volume, maximum diameter, average diameter, median intensity, and interquartile range of the intensities for the superior, inferior, medial, and lateral rectus muscles and total rectus muscle volume; (#21) Barrett muscle index; (#22-#25) volume, diameter, median intensity and the interquartile range of the intensities of the globe; (#26) orbital volume; (#27) volume crowding index; (#28) orbital angle; (#29) degree of proptosis; (#30) length along the optic nerve, and (#31-#36) traditional length, volume, average area, maximum diameter, median intensity, and the interquartile range of intensities of the optic nerve. All structural metrics are computed bilaterally, and their mean is used for analysis. The radiological imaging data vector $X_{Rad}$, in equation (1) is formed from these metrics,

$$X_{Rad} = [\, sm_1^{N_2\times 1}\ sm_2^{N_2\times 1}\ \dots\ sm_{36}^{N_2\times 1}] \tag{3}$$

where, $sm_k$ is the average of the $k^{th}$ structural metric of the left and right eye.

*2.5.2. Study 2: Diabetes*

In study 2, $\mathrm{X_{Rad}}$ is calculated from segmented MR images of the brain as described in [29]. A multi-atlas segmentation paradigm is used for the segmentation. Briefly, scans are first affinely registered [30] to the MNI305 atlas[31]. Then, multi-atlas segmentation is performed on each subject. 45 MPRAGE images from OASIS dataset are used as original atlases which are manually labeled with 133 labels (132 brain regions and 1 background) by the

BrainCOLOR protocol[32]. Multi-atlas segmentation produces regional masks for each of the 133 labels. The volume of each label, except the background, is computed by integrating the individual masks, and a radiological imaging data vector $\mathrm{X_{Rad}}$, is formed,

$$\mathrm{X_{Rad}} = [\ \mathrm{sm_1^{N_2\times 1}}\ \mathrm{sm_2^{N_2\times 1}}\ \ldots\ \mathrm{sm_{132}^{N_2\times 1}}]$$

where, $\mathrm{sm_k}$ is the volume of the $\mathrm{k^{th}}$ brain region.

2.6. Classification

A binary classifier is trained on dataset RE for each case-control experiment to distinguish subjects belonging to the disease group from the control group. The data vector $X_{RE}$, is calculated from equations 2, 3, and 4 as follows,

$$X_{RE} = [\ X_{Rad} \quad X_{EMR}]^{N_2\times P_2}$$

The relative performance of each of $X_{RE}$, $X_{Rad}$, and $X_{EMR}$ is evaluated. For each *X,* we model

$$P(y_i = disease|X = x_i) = \frac{e^{\beta_0+\beta^T x_i}}{1+e^{\beta_0+\beta^T x_i}}$$

We use logistic regression with elastic net regularization[22] . The following penalized log likelihood function is minimized to derive the $\beta$ for the logistic regression model,

$$\min\ -\left[\sum_{i=1}^{N} y_i(\beta_0+\beta^T x_i)\ -\log\left(1+e^{\beta_0+\beta^T x_i}\right)\right] + \lambda[(1-\alpha)\|\beta\|_2^2/2\ \ +\alpha\|\beta\|_1]$$

Here, $\alpha$ is fixed at 0.5, and $\lambda$ that provides the optimal test error is selected. $\|\beta\|_2^2$ is the ridge penalty and $\|\beta\|_1$ is the lasso penalty[33].

We use ROC curves to evaluate the performance of the model. We build the final model using all of the data, and the extra-sample ROC estimate is computed using bootstrapping for *optimism correction*[34]–[36]. The extra sample error is the generalization error of the model, which is estimated as described in this section.

From the original $X^{N_2 \times P_2}$, a sample, $X_s^{N_2 \times P_2}$, of size $N_2$ is selected with replacement and the model is fitted to it. Let $X_s^+$ denote the positive class samples, and $X_s^-$ denote negative class samples of $X_s$. Let $X_{orig}^+$ denote the positive class samples, and $X_{orig}^-$ denote negative class samples of $X$. Let $\widehat{f_s}$ be the classification rule for the model fitted over the dataset $X_s$. The true and false positive rates, $tpr_{s,s}(t)$ and $fpr_{s,s}(t)$, of applying $\widehat{f_s}$ to $X_s$ are calculated over thresholds $t$,

$$tpr_{s,s}(t) = \frac{1}{size(X_s^+)} \sum_{x_i, y_i \in X_s^+} I\left(\widehat{f_s}(x_i) \geq t\right), \qquad t \in [0,1]$$

$$fpr_{s,s}(t) = \frac{1}{size(X_s^-)} \sum_{x_i, y_i \in X_s^-} I\left(\widehat{f_s}(x_i) \geq t\right), \qquad t \in [0,1]$$

Similarly, the true and false positive rates, $tpr_{s,orig}(t)$ and $fpr_{s,orig}(t)$, of applying $\widehat{f_s}$ to $X$ are calculated over thresholds $t$,

$$tpr_{s,orig}(t) = \frac{1}{size\left(X_{orig}^+\right)} \sum_{x_i, y_i \in X_{orig}^+} I\left(\widehat{f_s}(x_i) \geq t\right), \qquad t \in [0,1]$$

$$fpr_{s,orig}(t) = \frac{1}{size\left(X_{orig}^-\right)} \sum_{x_i, y_i \in X_{orig}^-} I\left(\widehat{f_s}(x_i) \geq t\right), \qquad t \in [0,1]$$

Now, we model all of the points in the original dataset $X$ using elastic net regression to obtain a new classification rule, $\hat{f}_{orig}$. Next, we calculate the apparent true and false positive rates, $tpr_{app}(t)$ and $fpr_{app}(t)$, of applying $\hat{f}_{orig}$ to $X$. This would give an optimistic estimate, but we correct for that using the optimism correction approach[34]:

$$tpr_{corrected}(t) = tpr_{app}(t) - \left(\frac{1}{\text{num}_{\text{bootstrap}}}\right)\sum_{s}\left(tpr_{s,s}(t) - tpr_{orig,s}(t)\right), t \in [0,1]$$

$$fpr_{corrected}(t) = fpr_{app}(t) - \left(\frac{1}{\text{num}_{\text{bootstrap}}}\right)\sum_{s}\left(fpr_{s,s}(t) - fpr_{orig,s}(t)\right), t \in [0,1]$$

The optimism corrected receiver operating characteristic curve is given by,

$$ROC(.) = \left\{\left(tpr_{corrected}(t), fpr_{corrected}(t)\right)\middle| t \in 0{:}\,0.01{:}\,1\right\}$$

The corrected area under the curve is calculated for the ROC using the trapezoidal rule[23].

## 3. Results

### 3.1. Study 1: Diseases of the Optic Nerve

#### *3.1.1. Case-Control Experiment: Glaucoma vs. Healthy Controls*

Structural metrics derived from CT imaging are moderately successful in distinguishing disease and control groups with an AUC of 0.71 (Fig. 5A). The EMR context signature vector results in slightly more successful prediction with an AUC of 0.76. However, the model that was built upon using both sets of data has much higher prediction accuracy, with an AUC of 0.83.

#### *3.1.2. Case-Control Experiment: Intrinsic Optic Nerve Disease vs. Healthy Controls*

Structural metrics derived from CT imaging are successful in distinguishing disease and control groups with an AUC of 0.72 (Fig. 5B). The EMR context signature vector results in an AUC of 0.85. The model that was built upon using both sets of data has higher prediction accuracy than either of the former models, with an AUC of 0.91.

*3.1.3. Case-Control Experiment: Optic Nerve Edema vs. Healthy Controls*

Structural metrics derived from CT imaging have a high AUC of 0.96 (Fig. 5C). The EMR context signature vector results in an AUC of 0.95. The model that was built upon using both sets of data has slightly higher prediction accuracy with an AUC of 0.96.

*3.1.4. Case-Control Experiment: Thyroid Eye Disease vs. Healthy Controls*

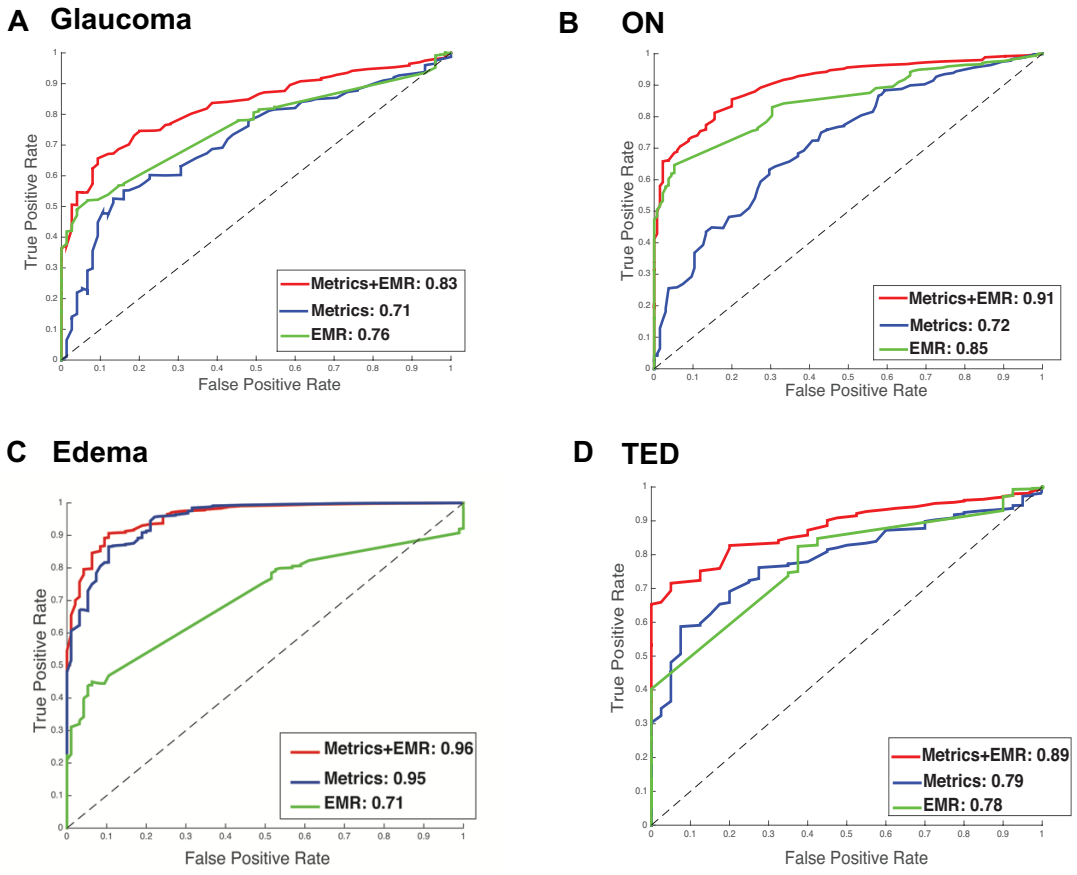



**Figure VII-5. Disease vs. healthy control results of elastic net classifier for study. Green line indicates the curve for EMR data, blue line indicates the curve for imaging data, and the red line indicates the curve for EMR + imaging data.1. 5A. shows the result for glaucoma, 5B shows the result for intrinsic optic nerve disease, 5C shows the result for optic nerve edema, and 5D shows the result for thyroid eye disease.**

Structural metrics derived from CT imaging are successful in distinguishing disease and control groups with an AUC of 0.79 (Fig. 5D). The EMR context signature vector results in an

AUC of 0.78. The model that was built upon using both sets of data has higher prediction accuracy than either of the former models, with an AUC of 0.89.

*3.1.5. Case-Control Experiment: Glaucoma vs. Other Diseases*

Structural metrics derived from CT imaging are mildly predictive in distinguishing disease and control groups with an AUC of 0.66 (Fig. 6A). The EMR context signature vector is more predictive with an AUC of 0.88. The predictive value of the combined model that was built using both sets of data has an improved prediction accuracy with an AUC of 0.91.

*3.1.6. Case-Control Experiment: Intrinsic Optic Nerve Disease vs. Other Diseases*

Structural metrics derived from CT imaging are mildly successful in distinguishing disease and control groups with an AUC of 0.68 (Fig. 6B). The EMR context signature vector results in an AUC of 0.85. The model that was built upon using both sets of data has higher prediction accuracy than either with an AUC of 0.90.

*3.1.7. Case-control experiment: Thyroid eye disease vs. other diseases*

Structural metrics derived from CT imaging are successful in distinguishing disease and control groups with an AUC of 0.74 (Fig. 6C). The EMR context signature vector results in an AUC of 0.73. The combined model has a much higher prediction accuracy than either with an AUC of 0.90.

3.2. Study 2: Diabetes

Structural metrics derived from MR are slightly better than chance, with an AUC of 0.58, as seen in Fig. 7. The EMR context signature vector has an AUC of 0.83. The model that was built

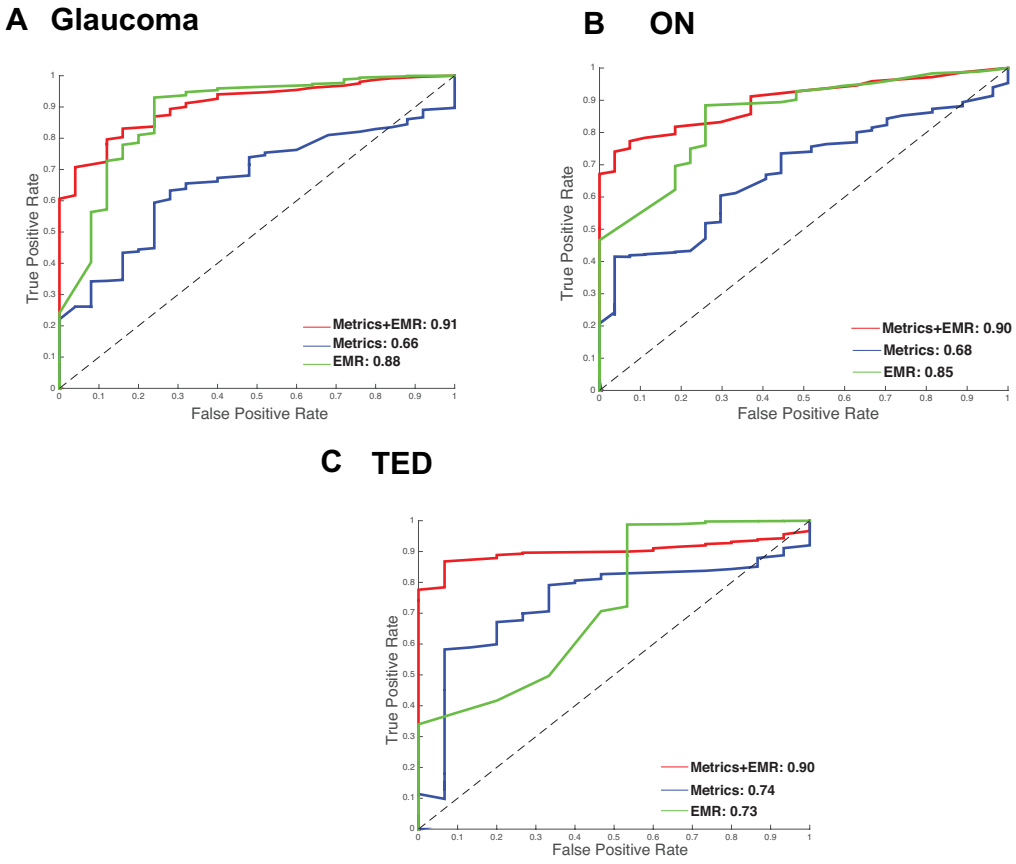



**Figure VII-6. Disease vs. other results of elastic net classifier for study 1. Green line indicates the curve for EMR data, blue line indicates the curve for imaging data, and the red line indicates the curve for EMR + imaging data.1. 6A. shows the result for glaucoma, 6B shows the result for intrinsic optic nerve disease, 6C shows the result for thyroid eye disease.**

upon using both sets of data does not select any imaging features in the elastic net model, and demonstrated an AUC of 0.85 using abnormal glucose as the only selected feature.

