DETERMINING THE USE OF ELECTRONIC MEDICAL RECORDS IN GENETIC STUDIES OF MUTLIPLE SCLEROSIS

By

Mary Feller Davis

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

Human Genetics

December, 2013

Nashville, Tennessee

Approved:

William S. Bush, Ph.D., M.S.

Jonathan L. Haines, Ph.D.

Subramaniam Sriram, M.B., B.S.

Joshua C. Denny, M.D., M.S.

Thomas M. Aune, Ph.D.

To my best friend and husband, Ryan

and my beautiful daughter, Maggie

ACKNOWLEDGEMENTS

This study was supported by grants NS032830 (to Dr. Jonathan Haines), LM010685 (to Dr. Joshua Denny), P32GM080178 (to Vanderbilt University) and UL1TR000445 (to Vanderbilt CTSA). The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of the National Center for Advancing Translation Sciences or the National Institutes of Health. Genotyping was performed by the DNA Resources Core at Vanderbilt University and genotype calling was performed by the Wellcome Trust Sanger Institute. The Center for Human Genetics Research core resources have been invaluable.

There are many people who have contributed significantly to this work, both in scientific efforts and as support in other aspects of life. I am thankful for all who have contributed and sincerely apologize if I have missed someone.

I would like to thank the patients at Vanderbilt University Medical Center for providing us with this exciting research opportunity. I would also like to thank the Vanderbilt Multiple Sclerosis Clinic for shedding further light on the EMR at VUMC and walking me through common clinical practices.

I am grateful for my Ph.D. advisor, Dr. Jonathan Haines, for his mentoring and for believing in my potential enough to allow me to travel back and forth between Tennessee and Texas for the second half of my Ph.D. He has provided me with exceptional opportunities to attend and present at conferences and to gain experience in large collaborations.

I am grateful to the other members of my Ph.D. committee—Drs. William Bush, Joshua Denny, Subramaniam Sriram, and Thomas Aune. Dr. Bush continually expands my view on what can be done with the data available and is always willing to discuss new ideas. Dr. Denny has spent numerous hours tutoring me in the basics of

programming and has painstakingly gone through code I have written line by line to assist me. Dr. Sriram has been an excellent teacher and shown great patience in helping me gain an accurate understanding of the clinical aspects of multiple sclerosis. The time spent in his clinic gave me insights I could not have gained in other way. Dr. Aune has been very gracious in sharing his expertise with me. His comments in my committee meetings were always greatly appreciated, not least for their directness and applicability to the issue at hand.

It is not possible to enumerate all of the ways the members of the Haines lab have benefitted me and this project. Dr. Nathalie Schnetz-Boutaud and Ping Mayo have taken me under their wings and provided a safe shelter where I could grow as a scientist. The postdoctoral fellows (Drs. William Bush, Brian Yaspan, and Jessica Cooke Bailey) and students (Drs. Rebecca Zuvich, Anna Cummings, and Olivia Veatch, Joshua Hoffman, and Laura D'Aoust) with whom I have overlapped have been wonderful to work with. Dr. Cummings was a very thorough and patient mentor when I joined the lab and is a wonderful friend. Dr. Veatch has been my comrade-in-arms from our first day in the Interdisciplinary Graduate Program at Vanderbilt and we have taken turns at encouraging one another through the challening times. When Laura joined the lab, she quickly became my go-to person every time my brain stalled, and she never let me down.

The support of family has been critical. Unwavering support from my family and my husband's family has given me the confidence I needed to pursue a PhD and then to continue persevering regardless of the obstacles. My dad, Dr. David Feller, has been an inspiration to me and I am grateful for his influence. The unending from encouragement from both my mom, Jolyn Feller, and my dad never ceases to motivate me. I have turned to family at every step of the way. The joy of seeing the happiness of my daughter, Maggie, always renewed me with energy and helped me find a smile on the most difficult days.

My husband, Ryan, has been with me at every point. His loyalty, support, and love has never faltered and he is always there for me to lean on. He has read every paper I have written, listened to practice talks, and helped me with innumerable hours of preparation for my qualifying exams. And throughout it all, he soaks it in and never ceases to want to know more of what I am doing. When my enthusiasm for genetics is lacking, he lets me borrow his.

TABLE OF CONTENTS

Page

Chapter

Appendix

LIST OF TABLES

LIST OF FIGURES

Figure Page

LIST OF ABBREVIATIONS

AAO	age at onset
ANA	anti-nuclear antibodies
BioVU	Vanderbilt biobank
BP	base pair position
BS	brain stem
CHR	chromosome
CNS	central nervous system
Coeff	coefficient
CSF	cerebrospinal fluid
CUI	concept unique identifiers
DZ	dizygotic twin pairs
EBV	Epstein-Barr virus
EDSS	Expanded Disability Status Scale
EMR	electronic medical record
GWAS	genome-wide association study
HLA	Human Leukocyte Antigen
IC	ImmunoChip
ICD	International Classification of Diseases
IMSGC	International Multiple Sclerosis Genetics Consortium
IRB	Institutional Review Board
IVMP	intravenous methylprednisone
kb	kilobase
LD	linkage disequilibrium
MA	minor allele