## 4. Discussion

Although success has been reported in several studies that were able to train diagnostic classifiers using medical imaging data in various domains, these studies do not consider the variability of presentation in radiology due to individual patient histories. One way to circumvent this problem is to use data available in electronic medical records to develop context-based

diagnostic classifiers. In this paper, we develop methods for multi-modal big data studies in medical image processing that use EMR information, i.e., data from medical images as well other EMR data such as ICD-9 codes. One of the challenging problems in such studies is to consolidate data from EMR records that can be integrated with imaging data. We develop PheDAS, a method to extract diagnostic phenotypes associated with a given condition and transform it into a binary EMR signature vector. In fact, our custom Python package provides provisions to transform the data using other aggregate measures such as counts of the diagnostic phenotypes, and duration of the diagnostic phenotypes[14].

Our main contribution in this paper is evaluation of these context-based diagnostic classifiers using two large-scale studies. In our two studies, we look at two different datasets: study 1 in which significant predictive value for imaging data is observed and study 2 in which the predictive value of imaging data is minimal. In study 1, we find that both structural metrics and PheWAS phenotypes are moderately successful in distinguishing disease groups from controls for diseases of the optic nerve. However, in all four of the disease groups, we saw that addition of the EMR context signature vectors increases the predictive power of the classifiers, making a strong case for context-based imaging studies that consider the whole story of the patient with the help of information readily available in EMR. In study 2, which was considered to be our validation dataset to evaluate the EMR-phenotype extraction, abnormal glucose, the most important clinical predictor of diabetes[18], [37], is selected as the only predictor by our final model, thus validating the methodology. The context-based classifier for diabetes does not select any imaging features and shows a final AUC of 0.85 using abnormal as a sole feature.

Upon further evaluation of the EMR and imaging phenotypes identified in study 1, several findings align with those reported in medical literature. Thyroid eye disease is associated with a

history of Grave's disease, an autoimmune disorder that causes problems with thyroid regulation [38], [39]. The final model for thyroid eye disease selects phenotypes associated with thyroid imbalances such as hyperthyroidism. The imaging phenotypes selected in this model are proptosis, optic nerve length, and inferior muscle volume and diameter. According to radiology literature, these are some of the most distinct features of TED [40], [41] and inferior muscle is the most commonly affected muscle[42].

The final model for glaucoma identified systemic problems such as diabetes, hyperlipidemia, hypertension, and miscellaneous visual changes as the most predictive, which coincide with common glaucoma co-morbidities reported in medical literature [43]. The EMR context vector for intrinsic optic nerve disease shows a history of multiple sclerosis and other demyelinating diseases of the central nervous system as significant, which is reported widely in medical literature as well[44], [45]. For both glaucoma and intrinsic optic nerve disease the structural changes from imaging calculated in this study were shown to be predictive in differentiating the subject population from controls.

The optic nerve edema disease group consists of papilledema and idiopathic intracranial hypertension. Previous studies show that volumetric changes of the optic nerve are important features used for diagnosis of these conditions [46]–[49]. Our final model identified optic nerve volume, length and intensity, as some of the important predictors of this disease group, thereby achieving a very high AUC of 0.95. The systemic causes for idiopathic intracranial hypertension are not known [50]. The presentation is variable and is associated with other co-morbidities such as obesity, migraine, nausea, depression and so on. Our final model identifies these EMR features as significant, however they do not significantly improve the final prediction over imaging features. The reason for this could be that EMR features for this condition are

inconsistent and variable, whereas imaging features are quite distinct and consistent across the subjects.

In future studies, we will develop methods to incorporate other data routinely available in EMR such as procedure codes and labs into the context vectors to develop comprehensive clinical decision support models.

## Chapter VIII. Discovering Novel Disease Comorbidities using Electronic Medical Records

### 1. Introduction

Electronic medical record (EMR) systems have increasingly been leveraged for clinical and medical research. EMR data provides several advantages including large sample sizes that are often required to obtain statistically significant results in observational studies. They contain a rich variety of data including lab results, medications, clinical notes, administrative and billing codes, images and so on. Large-scale EMR databases also provide access to a wider range of diseases to study. Moreover, models developed on real-world data will be more generalizable than carefully selected homogenous cohort studies. Applications of EMR-based models in clinical research range from studying adverse drug events [210], [211], performing pragmatic clinical trials[212], predicting clinical outcomes[213], predicting hospital readmission rates[214], and genome-wide association studies[215]. In this work, we introduce a tool to perform an association study to identify disease comorbidities using EMR databases at different timepoints.

Early association studies were prospective trials that recruited case and control subjects to identify genetic variants associated with a disease of interest, called GWAS (genome-wide association study). However, there has been significant effort recently to use existing data from other sources such as observational data from EMR databases. In 2007, the eMERGE (Electronic Medical Records and Genomics) consortium was formed to facilitate large-scale genomic research by linking DNA biorepositories to EMR systems in 7 clinical sites and 2 pediatric sites [216]. Denny et al have used EMR-based definitions of diseases to perform reverse GWAS studies i.e., to identify all clinical phenotypes association with a genetic variant of interest[180]. The concept of association studies has been expanded to study disease-wide comorbidities as well. Engels et al

used a Medicare claims database to study comorbidities associated with Non-Hodgkin's Lymphoma[217]. Hanauer et al developed a disease-wide comorbidity map, similar to the molecular concept map (MCM) using EMR records[218]. Roque et al used clinical text and ICD-10 codes to stratify patients by phenotypes, and identify clusters of co-occurring conditions[219]. Holmes et al used a combination of discharge summaries, ICD-9 codes, PubMed database and Wikipedia articles to identify co-morbidities in three rare diseases [220].

There are several challenges in the design and analysis of disease co-morbidity, including: 1) Identifying disease phenotypes from EMR, 2) Identifying comorbidities that occur at different stages: before diagnosis, co-occurring, and after diagnosis, 3) Identifying clinically relevant associations, 4) Identifying novel associations that were previously unknown or understudied. In this work, we describe phenome-disease association study or PheDAS, an EMR-based open-source association study tool to evaluate relationships between an index disease (disease of interest) and other clinical phenotypes.

A disease phenotype can be described as a set of characteristics or symptoms that consistently define the clinical presentation disease of interest. Several studies have been conducted to automatically extract phenotypes from EMR data such as billing codes and clinical summaries. Lasko et al describe a method that uses an unsupervised data-driven approach to identify latent phenotypes from EMR[221]. cTAKES was developed by Mayo clinic to extract phenotypes from clinical text[101]. The PheKB catalog provides algorithms to extract clinical phenotypes using multiple sources in EMR[102]. i2b2 is another natural language processing (NLP) system that used EMR to extract phenotypes[103]. While these studies have shown promise, there are still ongoing analyses to study the generalizability of these methods across all databases and in cases where there are missing data types. Meanwhile, the adaptation of phecodes has proven

successful in genetic association studies. Phecodes were defined by Denny et al, based on hierarchical categorization of ICD-9 codes[180]. ICD-9 codes are universally available in EMR databases, so phecodes can be readily adapted with any given use case. They have the added benefit of easy interpretation when compared to other methods that describe phenotypes through complex algorithms that use different types of data. In this work, we use phecodes to define our disease phenotypes.

Previous studies of disease comorbidity show the overall associations in terms of a graph, or a list. Often, it is a challenge to separate comorbid conditions to identify those that are predictive and those that are a consequence of an index disease. Our tool provides data cleaning and censoring abilities to address these questions. The longitudinal data available for each patient in an EMR database can be used to design various types of experiments including selection of specific age intervals, censoring the disease diagnosis to study only predictive factors before the diagnosis of the disease, or studying the future progression of a disease. In this paper, we will demonstrate a use case for each of these scenarios.

In such studies, there is a problem of multiple hypothesis testing, since several associations are studied at the same time. Previous association studies employed tradition multiple hypothesis corrections such as Bonferroni, false discovery rate (FDR), or permutation testing[222], [223]. Despite these measures, association studies have suffered from the problem of reproducibility[224]. GWAS studies are usually followed by meta analyses and replication studies to identify truly significant results[225]. For reduced false discoveries and improved interpretability, we use second generation p-values proposed by Blume et al [125]. Second generation p-value approach was introduced to tackle the common challenges that arise with

traditional p-values. In this approach, a clinically meaningful or relevant null-interval is set prior to the analysis.

Once the significant associations are identified, quantifying the quality and relevance of these associations remains a huge unaddressed challenge. Exploratory association studies can be validated when they find and replicate well-known results. In genetic association studies, meta analyses are used to substantiate results[225]. In previous disease co-morbidities, associations were manually reviewed by experts to recognize well-known results[220]. Another main advantage of association studies is hypothesis generation for future research questions. These tools could potentially be used to discover new empirical associations that have previously been unexplored by the scientific community. In our tool, we introduce a new measure called *Novelty Finding Index* that quantifies and ranks the novelty of a finding automatically by performing an automated clinical literature search using PubMed.

We demonstrate the use of this tool in three studies: 1) We will use this tool to identify comorbidities of Autism Spectrum Disorder (ASD) among subjects of ages 7 and above (excluding early childhood). 2) We will identify predictive clinical phenotypes for Alzheimer's disease that appear five years prior to a diagnosis. 3) We will identify the clinical phenotypes that present after a diagnosis of Optic Neuritis, that show that variability of prognosis in this condition.

## 2. Methods

### 2.1. Data

#### *2.1.1. Case-Control Study 1: Autism Spectrum Disorder*

The data for this study, including demographic and ICD-9 codes, was collected at the Vanderbilt Kennedy Center. The index disease group for this study was defined by patients with a

diagnosis of ASD (ICD-9 codes – 299.*). The control group is defined by subjects with typical development.

*2.1.2. Case-Control Study 2: Alzheimer's Disease*

The data for this study was collected through the Baltimore Longitudinal Study of Aging (BLSA), a study that collects longitudinal data of an aging population in order to examine changes in the brain as a person ages[198]. The data contains self-reported ICD-9 codes and demographic data. In this study, the index disease group is defined by individuals who were diagnosed with Alzheimer's disease. The control group for this study is individuals in BLSA who had no cognitive impairment.

*2.1.3. Case-Control Study 3: Optic Neuritis*

The data for this study was collected from Vanderbilt University's Synthetic Derivative under IRB approval. It contains EMR data including ICD-9 codes and demographic information. The index disease group is defined by patients with codes 377.30-377.39. The control group for this study are subjects with other disorders of the optic nerve or subjects with hearing loss.

2.2. Phenome-Disease Association Study

We developed a python tool to perform phenome-disease association studies (PheDAS). PheDAS is used to identify clinical phenotypes that are associated with a given index disease. A clinical phenotype or a *phecode* is a code based on hierarchical categorization of ICD-9 (International Classification of Disease - 9) codes, which describes a diagnostic "phenotype" by grouping a set of related ICD-9 codes. The ~15,000 ICD-9 codes are mapped to 1865 phecodes as described by Denny et al (2013). For example, "depression" phecode 296.2 groups the ICD-9 codes of "major depressive disorder, single episode, mild degree" (*ICD-9 = 296.21*), "major

depressive disorder, recurrent episode, mild degree" (*ICD-9 = 292.31*), and "depressive disorder NEC" (*ICD-9 = 311*). For each phecode, a set of exclusion codes are also defined which can be used to select a control cohort.

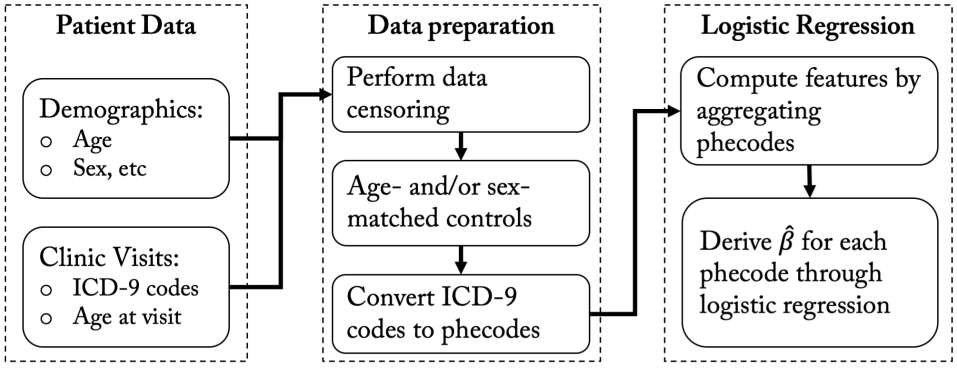



**Figure VIII-1. Flow chart for phenome-disease association study. The input patient data required for this analysis is demographic data and clinic visits data. The data is prepared by performing data censoring and control matching based on the experimental design. Next, the ICD-9 codes are converted to phecodes. Finally, logistic regression is performed for each phecode based on aggregate measures and demographic features as described in section 2.2.**

Given a disease group and a control group, the PheDAS tool performs a set of logistic regressions to identify significant phenotypes associated with the disease. A flowchart of the process is shown in Figure VIII-1. The ICD-9 codes for each clinical visit and other demographic information are extracted from each subject's electronic medical record (EMR). Optionally, the *time interval* for extraction of ICD-9 codes can also be adjusted according to the study design. This can be done by,

*Censoring by age-interval*: Selecting an age range within which to perform the analysis. (Ex. In *study 1*, we analyze the differences between ASD and control population after age 7); or

*Left-censoring with respect to diagnosis:* Selecting a time interval prior to year of diagnosis. (Ex. In s*tudy 2*, we analyze the differences between Alzheimer's and control population 0-5 years before the diagnosis of the disease); or

*Right-censoring with respect to diagnosis:* Selecting a time interval post the year of diagnosis. (Ex. In *study 3,* we analyze the differences between Optic Neuritis and control population 0-5 years after the diagnosis of the disease).

After defining the interval of the study, ICD-9 codes are extracted for all clinic visits during the period and converted to phecodes for every using the mapping provided by Denny et al. These codes are denoted by $C = \{c_k | k = 1 \ldots 1865\}$. For each code $c_k$, an aggregate measure $m_k$ is computed in order to perform the logistic regression. The regression tool can be set to one of the following options:

- *Binary measure:* aggregate codes to indicate the presence or absence of the phenotype $c_k$($m_k$ = 0 or 1),
- *Count measure:* aggregate codes to indicate the number of times $c_k$ was present in a subject's EMR ($m_k = n$), or
- *Duration measure:* aggregate codes to indicate the time interval between the first and the last time the phenotype $c_k$ was recorded in a subject's EMR ($m_k = t$).

Additional covariates such as age and sex can also be provided, if available. For each $c_k$, the following logistic regression is performed,

$$p(class = disease|c_k) = logit(\theta_0 + \theta_1 m_k + \theta_2 a_k + \theta_3 s_k),$$

where, $a_k$ is age and $s_k$ is sex. The co-efficient of the aggregate measure $\theta_1$ is used to determine the significance of the association between the disease and phenotype $c_k$. The

framework for determining the significance based on this point estimate is described in the following sections.

2.3. Second-Generation P-value

We used the second-generation *p*-value (SGPV) approach described by Blume et al (2018) to prioritize or rank potential associations. The SGPV framework requires (1) a pre-defined "indifference zone" or interval null hypothesis around the null effect to denote the set of effect magnitudes that would not be clinically meaningful and (2) an uncertainty interval for the observed association, e.g. a confidence interval, likelihood support interval, or credible interval. The SGPV, denoted by $p_\delta$, measures the overlap between the data-supported effect sizes (#2) and the interval null (#1). See Blume (2018) for details.

The SGPV equals 0 when #2 and #1 do not overlap. In this case the data only support effect sizes in the alternative hypothesis space. We take all cases where the SGPV is zero, $p_\delta = 0$, to be clinically interesting and statistically 'significant'. In contrast, when $p_\delta = 1$, the data support only effects that are null or nearly null and not of clinical interest. These results would confirm the lack of association. SGPVs between 0 and 1 are treated as inconclusive as the data support both null and alternative hypotheses.

2.4. Positive Predictive Value

It is of interest to know how reliable SGPV finding are when $p_\delta = 0$ and whether or not this finding as already known in the literature. To address these two important questions, we estimated the positive predictive value (PPV) when the SGPV is zero and developed a "novelty score" by scraping and searching relevant abstracts in PubMed. In order to compute the PPV (1 -

false discovery rate), we also need to know the point estimate and standard error for the association of interest, which is often available from interval #2.

Let $\hat{\theta}$ be the point estimate, $\hat{V}_n$ be the estimated variance, and let $(\hat{\theta}_l, \hat{\theta}_u)$ be the lower and upper bounds of the uncertainty interval. Define the interval null hypothesis as $H_0: \theta \in \Theta_0 = [\theta_0^-, \theta_0^+]$, as compared to the typical point null hypothesis of $H_0: \theta = \theta_0$ (here $\theta_0^- = \theta_0^+ = \theta_0$). Note that this yields an alternative hypothesis of the form $H_1: \theta \in \Theta_1 = (-\infty, \theta_0^-) \cup (\theta_0^+, \infty)$. The PPV is the probability that the null hypothesis is true, given that $SGPV = 0$. This is just

$$PPV = 1 - \left[1 + \frac{1-\tilde{\beta}}{\tilde{\alpha}} \frac{1-\pi_0}{\pi_0}\right]^{-1}$$

where $1 - \tilde{\beta} = P(p_\delta = 0 \mid H_1)$ is the power function of the SGPV averaged over the alternative space, and $\tilde{\alpha} = P(p_\delta = 0 \mid H_0)$ is the Type I Error function of the SGPV averaged over the null space, and $\pi_0 = P(H_0)$ is the analysts' a-priori probability of the null hypothesis before data were collected. Computational details require some integration and are provided in the appendix, and for our examples we set $\pi_0 = 0.5$ which is the default non-informative approach.

2.5. Novelty Finding Index

The 'novelty score' is intended to approximate the extent to which a finding is well-studied in the literature. We used published abstracts from the PubMed database to construct the 'novelty score' as follows: For each index disease, and for each phecode-disease pairing, we obtained the number of published papers in which these are mentioned in the title, abstract, or keywords section.

In order to search PubMed database, we convert phecodes to *search strings* using the metathesaurus database provided as a part of the unified medical language system (UMLS)[226], as shown in Figure VIII- 2. The UMLS metathesaurus defines unique medical concepts that are

unchanged over time, identified by the Concept Unique Identifier (CUI). It links strings with the same meaning from over 200 different source vocabularies to the same CUI. ICD-9 codes are included as a part of source vocabularies provided by UMLS. For each phecode, the ICD-9 codes attached to it are linked to a CUI. Next, all possible strings associated with the CUI are extracted from the metathesaurus to be used as search strings. Henceforth, we will take 'mentioned' to mean the CUI terms linked to a phecode to be mentioned in either the title, abstract, or keywords section.

We then compute the proportions of published papers that mention the phecode-disease pairing out of all published papers that mention the disease (termed the 'PubMed proportion'). This proportion measures whether the associations between the outcome and the predictor phecodes are well-studied in the literature. Note that well-studied does not necessarily mean well-

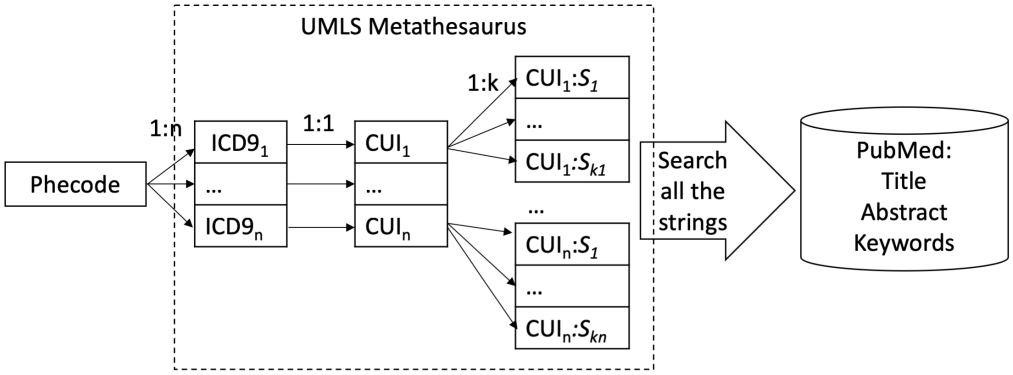



**Figure VIII-2. Searching PubMed for associations. For each Phecode, all the ICD-9 codes associated with it are mapped to their CUIs (concept unique identifiers). Next, all the strings associated with the CUIs in UMLS metathesaurus. These strings are used to search all the titles, abstracts and keywords in the PubMed database to identify the counts of academic research papers associated with each phecode.**

known to be associated (i.e., the PubMed proportion should not be interpreted as the estimated probability that an association exists). We denoted the novelty score by $N_s = 1 - \hat{F}(x)$, where

$\hat{F}(x)$ is the empirical cumulative distribution function estimated with the PubMed proportions under consideration.

Then, to provide a ranking that accounts for the reliability of the finding (PPV) and its relative novelty ($N_s$), we use a *Novel Finding Index* (NFI) defined as $NFI = (PPV \cdot N_s) * 100$. The NFI is used to rank associations; its magnitude is not important or interpretable in an absolute sense. Significant associations ($p_\delta = 0$) with NFI near 0 indicate that the finding is well-known and/or not reliable, while significant associations with NFI near 100 indicate that the finding is potentially novel and likely to be reliable.

## 3. Results

In this section we present the results of the PheDAS association study for three index diseases, with different censoring conditions.

### 3.1. Case-Control Study 1: Autism Spectrum Disorder

For ASD, we censored the data to remove all visits from early childhood. We look at significant associations after age 7 in patients diagnosed with ASD. This cohort consists of 1234 (926M and 308F) patients diagnosed with ASD and 1234 (932 M and 302F) age-matched controls. A binary aggregate measure was used for this analysis. Figure VIII-3 shows significant associations ranked by novel finding score. The null interval for this study was picked to be [0.3, 1.5]. We see several expected associations such as epilepsy and mood disorders. Among novel findings we see increased while blood count, hemorrhoids and glaucoma.

### 3.2. Case-Control Study 2: Alzheimer's Disease

For Alzheimer's disease, we right censored the data with respect to the year of diagnosis. We look at significant associations 0-5 years before the diagnosis of Alzheimer's disease. For this

study, we selected the binary aggregate measure for 242 (145M and 97F) subjects with Alzheimer's and 789 age- and sex-matched controls (499M and 290F). Figure VIII-4 shows significant associations ranked by novel finding score. The null interval for this study was picked to be [0.3, 1.1]. We see several expected associations such as psychosis, cerebral degenerations, and abnormality of gait. Among novel findings we see several musculoskeletal disorders and infections.

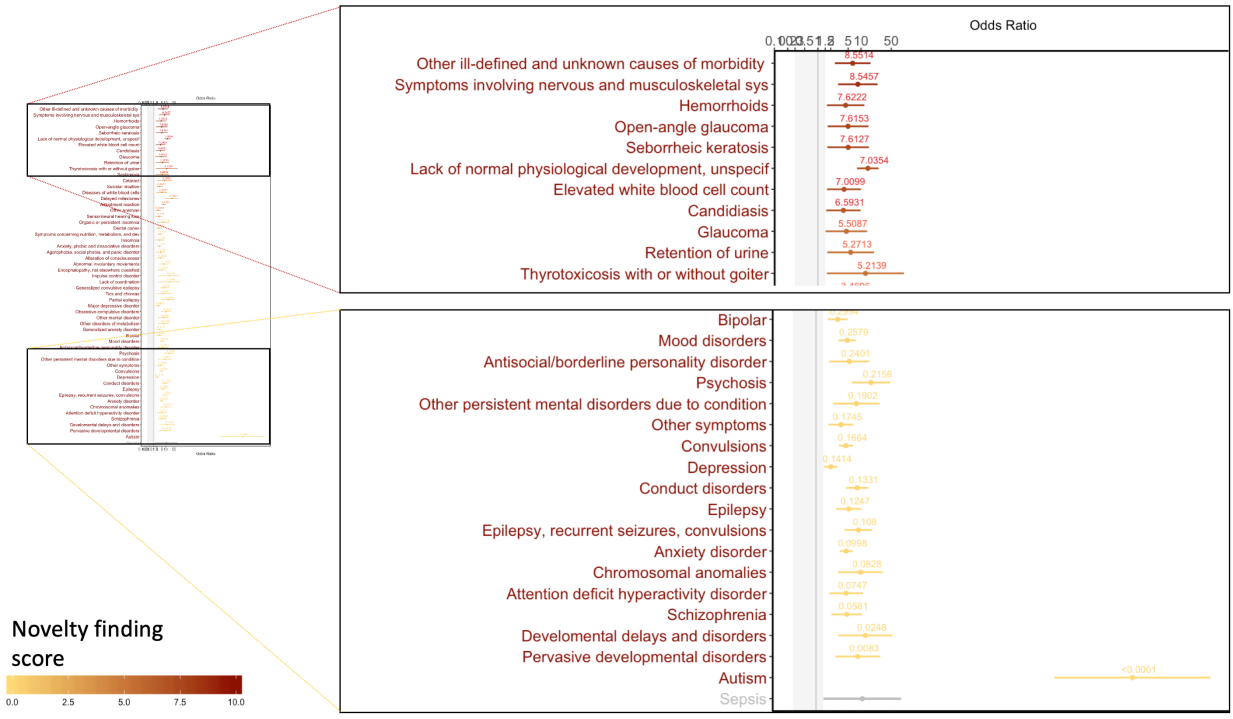



**Figure VIII-3. Significant associations for ASD**

## 3.3. Case-Control Study 2: Optic Neuritis

For optic neuritis, we left censored the data with respect to the year of diagnosis. We look at significant associations 0-5 years post the first optic neuritis diagnosis. We performed this study over 1085 (685M and 405F) subjects with optic neuritis and 1085 (685M and 405F) age- and sex-matched controls. Figure VIII-5 shows significant associations ranked by novel finding score. The

null interval for this study was picked to be [0.3, 2]. We see several expected associations that are seen in yellow including visual disturbances, blindness, and multiple sclerosis (MS). Among the novel findings, we see increased skull and face fractures, and subarachnoid hemorrhage. While these conditions are not directly linked to optic neuritis, these are associated with MS, which could be one of the complications associated with optic neuritis.