MAF	minor allele frequency
MHC	major histocompatibility complex
MRI	magnetic resonance imaging
MS	multiple sclerosis
MSFC	Multiple Sclerosis Functional Composite
MSSS	Multiple Sclerosis Severity Score
MZ	monozygotic twin pairs
n	number of individuals
NMO	neuromyelitis optica
NOS	not otherwise specified
ON	optic nerve
OR	odds ratio
P	p-value
PL	Problem List
PML	progressive multifocal leukoencephalopathy
PPMS	primary progressive multiple sclerosis
PPV	positive predictive value
PRMS	progressive-relapsing multiple sclerosis
QC	quality control
RA	risk allele
RAF	risk allele frequency
RRMS	relapsing-remitting multiple sclerosis
SC	spinal cord
SCID	severe combined immunodeficiency
SD	Synthetic Derivative
SNP	single nucleotide polymorphism

SPMS	secondary progressive multiple sclerosis
UK	United Kingdom
VU	Vanderbilt University
VUMC	Vanderbilt University Medical Center
WTCCC	Wellcome Trust Case Control Consortium
WTCCC2	Wellcome Trust Case Control Consortium 2
WTSI	Wellcome Trust Sanger Institute

CHAPTER I

INTRODUCTION

Multiple Sclerosis

Disease introduction

Multiple sclerosis (MS) is a demyelinating disease of the central nervous system (CNS). The etiology of MS appears to be complex, involving genetic and environmental factors. Although much effort has gone into understanding these underlying causes and new information has come to light in recent years, there is still much to learn.

MS is the second most common cause of neurologic disability in young adults, next to head trauma.(1;2) Diagnosis generally occurs early in life with the typical age of onset between 20 and 40 years, although juvenile forms do exist and some persons are diagnosed later in life. Females are affected more often than males—historically this gender ratio has been recorded as 2:1. Recent epidemiological studies suggest that the prevalence is increasing in women but not in men, driving the ratio closer to 3:1 in some areas of the world.(3) Prevalence rates vary by area of the world, from 190/100,000 in the north of Scotland(4), 110/100,000 in England and Wales(5), 62/100,000 in Hungary(6), 120/100,000 in western Greece(7), to 90/100,000 in Canada(8). The prevalence of the disease also appears to be increasing in many populations.(3) MS is most common in Caucasians, although other ethnic groups are affected to some degree. There is a relationship between increasing distance from the equator and increased prevalence of the disease.(9;10) Part of this appears to be from a genetic predisposition, although studies of individuals who have moved closer or farther to the equator early in life show the risk of developing MS is more linked to the place where the person spent their early years, as opposed to where their closer genetic relatives are from.

Several risk factors are known in the environment and in genetics.(11-14) Several environmental factors are described here and genetic risk factors are discussed later in this chapter. As discussed above, increased distance from the equator has been associated with risk of developing MS. Low levels of Vitamin D in the serum are associated with overall risk.(15;16) The relationship between distance from the equator, Vitamin D levels, and MS risk is unclear. Both factors appear to contribute risk.

Birth month has previously been associated with MS risk—with higher numbers of individuals with MS born in May, and decreased numbers in November(17), but a recent study suggests the association is due to confounding factors of year and place of birth.(18) Cigarette smoking is strongly correlated with MS(19;20), although this risk begins to diminish after a five year abstinence period.(21) Infections with Epstein-Barr virus (EBV) has been associated with MS in a number of studies.(22;23) Haahr et al. found the association only with late EBV infection; EBV infections occur later in developed countries, which could partly account for the greater incidence of MS in developed countries.(22)

Clinical diagnostic criteria

MS is a clinically defined disease and the diagnostic criteria have evolved over the last 30 years.(24-26) The current diagnostic criteria is the revised McDonald's criteria, requiring demonstration of disease dissemination in space and time. The most common way to meet these criteria is by magnetic resonance imaging (MRI).

Marked changes on an MRI that are highly suggestive of MS are critical to a diagnosis. MS lesions occur in the brain in all patients diagnosed with MS, and in the spinal cord of many patients. Patients who only exhibit lesions in the spinal cord are diagnosed with transverse myelitis, for which there is a fairly high conversion rate to MS in the first 3-5 years after diagnosis.(27) Lesions due to MS must be distinguished from other events that can alter an MRI, most notably cerebrovascular events. As individuals age, the accrual of a small number of lesions is normal. Distinguishing between normal aging changes and MS lesions makes diagnosis more difficult the older the person is—lesions in a younger person are a fairly good indicator of disease. The main MRI sequences in use to gain information regarding changes due to MS are the T1, T2, and T2 flair images. T1 images are best for looking at the anatomy of the brain and brain volume. Brains in patients with MS typically atrophy at a faster rate than in normal patients.(28) MS lesions are more easily seen in T2 images, although the T2 flair (in which the water sequence is inverted) provides the clearest image to view lesions. The actual lesions do not differ visually on images from those of other causes, but there are some characteristic areas for lesions specific to MS, such as lesions in the periventricular area. Many patients' MRIs show an accumulation of lesions around these ventricles over time. Juxtacortical lesions seen in a sagittal slice are often due to MS, as well as Dawson's fingers, which are small projections of lesions up from the corpus callosum. Lesions in the spinal cord (in addition to the brain) are highly indicative of MS.

Once an MS lesion is formed, it will generally be apparent on an MRI for the patient's life; however, a difference in newly formed, active lesions compared to older lesions can be seen by using gadolinium. Gadolinium usually stays in the blood vessels, but in active lesions it leaks into the brain tissue due to breaches in the blood-brain barrier; hence, gadolinium-enhanced lesions typically represent active lesions. Most lesions remain active for six to eight weeks, until the blood-brain barrier heals itself.

For diagnosis, dissemination in space requires lesions in at least two of the four important areas of the CNS (periventricular, juxtacortical, infratentorial, and spinal cord). The requirement for dissemination in time can be met with new active lesions seen on an MRI that were not present on a baseline MRI or the presence of both gadolinium-enhanced and non-enhanced lesions on the same scan.(26) The presence of oligoclonal bands in the cerebrospinal fluid and not in the serum could be used to confirm a diagnosis of MS when data from MRI was not conclusive in previous MS diagnostic criteria.(24) Oligoclonal bands are not part of the current MS diagnostic criteria; however, it is possible they will be included in future diagnostic criteria.(26) The current criteria suggest further research of oligoclonal bands to determine how they should be included.

Another test frequently performed prior to diagnosis of MS is a blood test to check for the presence of anti-nuclear antibodies (ANA), a type of auto-antibody found in patients with autoimmune diseases. The primary disease associated with a positive ANA is systemic lupus erythematosus, although the test is positive in 22.5% of patients with MS.(29) Positive ANA values can be further described by the laboratory based on the binding pattern of the antibodies (speckled, diffuse, etc.), but for MS there is no typical pattern. As with oligoclonal bands, not all patients with MS have a positive result. Also, a positive ANA is less specific for MS than the presence of oligoclonal bands.

Pathology

The underlying pathology of the disease is not fully understood, but research strongly suggests there are autoimmune, inflammatory, and neurodegenerative disease processes involved. Patients with MS have characteristic lesions in the brain and spinal cord. In the formation of these lesions, the blood-brain barrier is compromised, leading to an influx of cells into the brain tissue in an inflammatory response, and eventual neurodegeneration of axons occurs. Also occurring in the formation of the lesions is demyelination as the myelin sheath surrounding axons in the CNS is targeted for destruction, although whether this target is direct or indirect is unknown. There appears to be some level of remyelination that occurs in few lesions of MS, but it is not understood how this occurs, nor why it only occurs in some lesions.

Disease course

Clinical expression of the disease varies greatly among individuals. Temporary paralyses, optic neuritis, weakness in limbs, and cognitive difficulties are a few of the multitudinous ways MS can present. Progression of the disease also varies greatly and may be acute or gradual.

Initial presentation of neurological symptoms due to MS generally begins when patients are in their mid to late twenties, but the age range expands from early teens to mid-sixties.(3) Both the age and type of first neurological symptom aggregates in families, suggesting a genetic component to each.(1;3) There are no typical symptoms at onset—the range of symptoms at first presentation varies as widely as symptoms do throughout the disease course.

The range of symptoms that can be experienced due to MS is large. (**Table 1.1**) Patients may experience any combination of these symptoms and some patients tend to be afflicted by certain symptoms more frequently than others. Common neurological symptoms due to MS originate from the optic nerve, brain stem, and spinal cord.(30) A diurnal pattern of fatigue in the morning and afternoon is seen in many MS patients. Uhthoff's phenomenon, a worsening of symptoms when the patient becomes overheated, and Lhermitte's sign, an electrical sensation felt when bending the neck forward, are also characteristic symptoms. The "MS hug" felt by some patients is a squeezing sensation felt around the torso.

Table 1.1 Common neurological symptoms of MS

CNS	**Symptom**
Optic nerve	optic neuritis
	visual loss
Brain stem	speech diffiulties
	dysphagia
	diplopia
	trigeminal neuralgia
	tic douloureux
	incoordination
	ataxia
	vertigo
Spinal cord	limb weakness
	paresthesias
	Lhermitte's sign
	MS hug
	urinary incontinence
	erectile dysfunction

There are four recognized subdivisions of clinical MS (relapsing-remitting, primary progressive, secondary progressive, progressive-relapsing); a person is classified into one of these categories based on the presence or absence of relapses or episodes and the rate of continual progression of disability.

Relapses are episodes of acute neurological disability related to MS that last longer than 24 hours and resolve within days to weeks; often they can be attributed to active lesions in the corresponding area of the brain or spinal cord. The majority of patients (~85%) are initially classified as relapsing-remitting (RRMS); many of these will later convert to a secondary progressive phase (SPMS), which is classified as a stage of the disease with a lack of episodes, but a gradual worsening of symptoms over time.

About 10% of patients are initially diagnosed with primary progressive MS (PPMS). These patients never experience relapses, but have a continual progression of neurological disability from the onset of the disease. A small subset of MS patients with progressive-relapsing (PRMS) experience continual progression of disability interspersed with periods of greater acute disability.