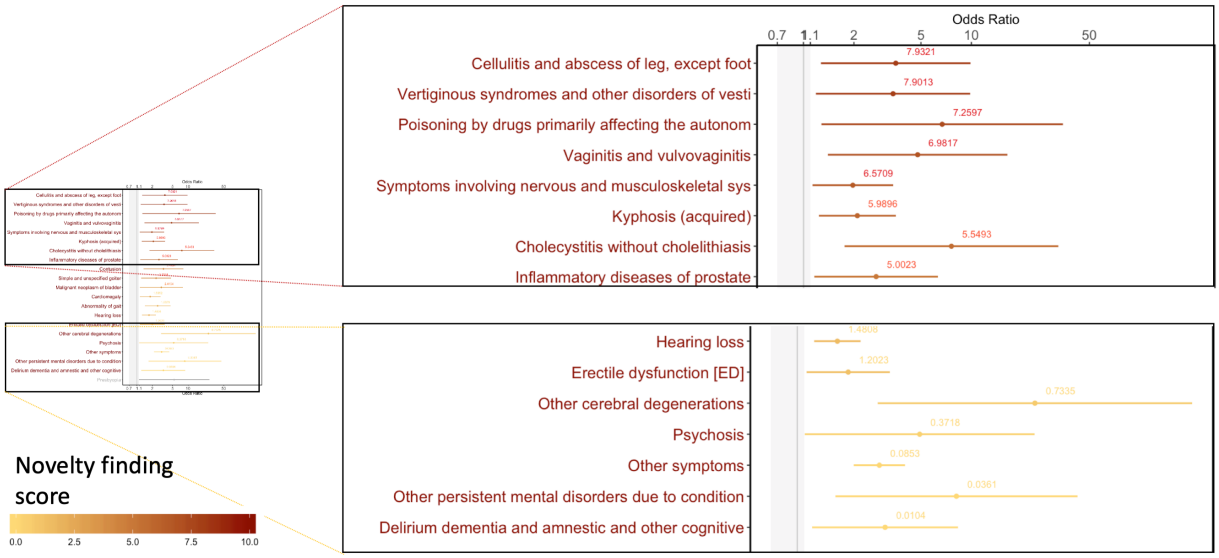



**Figure VIII-4. Significant associations in Alzheimer's disease**

## 4. Discussion

In this paper, we describe a general tool for discovering disease co-morbidities at different stages of presentation, from EMR data that is routinely collected during clinic visits. We utilized the concept of second-generation p-values to identify significant associations. In this approach, we specify a null interval for effects size that are scientifically uninteresting a priori to identify meaningful findings. The statistical interpretation of the second-generation p-values is intuitive, they describe the fraction of empirically derived intervals that overlap with the null-intervals. When $p_\delta = 0$, the null hypothesis is rejected and we properly conclude that the finding is

significant. The interval null hypothesis eliminates the need for adjustments that traditional p-values require to control false discovery, thus eliminating a major problem of reproducibility that most association studies have.

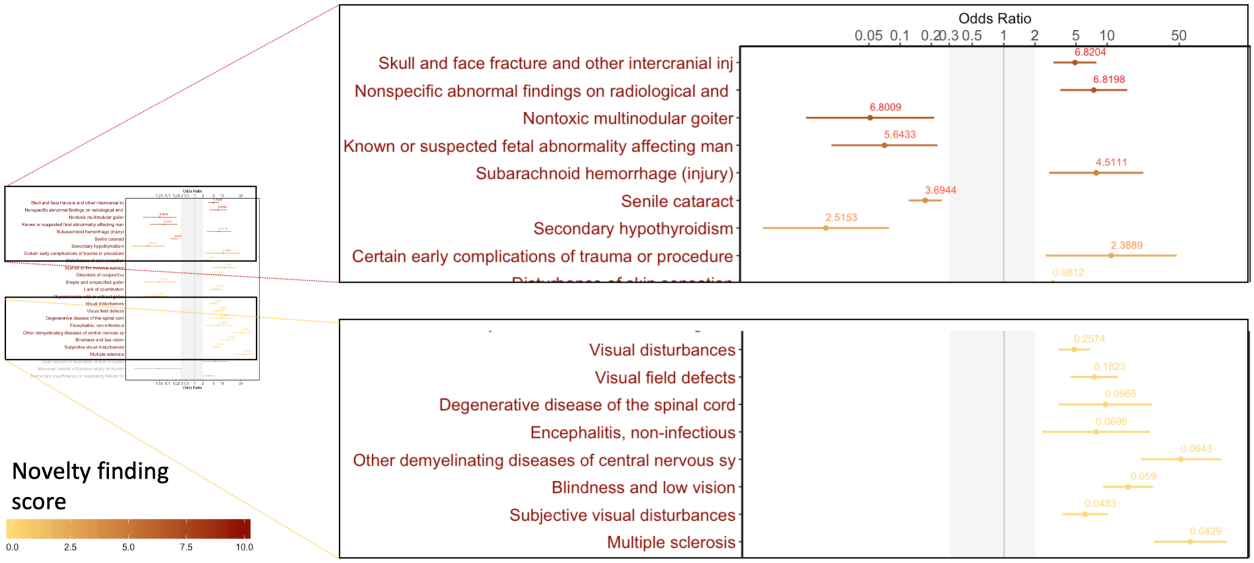



**Figure VIII-5. Significant conditions in Optic Neuritis**

In this paper, we introduce a novel approach to rank the significant findings based on prior scientific knowledge. We do this by comparing each disease-phecode finding to the number of papers that can be found on PubMed that mention both the index disease as well as the phecode as a proportion of the number of papers published on the index disease. We define a novelty score (see section 2.3), which moves the PubMed proportion, which in some sense is on an absolute scale, onto a relative scale. For example, Parkinson's disease is scored low on the novelty score (i.e. not considered to be a novel finding) because it is the 5th most frequent predictor phecode that is studied with Alzheimer's, despite the fact that it has only been studied in about 5% of the papers that mention Alzheimer's. A novelty finding index (NFI) is derived from novelty score and the positive predictive value of the association, to indicate the novelty and reliability of the finding.

This method has a two-fold benefit. The first, the methodology presented in this paper can be validated by the novelty finding index by automatically identifying phecode predictors that are well-known by the scientific community. The researcher is assured that that results of the experiment are indeed correct, thereby increasing confidence in the analysis. For instance, in our ASD example we see a majority of significant associations that have a low NFI and are well-understood by the ASD research community: psychiatric conditions, developmental disorders and seizure disorders. The second advantage of this method is that it could be used for hypothesis exploration. Conditions with a high NFI can be used to drive new studies to investigate associations that are empirically present in the data but have not been well-studied previously. For example, in the ASD study we found that "hemorrhoids" was a significant finding and had a high NFI. This direct link has not been studied extensively in medical literature. However, it is well-known that autism is associated with several gastro-intestinal problems, which could explain complications with hemorrhoids.

# Chapter IX. Conclusions and Future Work

## 1. Summary

The goal of this thesis was to develop a system that performs automated analysis of a patient's electronic medical history including past clinical visits and imaging, to identify salient features and develop models for predicting clinical outcomes. The rationale behind this work is that early detection and intervention in diseases of the optic nerve and orbit is important and necessary. Several studies have shown that the outcomes for conditions affecting the optic nerve can be severe, resulting in permanent blindness if timely intervention is not performed. We identified that one of the main problems was inundation of data from different sources, modalities and departments which makes the task of diagnosis and intervention planning a very difficult problem, especially in acute situations. In high-pressure environments, sifting through gigabytes of patient data to identify the right problem is challenging. Especially in the case of optic nerve conditions, where data from different sources needs to be evaluated simultaneously. Therefore, our solution was to develop automated analysis methods for CT imaging and EMR data. We found that information for early prediction of disease and disease function can be obtained from radiological imaging such as CT and MR and electronic medical records. We developed an image processing system to analyze imaging in the orbit to automatically detect important structures and their measurements. We developed an EMR analysis system to automatically identify potential co-morbidities and past risk factors that explain ophthalmological symptoms. We showed that this information has predictive value by developing models that showed correlations with prognosis markers such as visual disability scores. Furthermore, we showed that analyzing images in the context of a patient's story obtained from the EMR improves traditional images analyses. In this

work, we laid the groundwork to demonstrate for the first time that automated data analysis of orbital imaging and EMR have the potential to identify diseases early. This work can potentially impact identification and intervention in conditions related to the orbit and the optic nerve to prevent permanent vision loss. Additional analyses and retrospective trials are needed to evaluate if this system can be used in the clinic. If implemented successfully, it can streamline the workflow in ER departments and ophthalmology departments.

## 2. Image Analysis of the Eye Orbit

Diseases of the optic nerve such as glaucoma, optic neuropathy, papilledema, idiopathic intracranial hypertension, thyroid eye disease and other orbital infections affect millions of people every year. These conditions could lead to permanent blindness or severe vision loss if not treated in a timely manner. Quantitative assessment of the optic nerve and other structures of interest can improve early detection and thereby facilitate timely intervention and prevention of vision loss. In this work, we explored the use of automated image analysis methods to objectively assess clinically acquired imaging. We developed an automated image analysis pipeline to identify the structures in the eye orbit and compute quantitative metrics to evaluate the extent of disease.

### 2.1. Main Contributions

- We were the first to develop a fully automated image analysis pipeline to segment orbital structures from clinical CT images using a multi-atlas framework. The following structures were segmented: the optic nerve, the globe, extraocular muscles and orbital fat (Chapter III).
- We showed that Kalman filters can be used in medical image data processing to segment structures that are difficult to identify with multitask segmentation. We used Kalman filters to track and separate the multi-atlas extraocular rectus muscles label into the four constituent

muscles: superior, inferior, lateral and medial rectus muscles (Chapter III). As a result of this work, we were able to demonstrate the differential involvement of the four extraocular muscles in thyroid eye disease. This result has previously been postulated in clinical literature, but it was never shown empirically in a large cohort before.

- Assessing the quality of non-rigid registration is an extremely challenging problem. We developed a method called QUADRATIC (Quality of Dice in Registration Circuits) to estimate the Dice error associated with non-rigid registration, using the concept of registration circuits (Chapter IV). The ability to have a quantitative measure of non-rigid registration can lead to improved interpretation of data analyses that rely on non-rigid registration to draw conclusions about disease populations. It could also be used in applications such as atlas selection for multi-atlas methods.
- We identified imaging phenotypes for Thyroid Eye disease using principal component analysis. We showed that the imaging phenotypes derived from clinical data corresponded to the hypothesized presentation of thyroid eye disease in clinical literature (Chapter V).

## 3. Extracting Phenotypes from Electronic Medical Records

Adaptation of digitized electronic medical records (EMR) at large hospitals and university medical centers provides unique opportunities to perform population level analyses for disease etiology and progression. In this work, we present a methodology that uses ICD-9 codes from large clinical databases to learn relationships between diseases. Often, medically relevant information is present in the past records of a patient's EMR. Most clinical decisions are made after a thorough review of a patient's historical diagnoses, labs and procedures. We developed tools to automatically identify significant co-morbidities and conditions associated with an index disease using historical data from EMR databases.

3.1. Main Contributions

- We released a custom open-source python package called pyPheWAS to extract diagnostic and procedural phenotypes from EMR data. This is a highly generalizable tool, which is used learn associations between a disease and EMR phenotypes with various options for selection of time interval, age interval, censoring, age matching, and plotting (Chapter VII). This tool can be used by researchers interested in learning about co-morbidities in any medical condition, as long as they have access to an EMR database that records ICD-9 codes.
- With the help of new statistical measures (second generation p-values and novel finding index), we showed that significant co-morbidities identified through the pyPheWAS tool can be automatically ranked by novelty of the finding based on a novel method that searches the PubMed database (Chapter VIII). As a result of this work, we developed a custom shiny app which can be used for exploratory analyses to discover novel hypotheses for research.

## 4. Contextualizing Medical Image Analyses with Electronic Health Histories

Medical image analysis studies are used to study anatomical structures of interest using imaging modalities such as computed tomography, magnetic resonance imaging, positron emission tomography and so on. Traditionally, these studies focused on identify interesting imaging features such as shape, volume, or intensity to identify disease or function. More recently, deep learning methodologies have been used to automatically identify disease or function with variants of convolutional neural networks. However, in several medical problems, the answers lie in several different sources of data, including a patient's medical history, past diagnoses and procedures, medications and so on. Often, clinical imaging of patients is interpreted within the context of the entire patient story. Therefore, there is a need to develop models that integrate data

from multiple sources to form composite models of disease prediction. In this work, we demonstrate that addition of data from electronic health records improves traditional image processing studies.

### 4.1. Main Contributions

- This work was the first to show that features extracted from CT imaging of the eye orbit are correlated with visual function scores. Further we demonstrated that addition of EMR phenotypes to the imaging data improved the correlations with visual function (chapter VI). This result was an important first-step to show that context-based models that integrate patient histories from EMR can be used to improve image processing studies.
- We were also the first to show that imaging features from the eye orbit can be used to classify control CT scans from scans of diseases of the optic nerve. In this analysis, we once again tested and proved that addition of EMR phenotypes, namely the diagnostic phenotypes as additional features to the classifier improves the classification accuracy of the model (chapter VII).

## 5. Concluding Remarks

The main contribution of this work was to show that context-aware models that include EMR phenotypes along with imaging markers consistently perform better than traditional analysis using imaging markers alone. While imaging plays an important role in the diagnosis of several diseases, the patient's story needs to be tied together across different departments to form a composite picture of complete medical history. Most studies in medical image analysis develop models that focus on learning geometrical and intensity-based features in medical images such as CT and MR to develop predictive models. On the other hand, medical informatics community is

focused on developing models that use clinical data alone to study phenotype extraction, and provide clinical decision support. This is the first work of its kind that combines methods from the two fields to show that integrated models can be used predict optic nerve diseases better than standalone methods. As a result of this work, we released an open-source python package called pyPheWAS that can be installed and used by researchers. The concepts discussed in this work could be easily extended to study other clinically relevant problems, for example intervention planning and risk management. In chapter VII and chapter VIII, we show examples of this work being applied to other diseases diabetes, autism spectrum disorder, and Alzheimer's disease. The image and EMR analysis methods developed in this work can be generalized to any disease area where the presentation is complex and requires a thorough evaluation of different data resources. In order to implement systems that can be translated into the clinic, we propose that the direction of big data in medicine should be focused on developing integrated models that learn from a composite picture derived from multiple sources of data.

## Appendix A: Publications

1. Journal Articles

1. **Shikha Chaganti**, Louise A. Mawn, Hakmook Kang, Josephine Egan, Susan M. Resnick, Lori L. Beason-Held, Bennett A. Landman, Thomas A. Lasko. "Electronic Medical Record Context Signatures Improve Diagnostic Classification using Medical Image Computing". *Journal of Biomedical and Health Informatics*. doi: 10.1109/JBHI.2018.2890084
2. **Shikha Chaganti**, Katrina Nelson, Kevin Mundy, Robert Harrigan, Robert Galloway, Louise A. Mawn, and Bennett A. Landman. "Imaging Biomarkers in Thyroid Eye Disease and their Clinical Associations." *Journal of Medical Imaging* 5, no. 4 (2018): 044001.
3. **Shikha Chaganti**, Kevin Mundy, Michael P. DeLisi, Katrina Nelson, Robert Harrigan, Robert Galloway, Bennett A. Landman, and Louise A. Mawn. "Assessment of Orbital Computed Tomography (CT) Imaging Biomarkers in Patients with Thyroid Eye Disease." *Journal of Digital Imaging*.
4. Harrigan, Robert L., Swetasudha Panda, Andrew J. Asman, Katrina M. Nelson, **Shikha Chaganti**, Michael P. DeLisi, Benjamin CW Yvernault et al. "Robust optic nerve segmentation on clinically acquired computed tomography." *Journal of Medical Imaging* 1, no. 3 (2014): 034006.
5. Yuankai Huo*, Justin Blaber*, Stephen M. Damon, Brian D. Boyd, Shunxing Bao, Prasanna Parvathaneni, Camilo Bermudez Noguera, Shikha Chaganti, Vishwesh Nath, Jasmine M. Greer, Ilwoo Lyu, William R. French, Allen T. Newton, Baxter P. Rogers, and Bennett A. Landman. "Towards Portable Large-scale Image Processing with High-Performance Computing". *Journal of Digital Image* 31, no. 3 (2018): 304-314.

2. Conference Publications

1. **Chaganti, Shikha**, Katrina Nelson, Kevin Mundy, Yifu Luo, Robert L. Harrigan, Steve Damon, Daniel Fabbri, Louise Mawn, and Bennett Landman. "Structural functional associations of the orbit in thyroid eye disease: Kalman filters to track extraocular rectal muscles." In Medical Imaging 2016: Image Processing, vol. 9784, p. 97841G. International Society for Optics and Photonics, 2016.
2. **Chaganti, Shikha**, and Bennett A. Landman. "Quadratic: quality of dice in registration circuits." In Medical Imaging 2018: Image Processing, vol. 10574, p. 105740P. International Society for Optics and Photonics, 2018.
3. **Chaganti, Shikha**, Jamie R. Robinson, Camilo Bermudez, Thomas Lasko, Louise A. Mawn, and Bennett 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, pp. 373-381. Springer, Cham, 2017.
4. **Chaganti, Shikha**, Kunal P. Nabar, Katrina M. Nelson, Louise A. Mawn, and Bennett A. Landman. "Phenotype analysis of early risk factors from electronic medical records improves image-derived diagnostic classifiers for optic nerve pathology." In Medical Imaging 2017: Imaging Informatics for Healthcare, Research, and Applications, vol. 10138, p. 101380F. International Society for Optics and Photonics, 2017.
5. **Chaganti, Shikha**, Andrew J. Plassard, Laura Wilson, Miya A. Smith, Mayur B. Patel, and Bennett A. Landman. "A Bayesian framework for early risk prediction in traumatic brain injury." In Medical Imaging 2016: Image Processing, vol. 9784, p. 978422. International Society for Optics and Photonics, 2018.

## REFERENCES


[1] C.-J. Hsiao and E. Hing, "Use and Characteristics of Electronic Health Record Systems Among Office-Based Physician Practices, United States, 2001-2012," 2012.

[2] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Heal. Inf. Sci. Syst.*, vol. 2, no. 1, p. 3, 2014.

[3] R. Margolis, L. Derr, M. Dunn, M. Huerta, J. Larkin, J. Sheehan, M. Guyer, and E. D. Green, "The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data," *J. Am. Med. Informatics Assoc.*, vol. 21, no. 6, pp. 957–958, 2014.

[4] S. Saria, "A $ 3 Trillion Challenge to Computational Scientists : Transforming Healthcare Delivery," pp. 2–7, 2014.

[5] T. D. Wilson, "Information overload: implications for healthcare services," *Health Informatics J.*, vol. 7, no. 2, pp. 112–117, 2001.

[6] J. S. Ash, M. Berg, and E. Coiera, "Some unintended consequences of information technology in health care: the nature of patient care information system-related errors," *J. Am. Med. Informatics Assoc.*, vol. 11, no. 2, pp. 104–112, 2004.

[7] H. Singh, C. Spitzmueller, N. J. Petersen, M. K. Sawhney, and D. F. Sittig, "Information overload and missed test results in electronic health record–based settings," *JAMA Intern. Med.*, vol. 173, no. 8, pp. 702–704, 2013.

[8] B. I. Reiner and E. Krupinski, "The insidious problem of fatigue in medical imaging practice," *J. Digit. Imaging*, vol. 25, no. 1, pp. 3–6, 2012.

[9] A. P. Kansagra, K. Liu, and J. Y. John-Paul, "Disruption of radiologist workflow," *Curr. Probl. Diagn. Radiol.*, vol. 45, no. 2, pp. 101–106, 2016.

[10] J. Y. John-Paul, A. P. Kansagra, and J. Mongan, "The radiologist's workflow environment: evaluation of disruptors and potential implications," *J. Am. Coll. Radiol.*, vol. 11, no. 6, pp. 589–593, 2014.

[11] C. S. Lee, P. G. Nagy, S. J. Weaver, and D. E. Newman-Toker, "Cognitive and system factors contributing to diagnostic errors in radiology," *Am. J. Roentgenol.*, vol. 201, no. 3, pp. 611–617, 2013.

[12] B. I. Reiner and E. Krupinski, "Innovation strategies for combating occupational stress and fatigue in medical imaging," *J. Digit. Imaging*, vol. 25, no. 4, pp. 445–448, 2012.

[13] M. Recht and R. N. Bryan, "Artificial Intelligence: Threat or Boon to Radiologists?," *J. Am. Coll. Radiol.*, 2017.

[14] S. Wang and R. M. Summers, "Machine learning and radiology," *Med. Image Anal.*, vol. 16, no. 5, pp. 933–951, 2012.

[15] M. De Bruijne, "Machine learning approaches in medical image analysis : From detection to diagnosis," *Med. Image Anal.*, vol. 33, pp. 94–97, 2016.

[16] A. Criminisi, D. Robertson, E. Konukoglu, J. Shotton, S. Pathak, S. White, and K. Siddiqui, "Regression forests for efficient anatomy detection and localization in computed tomography scans," *Med. Image Anal.*, vol. 17, no. 8, pp. 1293–1303, 2013.

[17] M. N. Gurcan, B. Sahiner, N. Petrick, H. Chan, E. A. Kazerooni, P. N. Cascade, and L. Hadjiiski, “Lung nodule detection on thoracic computed tomography images: Preliminary evaluation of a computer-aided diagnosis system,” *Med. Phys.*, vol. 29, no. 11, pp. 2552–2558, 2002.

[18] S. G. Armato, F. Li, M. L. Giger, H. MacMahon, S. Sone, and K. Doi, “Lung cancer: performance of automated lung nodule detection applied to cancers missed in a CT screening program,” *Radiology*, vol. 225, no. 3, pp. 685–692, 2002.

[19] M. S. Brown, M. F. McNitt-Gray, J. G. Goldin, R. D. Suh, J. W. Sayre, and D. R. Aberle, “Patient-specific models for lung nodule detection and surveillance in CT images,” *IEEE Trans. Med. Imaging*, vol. 20, no. 12, pp. 1242–1250, 2001.

[20] D. L. Langer, T. H. van der Kwast, A. J. Evans, J. Trachtenberg, B. C. Wilson, and M. A. Haider, “Prostate cancer detection with multi-parametric MRI: Logistic regression analysis of quantitative T2, diffusion-weighted imaging, and dynamic contrast-enhanced MRI,” *J. Magn. Reson. imaging*, vol. 30, no. 2, pp. 327–334, 2009.

[21] M. Moradi, Y. Guo, Y. Gur, M. Negahdar, and T. Syeda-Mahmood, “A Cross-Modality Neural Network Transform for Semi-automatic Medical Image Annotation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2016, pp. 300–307.

[22] M. Moradi, Y. Gur, H. Wang, P. Prasanna, and T. Syeda-Mahmood, “A hybrid learning approach for semantic labeling of cardiac CT slices and recognition of body position,” in *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*, 2016, pp. 1418–1421.

[23] F. Ciompi, B. de Hoop, S. J. van Riel, K. Chung, E. T. Scholten, M. Oudkerk, P. A. de Jong, M. Prokop, and B. van Ginneken, “Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box,” *Med. Image Anal.*, vol. 26, no. 1, pp. 195–202, 2015.

[24] F. Ciompi, K. Chung, S. J. Van Riel, A. A. A. Setio, P. K. Gerke, C. Jacobs, E. T. Scholten, C. Schaefer-Prokop, M. M. W. Wille, and A. Marchianò, “Towards automatic pulmonary nodule management in lung cancer screening with deep learning,” *Sci. Rep.*, vol. 7, p. 46479, 2017.

[25] W. Shen, M. Zhou, F. Yang, D. Dong, C. Yang, Y. Zang, and J. Tian, “Learning from experts: developing transferable deep features for patient-level lung cancer prediction,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2016, pp. 124–131.

[26] T. Kooi, G. Litjens, B. van Ginneken, A. Gubern-Mérida, C. I. Sánchez, R. Mann, A. den Heeten, and N. Karssemeijer, “Large scale deep learning for computer aided detection of mammographic lesions,” *Med. Image Anal.*, vol. 35, pp. 303–312, 2017.

[27] M. Giger, K. Tipton, G. Newstead, R. Tomek, and B. Luerrsen, “Quantitave Insights FDA Clearance,” 2017. .

[28] A. Blum and M. Zins, “Radiology: Is its future bright?” Elsevier, 2017.

[29] S. B. Marzoli and V. Martinelli, “Optic neuritis: differential diagnosis,” *Neurol. Sci.*, vol. 22, no. 2, pp. S52–S54, 2001.

[30] J. Rootman, *Diseases of the orbit: a multidisciplinary approach*. Lippincott Williams & Wilkins, 2003.

[31] R. J. Casson, G. Chidlow, J. P. M. Wood, J. G. Crowston, and I. Goldberg, "Definition of glaucoma: clinical and experimental concepts," *Clin. Experiment. Ophthalmol.*, vol. 40, no. 4, pp. 341–349, 2012.

[32] C. Cook and P. Foster, "Epidemiology of glaucoma: what's new?," *Can. J. Ophthalmol.*, vol. 47, no. 3, pp. 223–226, 2012.

[33] M. A. Kass, D. K. Heuer, E. J. Higginbotham, C. A. Johnson, J. L. Keltner, J. P. Miller, R. K. Parrish, M. R. Wilson, and M. O. Gordon, "The Ocular Hypertension Treatment Study: a randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma," *Arch. Ophthalmol.*, vol. 120, no. 6, pp. 701–713, 2002.

[34] M. O. Gordon, J. A. Beiser, J. D. Brandt, D. K. Heuer, E. J. Higginbotham, C. A. Johnson, J. L. Keltner, J. P. Miller, R. K. Parrish, and M. R. Wilson, "The Ocular Hypertension Treatment Study: baseline factors that predict the onset of primary open-angle glaucoma," *Arch. Ophthalmol.*, vol. 120, no. 6, pp. 714–720, 2002.

[35] A. E. Kolker, "Visual prognosis in advanced glaucoma: a comparison of medical and surgical therapy for retention of vision in 101 eyes with advanced glaucoma.," *Trans. Am. Ophthalmol. Soc.*, vol. 75, p. 539, 1977.

[36] O. N. S. Group, "The clinical profile of optic neuritis: experience of the Optic Neuritis Treatment Trial," *Arch Ophthalmol*, vol. 109, pp. 1673–1678, 1991.

[37] C. N. Pedro-Egbe, "Visual outcome following optic neuritis: A 5-year review," *Niger. J. Clin. Pract.*, vol. 15, no. 3, pp. 311–314, 2012.

[38] B. J. Osborne and N. J. Volpe, "Optic neuritis and risk of MS: differential diagnosis and management.," *Cleve. Clin. J. Med.*, vol. 76, no. 3, pp. 181–190, 2009.

[39] S. J. Hickman, C. M. H. Brierley, P. A. Brex, D. G. MacManus, N. J. Scolding, D. A. S. Compston, and D. H. Miller, "Continuing optic nerve atrophy following optic neuritis: a serial MRI study," *Mult. Scler. J.*, vol. 8, no. 4, pp. 339–342, 2002.

[40] D. I. Kaufman, J. D. Trobe, E. R. Eggenberger, and J. N. Whitaker, "Practice parameter: The role of corticosteroids in the management of acute monosymptomatic optic neuritis Report of the Quality Standards Subcommittee of the American Academy of Neurology," *Neurology*, vol. 54, no. 11, pp. 2039–2044, 2000.

[41] M. X. Repka, P. J. Savino, N. J. Schatz, and R. C. Sergott, "Clinical profile and long-term implications of anterior ischemic optic neuropathy," *Am. J. Ophthalmol.*, vol. 96, no. 4, pp. 478–483, 1983.

[42] S. R. Lambert, C. S. Hoyt, and M. H. Narahara, "Optic nerve hypoplasia," *Surv. Ophthalmol.*, vol. 32, no. 1, pp. 1–9, 1987.

[43] S. R. Lambert, T. E. Johnson, and C. S. Hoyt, "Optic nerve sheath and retinal hemorrhages associated with the shaken baby syndrome," *Arch Ophthalmol*, vol. 104, no. 10, pp. 1509–1512, 1986.

[44] J. J. Ney, N. J. Volpe, G. T. Liu, L. J. Balcer, M. L. Moster, and S. L. Galetta, "Functional visual loss in idiopathic intracranial hypertension," *Ophthalmology*, vol. 116, no. 9, pp. 1808–1813, 2009.

[45] A. K. Ball and C. E. Clarke, "Idiopathic intracranial hypertension," *Lancet Neurol.*, vol. 5, no. 5, pp. 433–442, 2006.

[46] L. L. C. D. Bursztyn, S. Sharan, L. Walsh, G. R. LaRoche, J. Robitaille, and I. De Becker, "Has rising pediatric obesity increased the incidence of idiopathic intracranial hypertension in children?," *Can. J. Ophthalmol. Can. d'Ophtalmologie*, vol. 49, no. 1, pp. 87–91, 2014.

[47] W. R. Nunery, C. W. Nunery, R. T. Martin, T. V Truong, and D. R. Osborn, "The risk of diplopia following orbital floor and medial wall decompression in subtypes of ophthalmic Graves' disease.," *Ophthalmic plastic and reconstructive surgery*, vol. 13, no. 3. pp. 153–60, 1997.

[48] G. B. BARTLEY and C. A. GORMAN, "Diagnostic Criteria for Graves' Ophthalmopathy," *Am. J. Ophthalmol.*, vol. 119, no. 6, pp. 792–795, 1995.