It is not understood if these forms of MS represent different expressions of the same underlying pathology or different disease pathologies with similar clinical phenotypes. In MS studies, a major drawback to understanding differences between the subtypes is the small sample sizes available for patients with PPMS and RPMS.

The symptoms experienced by individuals with MS vary greatly, as does the rate of progression of disability. Some patients may be able to walk independently 40 years after diagnosis while others may be wheelchair-bound within 5 years. The factors influencing the rate of progression are unknown. Part of the rate of disability is described by the subtype of MS, but there is a large amount of inter-individual variability even within subtypes. The stage of progression of disability due to MS is most commonly reported using the Expanded Disability Status Scale (EDSS) from 0 (no disability due to MS) to 10 (death due to MS).(31) It is reported in half point increments based on neurological and physical examinations. Many of the higher scores represent milestones easily recognized—6: reliance on a cane or other walking assistance, 7: wheelchair bound, 9: confined to bed. However, intensive neurological examination may be required in some cases to determine an exact EDSS. Scores of 4-5.5 are difficult to determine in a clinic setting because they require testing how far the patient can walk unaided, up to 500 meters. This type of space and amount of time is often not available in a routine clinic visit. The lower scores, 0-4, can be scored in clinic visits during routine examinations.

The EDSS reports the level of disability a patient currently has but does not take into account the length of time the patient has had the disease, so it cannot accurately reflect the progression of the disease. To address this issue, the Multiple Sclerosis Severity Score (MSSS) was devised.(32) The MSSS assigns a patient a new index (0-10) that describes how fast of a “progressor” that person is in comparison to others with the same length of disease time. An index of 10 represents the person has progressed faster than 100% of patients, whereas an index of 3 represents a patient that has progressed faster than only 30% of patients with the same disease duration. The MSSS algorithm uses a patient’s EDSS score and the amount of time passed between the EDSS measurement and initial disease, and then compares this to a distribution of other patients to create the MSSS. The distribution used for comparison can be from the patients in a local dataset or a standard distribution provided with the software can be used.(32)

Other scales can be used to measure disability in MS patients. Some scales focus on specific aspects of disability (e.g. bladder control scale), while others try to measure disability as a whole, similar to the EDSS. The Multiple Sclerosis Functional Composite (MSFC) was created for assessment in clinical studies to include often overlooked clinical dimensions, such as cognition, in an overall disability score.(33) One scale used independently and in the MSFC is the timed 25 foot walk. The timed walk is a simple test easily measured in the clinic setting—patients are timed while they walk 25 feet. Use of walking aids is permitted. Longer walks indicate greater disease disability. Walking times are recorded and can be used to monitor disability, specifically of the lower limbs, over time.

Current treatments

There is no cure for MS and current treatments are focused only on easing the symptoms of the disease. There are several treatments currently approved for MS, but all are for patients with relapsing forms of the disease. No FDA-approved treatment is available for patients with PPMS. Treatments for relapsing forms aim to decrease the rate of exacerbations. FDA-approved treatments include interferon β, glatiramer acetate, natalizumab, mitoxantrone, fingolimod , teriflunomide , and dimethyl fumarate.

Interferon β treatments were approved for treatment of MS in the 1990s.(34) There are multiple forms of interferon β—IFN β-1a (brand names Avonex, Rebif) and IFN β-1b (brand names Betaseron, Extavia); all are self-administered injections. Avonex, Betaseron, and Extavia are given as subcutaneous injections once a week. Rebif is administered as an intramuscular injection three times a week. Due to the nature of interferons, many patients experience flu-like symptoms upon injection. These symptoms can be combated by taking ibuprofen or acetaminophen before injection. However, some patients cannot tolerate the symptoms from one or all of the interferon treatments and must be changed to other treatments. As with all types of injections, patients must rotate injection sites to decrease bruising, pain, and permanent damage. One difficult side effect of Rebif is that it dissolves the tissue around where it has been injected repeatedly; many patients have “craters” across their bodies. Most new patients have historically been offered interferons as their first line of treatment, due to their effectiveness in decreasing relapses and few life-threatening side effects. The major complications that result from interferons are possible liver damage and neutropenia, so patients on these treatments must have their liver levels and blood counts checked every three to six months.

Glatiramer acetate (brand name Copaxone) has been approved for treatment in RRMS patients and is self-administered as a subcutaneous injection. Patients on glatiramer acetate are not required to have any routine blood work done to test for serious side effects, such as liver failure, but patients do tend to experience greater injection site reactions than with interferon treatments.