[49] T. Cawood, P. Moriarty, and D. O'shea, "Recent developments in thyroid eye disease," *BMJ Br. Med. J.*, vol. 329, no. 7462, p. 385, 2004.

[50] J. A. Mauriello Jr, N. Yepez, R. Mostafavi, J. Barofsky, R. Kapila, S. Baredes, and J. Norris, "Invasive rhinosino-orbital aspergillosis with precipitous visual loss.," *Can. J. Ophthalmol.*, vol. 30, no. 3, pp. 124–130, 1995.

[51] R. A. Yohai, J. D. Bullock, A. A. Aziz, and R. J. Markert, "Survival factors in rhino-orbital-cerebral mucormycosis," *Surv. Ophthalmol.*, vol. 39, no. 1, pp. 3–22, 1994.

[52] S. J. A. Yuen, "Idiopathic Orbital Inflammation," *Arch. Ophthalmol.*, vol. 121, no. 4, p. 491, 2003.

[53] M. Ozgen, Ali, Ariyurek, "Normative measurements of orbital structures using CT.," no. April, pp. 1093–1096, 1998.

[54] A. G. Lee, M. C. Johnson, B. A. Policeni, and W. R. K. Smoker, "Imaging for neuro-ophthalmic and orbital disease–a review," *Clin. Experiment. Ophthalmol.*, vol. 37, no. 1, pp. 30–53, 2009.

[55] S. Tian, Y. Nishida, B. Isberg, and G. Lennerstrand, "MRI measurements of normal extraocular muscles and other orbital structures.," *Graefes Arch. Clin. Exp. Ophthalmol.*, vol. 238, no. 5, pp. 393–404, May 2000.

[56] J. Hoffmann, H.-J. Huppertz, C. Schmidt, H. Kunte, L. Harms, R. Klingebiel, and E. Wiener, "Morphometric and volumetric MRI changes in idiopathic intracranial hypertension," *Cephalalgia*, p. 0333102413484095, 2013.

[57] J. Hoffmann, C. Schmidt, H. Kunte, R. Klingebiel, L. Harms, H.-J. Huppertz, L. Lüdemann, and E. Wiener, "Volumetric assessment of optic nerve sheath and hypophysis in idiopathic intracranial hypertension," *Am. J. Neuroradiol.*, vol. 35, no. 3, pp. 513–518, 2014.

[58] B. Shofty, L. Ben-Sira, S. Constantini, S. Freedman, and A. Kesler, "Optic nerve sheath diameter on MR imaging: establishment of norms and comparison of pediatric patients with idiopathic intracranial hypertension with healthy controls," *Am. J. Neuroradiol.*, vol. 33, no. 2, pp. 366–369, 2012.

[59] D. I. Friedman and D. M. Jacobson, “Diagnostic criteria for idiopathic intracranial hypertension,” *Neurology*, vol. 59, no. 10, pp. 1492–1495, 2002.

[60] H. Sotoudeh, M. Bowerson, M. Parsons, G. VanStavern, R. Viets, C. Hildebolt, and A. Sharma, “Effect of spatial resolution of T2-weighted imaging on diagnostic efficacy of MRI in detection of papilledema,” *Am. J. Roentgenol.*, vol. 204, no. 3, pp. 602–607, 2015.

[61] Y.-C. C. Chang, N. Alperin, A. M. Bagci, S. H. Lee, P. R. Rosa, G. Giovanni, and B. L. Lam, “Relationship Between Optic Nerve Protrusion Measured by OCT and MRI and Papilledema SeverityOptic NP Measured by OCT, MRI, and Papilledema Severity,” *Invest. Ophthalmol. Vis. Sci.*, vol. 56, no. 4, pp. 2297–2302, 2015.

[62] E. Weis, M. K. S. Heran, A. Jhamb, A. K. Chan, J. P. Chiu, M. C. Hurley, and J. Rootman, “Quantitative computed tomographic predictors of compressive optic neuropathy in patients with thyroid orbitopathy: a volumetric analysis.,” *Ophthalmology*, vol. 119, no. 10, pp. 2174–8, Oct. 2012.

[63] E. S. Hallin and S. E. Feldon, “Graves’ ophthalmopathy: II. Correlation of clinical signs with measures derived from computed tomography.,” *Br. J. Ophthalmol.*, vol. 72, no. 9, pp. 678–82, Sep. 1988.

[64] E. Sverker Hallin ’and and S. E. Feldon2, “Graves’ ophthalmopathy: I. Simple CT estimates of extraocular muscle volume,” *Br J Ophthalmol*, vol. 72, pp. 674–677, 1988.

[65] W. Müller-Forell and G. J. Kahaly, “Neuroimaging of Graves’ orbitopathy,” *Best Pract. Res. Clin. Endocrinol. Metab.*, vol. 26, no. 3, pp. 259–271, 2012.

[66] L. K. Gordon, “Diagnostic dilemmas in orbital inflammatory disease,” *Ocul. Immunol. Inflamm.*, vol. 11, no. 1, pp. 3–15, 2003.

[67] H. S. Eustis, M. F. Mafee, C. Walton, and J. Mondonca, “MR imaging and CT of orbital infections and complications in acute rhinosinusitis,” *Radiol. Clin. North Am.*, vol. 36, no. 6, pp. 1165–1183, 1998.

[68] A. L. Weber, L. V. Romo, and N. R. Sabates, “Pseudotumor of the orbit: clinical, pathologic, and radiologic evaluation,” *Radiol. Clin. North Am.*, vol. 37, no. 1, pp. 151–168, 1999.

[69] K. W. Preidler, W. Koele, and D. H. Szolar, “Bilateral orbital pseudotumour due to Ormond’s disease: MR imaging and CT® findings,” *Neuroradiology*, vol. 44, pp. 272–274, 2002.

[70] S. J. Hickman, A. T. Toosy, S. J. Jones, D. R. Altmann, K. A. Miszkiel, D. G. MacManus, G. J. Barker, G. T. Plant, A. J. Thompson, and D. H. Miller, “A serial MRI study following optic nerve mean area in acute optic neuritis,” *Brain*, vol. 127, no. 11, pp. 2498–2505, 2004.

[71] C. W. Howard, R. H. Osher, and R. L. Tomsak, “Computed tomographic features in optic neuritis,” *Am. J. Ophthalmol.*, vol. 89, no. 5, pp. 699–702, 1980.

[72] R. L. Harrigan, A. J. Plassard, F. W. Bryan, G. Caires, L. A. Mawn, L. M. Dethrage, S. Pawate, R. L. Galloway, S. A. Smith, and B. A. Landman, “Disambiguating the optic nerve from the surrounding cerebrospinal fluid: Application to MS-related atrophy,” *Magn. Reson. Med.*, vol. 75, no. 1, pp. 414–422, 2016.

[73] M. Becker, K. Masterson, J. Delavelle, M. Viallon, M.-I. Vargas, and C. D. Becker, “Imaging of the optic nerve,” *Eur. J. Radiol.*, vol. 74, no. 2, pp. 299–313, 2010.

[74] D. H. Miller, M. R. Newton, J. C. Van der Poel, E. Du Boulay, A. M. Halliday, B. E. Kendall, G. Johnson, D. G. MacManus, I. F. Moseley, and W. I. McDonald, "Magnetic resonance imaging of the optic nerve in optic neuritis," *Neurology*, vol. 38, no. 2, p. 175, 1988.

[75] S. A. Trip, P. G. Schlottmann, S. J. Jones, W.-Y. Li, D. F. Garway-Heath, A. J. Thompson, G. T. Plant, and D. H. Miller, "Optic nerve atrophy and retinal nerve fibre layer thinning following optic neuritis: evidence that axonal loss is a substrate of MRI-detected atrophy," *Neuroimage*, vol. 31, no. 1, pp. 286–293, 2006.

[76] J. Xu, S. Sun, R. T. Naismith, A. Z. Snyder, A. H. Cross, and S. Song, "Assessing optic nerve pathology with diffusion MRI: from mouse to human," *NMR Biomed.*, vol. 21, no. 9, pp. 928–940, 2008.

[77] D. K. Binder, J. C. Horton, M. T. Lawton, and M. W. McDermott, "Idiopathic intracranial hypertension," *Neurosurgery*, vol. 54, no. 3, pp. 538–552, 2004.

[78] P. Laurberg, D. C. Berman, I. Bülow Pedersen, S. Andersen, and A. Carlé, "Double vision is a major manifestation in moderate to severe graves' orbitopathy, but it correlates negatively with inflammatory signs and proptosis," *J. Clin. Endocrinol. Metab.*, vol. 100, no. 5, pp. 2098–2105, 2015.

[79] G. Ophthalmopathy, "Graves' Ophthalmopathy," 2010.

[80] H.-C. Lin, C.-W. Chien, C.-C. Hu, and J.-D. Ho, "Comparison of comorbid conditions between open-angle glaucoma patients and a control cohort: a case-control study," *Ophthalmology*, vol. 117, no. 11, pp. 2088–2095, 2010.

[81] L. K. Gordon, "Orbital inflammatory disease: a diagnostic and therapeutic challenge," *Eye*, vol. 20, no. 10, pp. 1196–1206, 2006.

[82] R. Datteri, A. J. Asman, B. A. Landman, and B. M. Dawant, "Applying the algorithm assessing quality using image registration circuits (aquirc) to multi-atlas segmentation," in *SPIE Medical Imaging*, 2014, p. 90341F–90341F.

[83] S. Chaganti, K. P. Nabar, and B. Landman, "pyPheWAS." .

[84] 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 *SPIE Medical Imaging*, 2017, p. 101380F–101380F.

[85] J. C. Denny, M. D. Ritchie, M. A. Basford, J. M. Pulley, L. Bastarache, K. Brown-Gentry, D. Wang, D. R. Masys, D. M. Roden, and D. C. Crawford, "PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations," *Bioinformatics*, vol. 26, no. 9, pp. 1205–1210, 2010.

[86] I. Danciu, J. D. Cowan, M. Basford, X. Wang, A. Saip, S. Osgood, J. Shirey-Rice, J. Kirby, and P. A. Harris, "Secondary use of clinical data: The Vanderbilt approach," *J. Biomed. Inform.*, vol. 52, pp. 28–35, 2014.

[87] R. A. Brooks, "A quantitative theory of the Hounsfield unit and its application to dual energy scanning.," *J. Comput. Assist. Tomogr.*, vol. 1, no. 4, pp. 487–493, 1977.

[88] W. H. Organization, *International statistical classification of diseases and related health problems*, vol. 1. World Health Organization, 2004.

[89] A. M. Association, *Current procedural terminology: CPT*. American Medical Association, 2007.

[90] M. P. Kawa, A. Machalińska, G. Wilk, and B. Machalinski, "Graves' Ophthalmopathy Imaging Evaluation," *Thyroid Disord. - Focus Hyperthyroidism*, 2014.

[91] A. C. P. Gonçalves, L. N. Silva, E. M. M. S. Gebrim, S. Matayoshi, and M. L. R. Monteiro, "Predicting dysthyroid optic neuropathy using computed tomography volumetric analyses of orbital structures," *Clinics*, vol. 67, pp. 891–896, 2012.

[92] R. D. Rondinelli, E. Genovese, and C. R. Brigham, *Guides to the evaluation of permanent impairment*. American Medical Association, 2008.

[93] D. Rueckert and J. A. Schnabel, "Medical image registration," in *Biomedical image processing*, Springer, 2010, pp. 131–154.

[94] M. Holden, "A review of geometric transformations for nonrigid body registration," *IEEE Trans. Med. Imaging*, vol. 27, no. 1, p. 111, 2008.

[95] K. Rohr, H. S. Stiehl, R. Sprengel, T. M. Buzug, J. Weese, and M. H. Kuhn, "Landmark-based elastic registration using approximating thin-plate splines," *IEEE Trans. Med. Imaging*, vol. 20, no. 6, pp. 526–534, 2001.

[96] J. M. Fitzpatrick and J. B. West, "The distribution of target registration error in rigid-body point-based registration," *IEEE Trans. Med. Imaging*, vol. 20, no. 9, pp. 917–927, 2001.

[97] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee, "A reproducible evaluation of ANTs similarity metric performance in brain image registration," *Neuroimage*, vol. 54, no. 3, pp. 2033–2044, 2011.

[98] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE Trans. Med. Imaging*, vol. 32, no. 7, pp. 1153–1190, 2013.

[99] D. L. Pham, C. Xu, and J. L. Prince, "Current methods in medical image segmentation," *Annu. Rev. Biomed. Eng.*, vol. 2, no. 1, pp. 315–337, 2000.

[100] J. Eugenio and M. R. Sabuncu, "Multi-atlas segmentation of biomedical images : A survey," *Med. Image Anal.*, vol. 24, no. 1, pp. 205–219, 2015.

[101] G. K. Savova, J. J. Masanz, P. V Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications," *J. Am. Med. Informatics Assoc.*, vol. 17, no. 5, pp. 507–513, 2010.

[102] J. C. Kirby, P. Speltz, L. V Rasmussen, M. Basford, O. Gottesman, P. L. Peissig, J. A. Pacheco, G. Tromp, J. Pathak, and D. S. Carrell, "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.

[103] K. P. Liao, T. Cai, G. K. Savova, S. N. Murphy, E. W. Karlson, and A. N. Ananthakrishnan, "Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ, 350, h1885." 2015.

[104] G. Surján, "Questions on validity of International Classification of Diseases-coded diagnoses," *Int. J. Med. Inform.*, vol. 54, no. 2, pp. 77–95, 1999.

[105] T. M. Mitchell, "Machine learning. WCB." McGraw-Hill Boston, MA:, 1997.

[106] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. .

[107] A. Ghodsi, "Dimensionality Reduction A Short Tutorial," 2006.

[108] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Stat. Soc. Ser. B (Statistical Methodol.*, vol. 67, no. 2, pp. 301–320, 2005.

[109] I. El-Naqa, Y. Yang, M. N. Wernick, N. P. Galatsanos, and R. M. Nishikawa, "A support vector machine approach for detection of microcalcifications," *IEEE Trans. Med. Imaging*, vol. 21, no. 12, pp. 1552–1563, 2002.

[110] L. Wei, Y. Yang, R. M. Nishikawa, and Y. Jiang, "A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications," *IEEE Trans. Med. Imaging*, vol. 24, no. 3, pp. 371–380, 2005.

[111] J. Ramírez, J. M. Górriz, F. Segovia, R. Chaves, D. Salas-Gonzalez, M. López, I. Álvarez, and P. Padilla, "Computer aided diagnosis system for the Alzheimer's disease based on partial least squares and random forest SPECT image classification," *Neurosci. Lett.*, vol. 472, no. 2, pp. 99–103, 2010.