Natalizumab (brand name Tysabri) was approved for use in RRMS patients in November 2004 as an IV infusion once every four weeks, but was pulled from the market shortly thereafter. It is extremely effective in decreasing the number of relapses in patients, but two members of the clinical trials later developed progressive multifocal leukoencephalopathy (PML) as a result of taking the drug.(34) PML is caused by the JC virus in patients with decreased immune systems. In 2006, the drug was reintroduced into the market with a black box warning. Patients with MS are able to be on the drug if approved by a physician participating in the TOUCH Prescribing Program, and are closely monitored for any sign of PML. Research on natalizumab and PML continues. Risk of PML development is greatest (1 per 90) in patients who are positive for antibodies to JC-virus and have received more than 24 doses.(35)

Mitoxantrone (brand name Novantrone) is given by intravenous infusion every three months, for up to 24 months. The goals of treatment are to reduce neurologic disability and relapses. Mitoxantrone can be used to minimize progression in addition to relapses; however, there is a lifetime cumulative dose limit due to possible cardiac toxicity, so it cannot be used throughout a patient's entire disease course.

In November 2010, the first oral medication for MS was approved, fingolimod (brand name Gilenya). Fingolimod is a capsule that should be swallowed once per day. One benefit of an oral medication is that there will, hopefully, be a greater compliance in taking the medication across the board. Since the introduction of fingolimod, two other oral medications have been approved by the FDA: dimethyl fumarate (brand name Tecfidera) and teriflunomide (brand name Aubagio).Dimethyl fumarate is taken by capsule two times per day, teriflunomide once per day. The primary aim of these drugs is the same as the injectable drugs, to prevent future relapses in a patient.

Although most clinicians follow the McDonald criteria to diagnose definitive MS, many physicians will start a patient with probable MS (following evidence of only one episode) on MS-related treatments. One of the prominent reasons expressed for this practice of early treatment is based on the observation that patients who have fewer relapses in the first three years of the disease tend to have a milder disease course throughout their lives. It is unknown whether the early relapses cause the more severe progression down the road or if relapses and progression are separate manifestations of the disease, but the hope is that if the number of relapses can be decreased early in the disease it may benefit patients over time.

Genetics of Multiple Sclerosis

Genetic epidemiology

Knowledge of the presence of genetic causes of MS is mostly founded on studies showing an increased relative recurrence risk, especially sibling recurrence risk. Sibling recurrence risk is 5%, and that of parents and children is 2%.(30) The frequency of conjugal multiple sclerosis is low (0.17%)(36) and lack of clinical concordance in published datasets of conjugal pairs concordant for MS suggest a genetic etiology, as does the increased risk in children of conjugal pairs.(37) Twin studies in the United States, Canada, the UK, and Denmark show higher concordance in monozygotic twins than in dizygotic twins.(38-42) (**Table 1.2**)

Table 1.2 Published twin studies of MS

Country	Number of twin pairs	MZ concordance (%)	DZ concordance (%)
Canada	390	25	5
USA/Canada	1123	13	4
UK	105	25	3
Denmark	178	24	3
Italy	216	15	4
France	54	6	3

MZ: monozygotic twin pairs; DZ: dizygotic twin pairs

While these studies provide convincing evidence of the genetic component of MS, it is interesting to note that twin studies from France and Italy do not show different concordance rates by zygosity.(43;44) The study of twins in France concluded concordance of "some form of clinical, radiological, or electrophysiological abnormality," whereas the majority of studies looked at concordance of clinically definite MS. It is possible some differences in outcomes may be because of this difference in clinical criteria. Heritability estimates of MS vary widely, but are in the range of 0.15 to 0.76.(45;46) Adoption and stepsibling studies show no increased risk of MS for individuals living with an individual with MS who is not genetically related.(47;48)

In summary, extensive familial studies such as those described here strongly support the hypothesis of a genetic component for multiple sclerosis.

MS susceptibility loci

Understanding the specific genetic components of multiple sclerosis is an ongoing process. The Human Leukocyte Antigen (HLA) region on chromosome 6p21 was found to contribute significant risk to development of the disease in the 1970s.(12-14) The DRB1*15:01 allele is the strongest association and the HLA region accounts for 10-50% of the genetic component of MS.(14)

The decades following this exciting discovery were plagued with frustrations. Numerous linkage studies were performed that yielded no new insights into the genetic architecture of MS.(49;50) In 2007, a single nucleotide polymorphism (SNP) in the *IL7R* loci was discovered in a candidate gene study, and replicated in following studies.(51-53) Additional loci were discovered through genome-wide association studies (GWAS), including *IL2RA*, *RPL5*, *CD58*, *CLEC16A*, and *EVI5*.(52;54)

In the following years, additional loci emerged that were either confirmed by replication or reached genome-wide significance. Understanding of the genetics of MS broadened significantly because of collaborations between independent groups. One major collaboration of note is the International Multiple Sclerosis Genetics Consortium (IMSGC). Several analyses performed by this group have added greatly to the overall knowledge of the genetics of MS.