[112] S. Wang and R. M. Summers, "Machine learning and radiology," *Med. Image Anal.*, vol. 16, no. 5, pp. 933–951, 2012.

[113] H. Greenspan, B. Van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1153–1159, 2016.

[114] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[115] P. Agrawal, R. T. Whitaker, and S. Y. Elhabian, "Learning Deep Features for Automated Placement of Correspondence Points on Ensembles of Complex Shapes," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2017, pp. 185–193.

[116] Y. Guo, G. Wu, L. A. Commander, S. Szary, V. Jewells, W. Lin, and D. Shen, "Segmenting hippocampus from infant brains by sparse patch matching with deep-learned features," in *Medical image computing and computer-assisted intervention: MICCAI... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2014, vol. 17, no. 0 2, p. 308.

[117] H. R. Roth, L. Lu, A. Seff, K. M. Cherry, J. Hoffman, S. Wang, J. Liu, E. Turkbey, and R. M. Summers, "A new 2.5 D representation for lymph node detection using random sets of deep convolutional neural network observations," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2014, pp. 520–527.

[118] B. Q. Huynh, H. Li, and M. L. Giger, “Digital mammographic tumor classification using transfer learning from deep convolutional neural networks,” *J. Med. Imaging*, vol. 3, no. 3, p. 34501, 2016.

[119] A. Akselrod-Ballin, L. Karlinsky, S. Alpert, S. Hasoul, R. Ben-Ari, and E. Barkan, “A region based convolutional network for tumor detection and classification in breast mammography,” in *International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, 2016, pp. 197–205.

[120] S. Hwang, H.-E. Kim, J. Jeong, and H.-J. Kim, “A novel approach for tuberculosis screening based on deep convolutional neural networks,” *Med. Imaging*, vol. 9785, p. 97852W–1, 2016.

[121] H.-I. Suk and D. Shen, “Deep learning-based feature representation for AD/MCI classification,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2013, pp. 583–590.

[122] Q. Dou, H. Chen, L. Yu, J. Qin, and P.-A. Heng, “Multilevel Contextual 3-D CNNs for False Positive Reduction in Pulmonary Nodule Detection,” *IEEE Trans. Biomed. Eng.*, vol. 64, no. 7, pp. 1558–1567, 2017.

[123] J. Hauke and T. Kossowski, “Comparison of values of Pearson’s and Spearman’s correlation coefficients on the same sets of data,” *Quaest. Geogr.*, vol. 30, no. 2, pp. 87–93, 2011.

[124] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: a practical and powerful approach to multiple testing,” *J. R. Stat. Soc. Ser. B*, pp. 289–300, 1995.

[125] J. D. Blume, L. D. McGowan, W. D. Dupont, and R. A. Greevy Jr, “Second-generation p-values: Improved rigor, reproducibility, & transparency in statistical analyses,” *PLoS One*, vol. 13, no. 3, p. e0188299, 2018.

[126] A. C. P. Gonçalves, L. N. Silva, E. M. M. S. Gebrim, and M. L. R. Monteiro, “Quantification of orbital apex crowding for screening of dysthyroid optic neuropathy using multidetector CT,” *Am. J. Neuroradiol.*, vol. 33, no. 8, pp. 1602–1607, 2012.

[127] M. Al-Bakri, A. K. Rasmussen, C. Thomsen, and P. B. Toft, “Orbital Volumetry in Graves’ Orbitopathy: Muscle and Fat Involvement in relation to Dysthyroid Optic Neuropathy.,” *ISRN Ophthalmol.*, vol. 2014, p. 435276, 2014.

[128] B. da Rocha Lima and J. D. Perry, “Superior Ophthalmic Vein Enlargement and Increased Muscle Index in Dysthyroid Optic Neuropathy.,” *Ophthal. Plast. Reconstr. Surg.*, vol. XXX, no. Xxx, pp. 1–3, 2013.

[129] A. C. P. Gonçalves, L. N. Silva, E. M. M. S. Gebrim, S. Matayoshi, and M. L. R. Monteiro, “Predicting dysthyroid optic neuropathy using computed tomography volumetric analyses of orbital structures.,” *Clinics (Sao Paulo).*, vol. 67, no. 8, pp. 891–6, 2012.

[130] J. S. Kennerdell, a E. Rosenbaum, and M. H. El-Hoshy, “Apical optic nerve compression of dysthyroid optic neuropathy on computed tomography.,” *Arch. Ophthalmol.*, vol. 99, no. 5, pp. 807–809, 1981.

[131] S. E. Feldon, J. M. Weiner, and A. Ophthalmol, “Clinical Significance of Extraocular Muscle Volumes in Graves ’ Ophthalmopathy,” pp. 10–13, 2014.

[132] S. E. Feldon, C. P. Lee, S. K. Muramatsu, and J. M. Weiner, "Quantitative computed tomography of Graves' ophthalmopathy. Extraocular muscle and orbital fat in development of optic neuropathy.," *Arch. Ophthalmol. (Chicago, Ill. 1960)*, vol. 103, no. 2, pp. 213–5, 1985.

[133] L. Barrett, H. J. Glatt, R. M. Burde, and M. H. Gado, "Optic nerve dysfunction in thyroid eye disease: CT.," *Radiology*, vol. 167, no. 2, pp. 503–507, 1988.

[134] J. a Giaconi, M. Kazim, T. Rho, and C. Pfaff, "CT scan evidence of dysthyroid optic neuropathy.," *Ophthal. Plast. Reconstr. Surg.*, vol. 18, no. 3, pp. 177–82, 2002.

[135] M. L. R. Monteiro, A. C. P. Gonçalves, C. T. M. Silva, J. P. Moura, C. S. Ribeiro, and E. M. M. S. Gebrim, "Diagnostic ability of barrett's index to detect dysthyroid optic neuropathy using multidetector computed tomography," *Clinics*, vol. 63, no. 3, pp. 301–306, 2008.

[136] L. L. Chan, H. E. Tan, S. Fook-Chong, T. H. Teo, L. H. Lim, and L. L. Seah, "Graves ophthalmopathy: The bony orbit in optic neuropathy, its apical angular capacity, and impact on prediction of risk," *Am. J. Neuroradiol.*, vol. 30, no. 3, pp. 597–602, 2009.

[137] E. Weis, M. S. Heran, A. Jhamb, and et al., "Clinical and soft-tissue computed tomographic predictors of dysthyroid optic neuropathy: Refinement of the constellation of findings at presentation," *Arch. Ophthalmol.*, vol. 129, no. 10, pp. 1332–1336, 2011.

[138] R. L. Harrigan, S. Panda, A. J. Asman, K. M. Nelson, S. Chaganti, M. P. DeLisi, B. C. W. Yvernault, S. A. Smith, R. L. Galloway, and L. A. Mawn, "Robust optic nerve segmentation on clinically acquired computed tomography," *J. Med. Imaging*, vol. 1, no. 3, p. 34006, 2014.

[139] M. Gensheimer, A. Cmelak, K. Niermann, and B. M. Dawant, "Automatic delineation of the optic nerves and chiasm on CT images," 2007, vol. 6512, pp. 651210–651216.

[140] A. Isambert, F. Dhermain, F. Bidault, O. Commowick, P.-Y. Bondiau, G. Malandain, and D. Lefkopoulos, "Evaluation of an atlas-based automatic segmentation software for the delineation of brain organs at risk in a radiation therapy clinical context," *Radiother. Oncol.*, vol. 87, no. 1, pp. 93–99, 2008.

[141] R. A. Heckemann, J. V Hajnal, P. Aljabar, D. Rueckert, and A. Hammers, "Automatic anatomical brain MRI segmentation combining label propagation and decision fusion," *Neuroimage*, vol. 33, no. 1, pp. 115–126, 2006.

[142] R. L. Harrigan, B. C. Yvernault, B. D. Boyd, S. M. Damon, K. D. Gibney, B. N. Conrad, N. S. Phillips, B. P. Rogers, Y. Gao, and B. A. Landman, "Vanderbilt University Institute of Imaging Science Center for Computational Imaging XNAT: A multimodal data archive and processing environment," *Neuroimage*, vol. 124, pp. 1097–1101, 2016.

[143] D. Marcus, T. Olsen, M. Ramaratnam, and R. Buckner, "The extensible neuroimaging archive toolkit," *Neuroinformatics*, vol. 5, no. 1, pp. 11–33, 2007.

[144] A. J. Asman and B. a Landman, "Non-local statistical label fusion for multi-atlas segmentation.," *Med. Image Anal.*, vol. 17, no. 2, pp. 194–208, Feb. 2013.

[145] A. Ozgen and M. Ariyurek, "Normative measurements of orbital structures using CT," *AJR. Am. J. Roentgenol.*, vol. 170, no. 4, pp. 1093–1096, 1998.

[146] Z. Szucs-Farkas, J. Toth, E. Balazs, L. Galuska, K. D. Burman, Z. Karanyi, A. Leovey, and E. V. Nagy, "Using morphologic parameters of extraocular muscles for diagnosis and follow-up

of graves' ophthalmopathy: Diameters, areas, or volumes?," *Am. J. Roentgenol.*, vol. 179, no. 4, pp. 1005–1010, 2002.

[147] E. Pearce and H. Bridge, "Is orbital volume associated with eyeball and visual cortex volume in humans?," *Ann. Hum. Biol.*, vol. 40, no. 6, pp. 531–540, 2013.

[148] P. a Rubin, L. M. Watkins, S. Rumelt, F. C. Sutula, and R. L. Dallow, "Orbital computed tomographic characteristics of globe subluxation in thyroid orbitopathy.," *Ophthalmology*, vol. 105, no. 11, pp. 2061–4, Nov. 1998.

[149] R. G. Peyster, E. D. Hoover, B. L. Hershey, and M. E. Haskin, "High-resolution CT of lesions of the optic nerve," *Am. J. Roentgenol.*, vol. 140, no. 5, pp. 869–874, 1983.

[150] M. R. Sabuncu, B. T. T. Yeo, K. Van Leemput, B. Fischl, and P. Golland, "A generative model for image segmentation based on label fusion.," *IEEE Trans. Med. Imaging*, vol. 29, no. 10, pp. 1714–29, Oct. 2010.

[151] C. D. Good, R. I. Scahill, N. C. Fox, J. Ashburner, K. J. Friston, D. Chan, W. R. Crum, M. N. Rossor, and R. S. J. Frackowiak, "Automatic differentiation of anatomical patterns in the human brain: validation with studies of degenerative dementias," *Neuroimage*, vol. 17, no. 1, pp. 29–46, 2002.

[152] M. Sdika, "Combining atlas based segmentation and intensity classification with nearest neighbor transform and accuracy weighted vote," *Med. Image Anal.*, vol. 14, no. 2, pp. 219–226, 2010.

[153] C. Sjöberg and A. Ahnesjö, "Multi-atlas based segmentation using probabilistic label fusion with adaptive weighting of image similarity measures," *Comput. Methods Programs Biomed.*, vol. 110, no. 3, pp. 308–319, 2013.

[154] R. D. Datteri, A. J. Asman, B. A. Landman, and B. M. Dawant, "Estimation of Registration Accuracy Applied to Multi-Atlas Segmentation."

[155] R. D. Datteri, Y. Liu, P.-F. D'Haese, and B. M. Dawant, "Validation of a nonrigid registration error detection algorithm using clinical MRI brain data," *IEEE Trans. Med. Imaging*, vol. 34, no. 1, pp. 86–96, 2015.

[156] S. Chaganti, K. Nelson, K. Mundy, Y. Luo, R. L. Harrigan, S. Damon, D. Fabbri, L. Mawn, and B. Landman, "Structural functional associations of the orbit in thyroid eye disease: Kalman filters to track extraocular rectal muscles," in *Proceedings of SPIE--the International Society for Optical Engineering*, 2016, vol. 9784.

[157] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain," *Med. Image Anal.*, vol. 12, no. 1, pp. 26–41, 2008.

[158] P. Aljabar, R. A. Heckemann, A. Hammers, J. V Hajnal, and D. Rueckert, "Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy," *Neuroimage*, vol. 46, no. 3, pp. 726–738, 2009.

[159] Y. Cao, Y. Yuan, X. Li, B. Turkbey, P. L. Choyke, and P. Yan, "Segmenting images by combining selected atlases on manifold," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2011, pp. 272–279.

[160] A. K. H. Duc, M. Modat, K. K. Leung, M. J. Cardoso, J. Barnes, T. Kadir, S. Ourselin, and A. D. N. Initiative, "Using manifold learning for atlas selection in multi-atlas segmentation," *PLoS One*, vol. 8, no. 8, p. e70059, 2013.

[161] P. Aljabar, R. Heckemann, A. Hammers, J. V Hajnal, and D. Rueckert, "Classifier selection strategies for label fusion using large atlas databases," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2007, pp. 523–531.

[162] W. R. Nunery, "Ophthalmic Graves' disease: a dual theory of pathogenesis," *Ophthalmol Clin North Am*, vol. 4, pp. 73–87, 1991.

[163] W. R. Nunery, R. T. Martin, G. W. Heinz, and T. J. Gavin, "The association of cigarette smoking with clinical subtypes of ophthalmic Graves' disease.," *Ophthal. Plast. Reconstr. Surg.*, vol. 9, no. 2, pp. 77–82, 1993.

[164] W. M. Wiersinga, N. I. Regensburg, and M. P. Mourits, "Differential involvement of orbital fat and extraocular muscles in graves' ophthalmopathy.," *Eur. Thyroid J.*, vol. 2, no. 1, pp. 14–21, 2013.

[165] N. I. Regensburg, W. M. Wiersinga, T. T. J. M. Berendschot, P. Potgieser, and M. P. Mourits, "Do subtypes of graves' orbitopathy exist?," *Ophthalmology*, vol. 118, no. 1, pp. 191–196, 2011.

[166] I. B. Hales and F. F. Rundle, "Ocular changes in Graves' disease: a long-term follow-up study.," *Q. J. Med.*, vol. 29, pp. 113–126, 1960.

[167] M. P. Mourits, S. H. C. Lombardo, F. A. van der Sluijs, and S. Fenton, "Reliability of exophthalmos measurement and the exophthalmometry value distribution in a healthy Dutch population and in Graves' patients. An exploratory study," *Orbit*, vol. 23, no. 3, pp. 161–168, 2004.