Loci discovered between 2007 and 2011 by additional GWAS and meta-analyses include *CD226*, *CD6*, *IRF8*, *TNFRSF1A,* and *TYK2*.(53;55;56) While these positive results were promising and helped to move the understanding of MS pathophysiology forward, the published literature suggests only 3% of the total variance of MS risk is conferred by the variants described thus far.(57)

The density of polymorphisms and the extensive linkage disequilibrium at the HLA have made delving into the region difficult. However, larger sample sizes and advancements in genotyping technology made it possible to determine multiple independent effects in the HLA region in addition to HLA-DRB1*15:01, including HLA-A*02:01, HLA-DRB1*03, and HLA-DRB1*13:03.(58-60)

A collaborative effort by the IMSGC and Wellcome Trust Case Control Consortium (WTCCC) in 2011 produced the largest GWAS of MS to date.(60) This analysis of 27,148 individuals (9,772 cases, 17,374 controls) and 465,434 autosomal SNPs confirmed previously identified and strongly suspected MS loci and identified an additional 29 novel loci. However, the risk conferred by each of these variants is small (odds ratios (OR) ~1.1-1.3) and, unfortunately, when we combine knowledge from all of these variants there is still a great portion of the genetics of MS left to be discovered.(57)

A follow-up analysis was recently conducted by the IMSGC. In collaboration with other auto-immune disease groups, a custom genotyping array focusing on loci associated with auto-immune diseases was formed—the ImmunoChip. (The creation of the ImmunoChip is discussed in detail in Chapter III.) Dense coverage of known loci, including rare variants, and a substantial dataset allowed for fine-mapping of several of these regions. 14,498 MS patients and 24,091 controls were analyzed for 161,311 autosomal SNPs and identified 135 potentially associated regions.(61) Replication was performed by combining the dataset with previous datasets for a combined total of 80,094 individuals. 48 new risk variants for MS were identified, bringing the total number to 110 variants in 103 loci (**Table 1.3**), outside of the HLA region. Estimates suggest these variants explain 20% of the sibling recurrence risk.(62)

The progress in the field of human genetics for MS has grown tremendously in the past six years. The loci identified paint a broader picture of the underlying genetic architecture of the disease. Even though the number of loci tops 100, these loci cannot explain all of the genetics of MS. Additional work is required. The inability to explain the genetics of disease even with numerous loci is not a situation unique to MS, but is a problem in the spotlight throughout the field of human genetics. This problem of “missing heritability” may be tackled by searching for additional variants, both rare and common. The discovery of any additional common variants will require datasets even larger than the tens of thousands of samples already studied, and will likely have very small effects.

Table 1.3 Current list of known MS loci outside of the HLA region

CHR	rsID	BP	RAF	RA	OR (95% CI)	P-value	Function	Published rsID	Published Gene
1	rs3748817	2525665	0.6412	A	1.14 (1.10-1.18)	1.33E-12	intronic	rs4648356	MMEL1
1	rs3007421	6530189	0.123	A	1.12 (1.07-1.18)	9.61E-07	intronic	-	-
1	rs12087340	85746993	0.08673	A	1.22 (1.15-1.29)	5.13E-12	intergenic	-	-
1	rs11587876	85915183	0.7931	A	1.12 (1.07-1.17)	8.4E-08	intronic	-	-
1	rs41286801	92975464	0.1441	A	1.20 (1.15-1.25)	7.92E-16	UTR3	rs11810217	EVI5
1	rs7552544	101240893	0.5583	A	1.08 (1.05-1.12)	3.67E-06	intergenic	rs12048904	VCAM1
1	rs11581062	101407519	0.2947	G	1.05 (1.01-1.09)	0.012	intronic	rs11581062	VCAM1
1	rs6677309	117080166	0.8786	A	1.34 (1.27-1.41)	1.45E-28	intronic	rs1335532	CD58
1	rs666930	120258970	0.5268	G	1.09 (1.06-1.13)	7.49E-08	intronic	-	-
1	rs2050568	157770241	0.5342	G	1.08 (1.05-1.12)	1.33E-06	intronic	-	-
1	rs35967351	160711804	0.6726	A	1.09 (1.05-1.13)	1.7E-06	intronic	-	-
1	rs1359062	192541472	0.8164	C	1.18 (1.13-1.23)	1.84E-13	intergenic	rs1323292	RGS1
1	rs55838263	200874728	0.7057	A	1.12 (1.08-1.17)	1.41E-09	intronic	rs7522462	KIF21B
2	rs4665719	25017860	0.2532	G	1.09 (1.05-1.13)	6.8E-06	intronic	-	-
2	rs2163226	43361256	0.7147	A	1.10 (1.07-1.15)	7.02E-08	intergenic	rs12466022	No gene
2	rs842639	61095245	0.653	A	1.11 (1.08-1.15)	1.7E-09	ncRNA_intronic	-	-
2	rs7595717	68587477	0.264	A	1.10 (1.06-1.14)	3.29E-07	intergenic	rs7595037	PLEK
2	rs17174870	112665201	0.7581	G	1.03 (1.00-1.07)	0.08835	intronic	rs17174870	MERTK
2	rs9967792	191974435	0.619	G	1.11 (1.07-1.15)	1.8E-09	intronic	-	-
2	rs9989735	231115454	0.1822	C	1.17 (1.12-1.22)	7.84E-14	intronic	rs10201872	SP140
3	rs11719975	18785585	0.2693	C	1.09 (1.05-1.13)	5.39E-06	intergenic	-	-
3	rs2371108	27757018	0.3835	A	1.08 (1.05-1.12)	2.06E-06	downstream	rs11129295	EOMES
3	rs1813375	28078571	0.4695	A	1.15 (1.12-1.19)	5.75E-18	intergenic	rs669607	-
3	rs4679081	33013483	0.5222	G	1.08 (1.04-1.11)	1.2E-05	intergenic	-	-
3	rs9828629	71530346	0.6236	G	1.08 (1.05-1.12)	5.49E-06	intronic	-	-

Table 1.3 *continued*

CHR	rsID	BP	RAF	RA	OR (95% CI)	P-value	Function	Published rsID	Published Gene
3	rs2028597	105558837	0.92028	G	1.04 (0.98-1.11)	0.1786	intronic	rs2028597	CBLB
3	rs1131265	119222456	0.8037	C	1.19 (1.14-1.24)	1.97E-15	exonic	rs2293370	TMEM39A
3	rs1920296	121543577	0.6441	C	1.14 (1.11-1.18)	6.75E-15	intronic	rs4285028	CD86
3	rs2255214	121770539	0.518	C	1.11 (1.08-1.15)	1.72E-10	intergenic	rs4308217	CD86
3	rs9282641	121796768	0.91882	G	1.12 (1.05-1.19)	0.000586	UTR5	rs9282641	CD86
3	rs1014486	159691112	0.4335	G	1.11 (1.07-1.14)	1.16E-09	intergenic	rs2243123	IL12A
4	rs7665090	103551603	0.5173	G	1.08 (1.05-1.12)	2.41E-06	intergenic	rs228614	NFKB1
4	rs2726518	106173199	0.55	C	1.09 (1.05-1.13)	1.23E-05	intronic	-	-
5	rs6881706	35879156	0.7246	C	1.12 (1.08-1.16)	4.87E-09	intergenic	rs6897932	IL7R
5	rs6880778	40399096	0.598	G	1.10 (1.06-1.14)	1.7E-08	intergenic	rs4613763	PTGER4
5	rs71624119	55440730	0.7552	G	1.12 (1.08-1.17)	2.7E-09	intronic	rs6859219	ANKRD55
5	rs756699	133446575	0.8729	A	1.12 (1.07-1.18)	2.97E-06	intergenic	-	-
5	none	141506564	0.6096	C	1.07 (1.04-1.11)	5.96E-05	intronic	-	-
5	rs2546890	158759900	0.5234	A	1.06 (1.02-1.09)	0.000659	ncRNA_exonic	rs2546890	IL12B
5	rs4976646	176788570	0.3399	G	1.13 (1.09-1.17)	1.04E-12	intronic	-	-
6	rs17119	14719496	0.812	A	1.11 (1.06-1.15)	1.91E-06	intergenic	-	-
6	rs941816	36375304	0.1813	G	1.13 (1.08-1.18)	4.47E-09	intronic	-	-
6	rs72928038	90976768	0.174	A	1.11 (1.07-1.16)	7.63E-07	intronic	rs12212193	BACH2
6	rs802734	128278798	0.6868	A	1.03 (0.99-1.06)	0.1577	intergenic	rs802734	THEMIS
6	rs11154801	135739355	0.366	A	1.11 (1.07-1.15)	2.35E-09	intronic	rs11154801	MYB
6	rs17066096	137452908	0.2294	G	1.14 (1.10-1.18)	5.91E-12	intergenic	rs17066096	IL22RA2
6	rs7769192	137962655	0.5464	G	1.08 (1.04-1.12)	0.000013	intergenic	-	-
6	rs67297943	138244816	0.7841	A	1.12 (1.07-1.16)	4.83E-08	intergenic	rs13192841	OLIG3
6	rs212405	159470559	0.6228	T	1.15 (1.11-1.19)	1.43E-15	intergenic	rs1738074	TAGAP
7	rs1843938	3113034	0.4382	A	1.08 (1.05-1.12)	2.21E-06	intergenic	-	-