[168] M. P. H. Mourits, L. Koornneef, W. M. Wiersinga, M. F. Prummel, A. Berghout, and R. Van Der Gaag, "Clinical criteria for the assessment of disease activity in Graves' ophthalmopathy: a novel approach.," *Br. J. Ophthalmol.*, vol. 73, no. 8, pp. 639–644, 1989.

[169] L. Wartofsky, "Classification of sys changes of Graves' disease.," *Thyroid*, vol. 2, pp. 235–236, 1992.

[170] J. Thornton, S. P. Kelly, R. A. Harrison, and R. Edwards, "Cigarette smoking and thyroid eye disease: a systematic review," *Eye*, vol. 21, no. 9, p. 1135, 2007.

[171] B. A. L. Andrew J. Asman, Michael P. DeLisi, Mawn, Robert L. Galloway, "Robust Non-Local Multi-Atlas Segmentation of the Optic Nerve," vol. 4, no. 164, pp. 1–13, 2011.

[172] R. a. Nugent, R. I. Belkin, J. M. Neigel, J. Rootman, W. D. Robertson, J. Spinelli, and D. a. Graeb, "Graves Orbitopathy: Correlation of CT Orbitopathy : and Clinical Correlation," *Radiology*, vol. 177, pp. 675–682, 1990.

[173] J. M. Neigel, J. Rootman, R. I. Belkin, R. A. Nugent, S. M. Drance, C. W. Beattie, and J. A. Spinelli, "Dysthyroid optic neuropathy: the crowded orbital apex syndrome," *Ophthalmology*, vol. 95, no. 11, pp. 1515–1521, 1988.

[174] R. S. Bahn, "Graves' ophthalmopathy," *N. Engl. J. Med.*, vol. 362, no. 8, pp. 726–738, 2010.

[175] N. I. Regensburg, W. M. Wiersinga, T. T. J. M. Berendschot, P. Saeed, and M. P. Mourits, "Effect of smoking on orbital fat and muscle volume in Graves' orbitopathy," *Thyroid*, vol. 21, no. 2, pp. 177–181, 2011.

[176] D. B. Rein, P. Zhang, K. E. Wirth, P. P. Lee, T. J. Hoerger, N. McCall, R. Klein, J. M. Tielsch, S. Vijan, and J. Saaddine, "The economic burden of major adult visual disorders in the United States," *Arch. Ophthalmol.*, vol. 124, no. 12, pp. 1754–1760, 2006.

[177] X. Yao, S. Chagantib, K. P. Nabarb, K. Nelsonc, A. Plassardb, R. L. Harriganc, L. A. Mawnd, and B. A. Landmanb, "Structural-Functional Relationships Between Eye Orbital Imaging Biomarkers and Clinical Visual Assessments," in *SPIE Medical Imaging*, 2017, p. 101331F–101331F.

[178] I. M. Xierali, C.-J. Hsiao, J. C. Puffer, L. A. Green, J. C. B. Rinaldo, A. W. Bazemore, M. T. Burke, and R. L. Phillips, "The rise of electronic health record adoption among family physicians," *Ann. Fam. Med.*, vol. 11, no. 1, pp. 14–19, 2013.

[179] V. Patel, E. Jamoom, C.-J. Hsiao, M. F. Furukawa, and M. Buntin, "Variation in electronic health record adoption and readiness for meaningful use: 2008–2011," *J. Gen. Intern. Med.*, vol. 28, no. 7, pp. 957–964, 2013.

[180] J. C. Denny, L. Bastarache, M. D. Ritchie, R. J. Carroll, R. Zink, J. D. Mosley, J. R. Field, J. M. Pulley, A. H. Ramirez, and E. Bowton, "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. 1102–1111, 2013.

[181] H.CUP, "Clinical Classification Software for Services and Procedures." [Online]. Available: https://www.hcup-us.ahrq.gov/toolssoftware/ccs_svcsproc/ccssvcproc.jsp.

[182] J. Shlens, "A tutorial on principal component analysis," *arXiv Prepr. arXiv1404.1100*, 2014.

[183] H. Abdi and D. Valentin, "Multiple correspondence analysis," *Encycl. Meas. Stat.*, pp. 651–657, 2007.

[184] P. McCullagh, "Generalized linear models," *Eur. J. Oper. Res.*, vol. 16, no. 3, pp. 285–292, 1984.

[185] N. R. Draper, H. Smith, and E. Pownell, *Applied regression analysis*, vol. 3. Wiley New York, 1966.

[186] K. Doi, "Computer-aided diagnosis in medical imaging: historical review, current status and future potential," *Comput. Med. imaging Graph.*, vol. 31, no. 4, pp. 198–211, 2007.

[187] S. Bauer, R. Wiest, L.-P. Nolte, and M. Reyes, "A survey of MRI-based medical image analysis for brain tumor studies," *Phys. Med. Biol.*, vol. 58, no. 13, p. R97, 2013.

[188] S. Chaganti, A. J. Plassard, L. Wilson, M. A. Smith, M. B. Patel, and B. A. Landman, "A Bayesian Framework for Early Risk Prediction in Traumatic Brain Injury," no. Table 1.

[189] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: application to breast MR images," *IEEE Trans. Med. Imaging*, vol. 18, no. 8, pp. 712–721, 1999.

[190] J. Wiens, E. Horvitz, and J. V Guttag, "Patient risk stratification for hospital-associated c. diff as a time-series classification task," in *Advances in Neural Information Processing Systems*, 2012, pp. 467–475.

[191] Z. Shahn, P. Ryan, and D. Madigan, "Predicting health outcomes from high-dimensional longitudinal health histories using relational random forests," *Stat. Anal. Data Min. ASA Data Sci. J.*, vol. 8, no. 2, pp. 128–136, 2015.

[192] R. J. Carroll, W. K. Thompson, A. E. Eyler, A. M. Mandelin, T. Cai, R. M. Zink, J. A. Pacheco, C. S. Boomershine, T. A. Lasko, and H. Xu, "Portability of an algorithm to identify rheumatoid arthritis in electronic health records," *J. Am. Med. Informatics Assoc.*, vol. 19, no. e1, pp. e162–e169, 2012.

[193] S. Chaganti, K. Nabar, and B. Landman, "pyPheWAS," 2017. [Online]. Available: https://github.com/BennettLandman/pyPheWAS.

[194] R. Schmidt, L. J. Launer, L.-G. Nilsson, A. Pajak, S. Sans, K. Berger, M. M. Breteler, M. de Ridder, C. Dufouil, and R. Fuhrer, "Magnetic resonance imaging of the brain in diabetes," *Diabetes*, vol. 53, no. 3, pp. 687–692, 2004.

[195] N. D. D. Group, "Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance," *Diabetes*, vol. 28, no. 12, pp. 1039–1057, 1979.

[196] B. Van Harten, F.-E. de Leeuw, H. C. Weinstein, P. Scheltens, and G. J. Biessels, "Brain imaging in patients with diabetes," *Diabetes Care*, vol. 29, no. 11, pp. 2539–2548, 2006.

[197] V. CPM, "PheWAS Resources." [Online]. Available: https://phewascatalog.org/phecodes.

[198] S. M. Resnick, D. L. Pham, M. A. Kraut, A. B. Zonderman, and C. Davatzikos, "Longitudinal magnetic resonance imaging studies of older adults: a shrinking brain," *J. Neurosci.*, vol. 23, no. 8, pp. 3295–3301, 2003.

[199] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159, 1997.

[200] M. P. Heinrich, M. Jenkinson, M. Brady, and J. A. Schnabel, "MRF-based deformable registration and ventilation estimation of lung CT," *IEEE Trans. Med. Imaging*, vol. 32, no. 7, pp. 1239–1248, 2013.

[201] Y. Huo, A. J. Plassard, A. Carass, S. M. Resnick, D. L. Pham, J. L. Prince, and B. A. Landman, "Consistent cortical reconstruction and multi-atlas brain segmentation," *Neuroimage*, vol. 138, pp. 197–210, 2016.

[202] S. Ourselin, A. Roche, G. Subsol, X. Pennec, and N. Ayache, "Reconstructing a 3D structure from serial histological sections," *Image Vis. Comput.*, vol. 19, no. 1–2, pp. 25–31, 2001.

[203] A. C. Evans, D. L. Collins, S. R. Mills, 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., 1993 IEEE Conference Record.*, 1993, pp. 1813–1817.

[204] A. Klein, T. Dal Canton, S. S. Ghosh, B. Landman, J. Lee, and A. Worth, "Open labels: online feedback for a public resource of manually labeled brain images," in *16th Annual Meeting for the Organization of Human Brain Mapping*, 2010.

[205] J. Friedman, T. Hastie, and R. Tibshirani, "glmnet: Lasso and elastic-net regularized generalized linear models," *R Packag. version*, vol. 1, no. 4, 2009.

[206] B. Efron and G. Gong, "A leisurely look at the bootstrap, the jackknife, and cross-validation," *Am. Stat.*, vol. 37, no. 1, pp. 36–48, 1983.

[207] F. E. Harrell Jr, *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.

[208] P. Poulsen, K. Ohm Kyvik, A. Vaag, and H. Beck-Nielsen, "Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance–a population-based twin study," *Diabetologia*, vol. 42, no. 2, pp. 139–145, 1999.

[209] M. C. Villadolid, N. Yokoyama, M. Izumi, T. Nishikawa, H. Kimura, K. Ashizawa, T. Kiriyama, M. Uetani, and S. Nagataki, "Untreated Graves' disease patients without clinical ophthalmopathy demonstrate a high frequency of extraocular muscle (EOM) enlargement by magnetic resonance," *J. Clin. Endocrinol. Metab.*, vol. 80, no. 9, pp. 2830–2833, 1995.

[210] X. Wang, G. Hripcsak, M. Markatou, and C. Friedman, "Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study," *J. Am. Med. Informatics Assoc.*, vol. 16, no. 3, pp. 328–337, 2009.

[211] P. M. Coloma, M. J. Schuemie, G. Trifirò, R. Gini, R. Herings, J. Hippisley-Cox, G. Mazzaglia, C. Giaquinto, G. Corrao, and L. Pedersen, "Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project," *Pharmacoepidemiol. Drug Saf.*, vol. 20, no. 1, pp. 1–11, 2011.

[212] R. L. Richesson, W. E. Hammond, M. Nahm, D. Wixted, G. E. Simon, J. G. Robinson, A. E. Bauck, D. Cifelli, M. M. Smerek, J. Dickerson, R. L. Laws, R. A. Madigan, S. A. Rusincovitch, C. Kluchar, and R. M. Califf, "Electronic health records based phenotyping in next-generation clinical trials: A perspective from the NIH health care systems collaboratory," *J. Am. Med. Informatics Assoc.*, vol. 20, no. E2, 2013.

[213] 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.

[214] T. A. Kellogg, T. Swan, D. A. Leslie, H. Buchwald, and S. Ikramuddin, "Patterns of readmission and reoperation within 90 days after Roux-en-Y gastric bypass," *Surg. Obes. Relat. Dis.*, vol. 5, no. 4, pp. 416–423, 2009.

[215] S. D. Turner, R. L. Berg, J. G. Linneman, P. L. Peissig, D. C. Crawford, J. C. Denny, D. M. Roden, C. A. McCarty, M. D. Ritchie, and R. A. Wilke, "Knowledge-driven multi-locus analysis reveals gene-gene interactions influencing HDL cholesterol level in two independent EMR-linked biobanks," *PLoS One*, vol. 6, no. 5, 2011.

[216] O. Gottesman, H. Kuivaniemi, G. Tromp, W. A. Faucett, R. Li, T. A. Manolio, S. C. Sanderson, J. Kannry, R. Zinberg, M. A. Basford, M. Brilliant, D. J. Carey, R. L. Chisholm, C. G. Chute, J. J. Connolly, D. Crosslin, J. C. Denny, C. J. Gallego, J. L. Haines, H. Hakonarson, J. Harley, G. P. Jarvik, I. Kohane, I. J. Kullo, E. B. Larson, C. McCarty, M. D. Ritchie, D. M. Roden, M. E. Smith, E. P. Böttinger, and M. S. Williams, "The Electronic Medical Records and Genomics (eMERGE) Network: Past, present, and future," *Genet. Med.*, vol. 15, no. 10, pp. 761–771, 2013.

[217] E. A. Engels, R. Parsons, C. Besson, L. M. Morton, L. Enewold, W. Ricker, E. L. Yanik, H. Arem, A. A. Austin, and R. M. Pfeiffer, "Comprehensive evaluation of medical conditions associated with risk of non-Hodgkin lymphoma using Medicare claims (' MedWAS')," *Cancer Epidemiol. Prev. Biomarkers*, p. cebp-0212, 2016.

[218] D. A. Hanauer, D. R. Rhodes, and A. M. Chinnaiyan, "Exploring clinical associations using '-omics' based enrichment analyses," *PLoS One*, vol. 4, no. 4, p. e5203, 2009.

[219] F. S. Roque, P. B. Jensen, H. Schmock, M. Dalgaard, M. Andreatta, T. Hansen, K. Søeby, S. Bredkjær, A. Juul, and T. Werge, "Using electronic patient records to discover disease correlations and stratify patient cohorts," *PLoS Comput. Biol.*, vol. 7, no. 8, p. e1002141, 2011.

[220] A. B. Holmes, A. Hawson, F. Liu, C. Friedman, H. Khiabanian, and R. Rabadan, "Discovering disease associations by integrating electronic clinical data and medical literature," *PLoS One*, vol. 6, no. 6, p. e21132, 2011.

[221] 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, 2013.

[222] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. De Bakker, and M. J. Daly, "PLINK: a tool set for whole-genome association and population-based linkage analyses," *Am. J. Hum. Genet.*, vol. 81, no. 3, pp. 559–575, 2007.

[223] B. L. Browning, "PRESTO: rapid calculation of order statistic distributions and multiple-testing adjusted P-values via permutation for one and two-stage genetic association studies," *BMC Bioinformatics*, vol. 9, no. 1, p. 309, 2008.

[224] P. Kraft, E. Zeggini, and J. P. A. Ioannidis, "Replication in genome-wide association studies," *Stat. Sci. A Rev. J. Inst. Math. Stat.*, vol. 24, no. 4, p. 561, 2009.

[225] R. M. Cantor, K. Lange, and J. S. Sinsheimer, "Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application," *Am. J. Hum. Genet.*, vol. 86, no. 1, pp. 6–22, 2010.

[226] A. R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.," in *Proceedings of the AMIA Symposium*, 2001, p. 17.