Table 1.3 *continued*

CHR	rsID	BP	RAF	RA	OR (95% CI)	P-value	Function	Published rsID	Published Gene
7	rs706015	27014988	0.1819	C	1.14 (1.09-1.19)	1.29E-09	intergenic	-	-
7	rs917116	28172739	0.2045	C	1.12 (1.07-1.16)	2.07E-08	intronic	-	-
7	rs60600003	37382465	0.1033	C	1.16 (1.10-1.22)	2.53E-08	intronic	-	-
7	rs201847125	50325567	0.6952	G	1.11 (1.07-1.15)	2.91E-08	intergenic	-	-
7	rs354033	149289464	0.7427	G	1.03 (1.00-1.07)	0.07696	ncRNA_intronic	rs354033	ZNF767
8	rs1021156	79575804	0.2425	A	1.12 (1.08-1.16)	5.6E-10	intergenic	rs1520333	IL7
8	rs2456449	128192981	0.3637	G	1.10 (1.06-1.14)	2.21E-08	intergenic	-	-
8	rs4410871	128815029	0.7175	G	1.12 (1.08-1.16)	1.98E-09	intergenic	rs4410871	MYC
8	rs759648	129158945	0.3068	C	1.09 (1.05-1.13)	2.82E-06	intergenic	rs2019960	PVT1
9	rs2150702	5893861	0.49	G	1.16 (1.10-1.22)	3.3E-08	intronic	rs2150702	MLANA
10	rs2104286	6099045	0.7215	A	1.21 (1.16-1.26)	7.61E-23	intronic	rs3118470	IL2RA
10	rs793108	31415106	0.5036	A	1.09 (1.06-1.13)	5.61E-08	intergenic	-	-
10	rs2688608	75658349	0.5492	A	1.07 (1.03-1.10)	6.37E-05	intergenic	-	-
10	rs1782645	81048611	0.4339	A	1.09 (1.05-1.13)	4.3E-07	intronic	rs1250550	ZMIZ1
10	rs7923837	94481917	0.6118	G	1.11 (1.07-1.14)	4.58E-09	intergenic	rs7923837	HHEX
11	rs7120737	47702395	0.1454	G	1.13 (1.08-1.18)	7.61E-08	intronic	-	-
11	rs34383631	60793330	0.3961	A	1.11 (1.07-1.15)	5.69E-10	intergenic	rs650258	CD6
11	rs694739	64097233	0.6161	A	1.08 (1.04-1.11)	1.3E-05	intergenic	-	-
11	rs533646	118566746	0.6845	G	1.10 (1.06-1.14)	3.6E-07	intergenic	-	-
11	rs9736016	118724894	0.6275	T	1.10 (1.07-1.14)	2.2E-08	intergenic	-	-
11	rs523604	118755738	0.5259	A	1.09 (1.05-1.13)	2.5E-07	intronic	rs630923	CXCR5
12	rs1800693	6440009	0.3979	G	1.14 (1.11-1.18)	6.92E-16	intronic	rs1800693	TNFRSF1A
12	rs12296430	6503500	0.189	C	1.14 (1.09-1.18)	3.62E-10	intergenic	-	-
12	rs11052877	9905690	0.3642	G	1.10 (1.07-1.14)	5.37E-09	UTR3	rs10466829	CLECL1
12	rs201202118	58182062	0.6677	A	1.14 (1.10-1.18)	7.4E-13	intronic	rs12368653	CYP27B1

Table 1.3 *continued*

CHR	rsID	BP	RAF	RA	OR (95% CI)	P-value	Function	Published rsID	Published Gene
12	rs7132277	123593382	0.1865	A	1.10 (1.06-1.15)	1.88E-06	intronic	rs949143	MPHOSPH9
13	rs4772201	100086259	0.8189	A	1.12 (1.07-1.17)	1.67E-07	intergenic	-	-
14	rs2236262	69261472	0.4974	A	1.08 (1.04-1.11)	1.16E-05	intronic	rs4902647	ZFP36L1
14	rs4903324	75961511	0.1897	A	1.10 (1.05-1.14)	9.62E-06	intergenic	rs2300603	BATF
14	rs74796499	88432328	0.95409	C	1.31 (1.21-1.42)	8.47E-11	intronic	rs2119704	GALC
14	rs12148050	103263788	0.3491	A	1.08 (1.04-1.11)	1.47E-05	intronic	-	-
15	rs59772922	79207466	0.8281	A	1.11 (1.06-1.15)	4.02E-06	intergenic	-	-
15	rs8042861	90977333	0.4405	A	1.08 (1.05-1.12)	9.8E-07	intronic	-	-
16	rs2744148	1073552	0.1767	G	1.09 (1.04-1.13)	0.000102	intergenic	rs2744148	SOX8
16	rs12927355	11194771	0.6779	G	1.21 (1.17-1.26)	8.19E-27	intronic	rs7200786	CLEC16A
16	rs4780346	11288806	0.2328	A	1.09 (1.05-1.13)	6.8E-06	intergenic	-	-
16	rs6498184	11435990	0.8132	G	1.15 (1.10-1.21)	2.07E-10	intergenic	-	-
16	rs7204270	30156963	0.5049	G	1.09 (1.06-1.13)	9.32E-08	intergenic	-	-
16	rs1886700	68685905	0.1402	A	1.11 (1.06-1.16)	8.76E-06	intronic	-	-
16	rs12149527	79110596	0.4706	A	1.08 (1.05-1.12)	1.74E-06	intronic	-	-
16	rs7196953	79649394	0.2875	A	1.08 (1.04-1.12)	2.65E-05	intergenic	-	-
16	rs35929052	85994484	0.8902	G	1.14 (1.09-1.20)	3.32E-07	intergenic	rs13333054	IRF8
17	rs12946510	37912377	0.4748	A	1.08 (1.04-1.11)	8.51E-06	intergenic	-	-
17	rs4796791	40530763	0.3646	A	1.10 (1.06-1.14)	1.81E-08	intronic	rs9891119	STAT3
17	rs4794058	45597098	0.4999	A	1.07 (1.04-1.11)	1.63E-05	intergenic	-	-
17	rs8070345	57816757	0.4533	A	1.14 (1.11-1.18)	5.43E-16	intronic	rs180515	RPS6KB1
18	rs7238078	56384192	0.7702	A	1.05 (1.02-1.10)	0.006288	intronic	rs7238078	MALT1
19	rs1077667	6668972	0.7858	G	1.16 (1.12-1.21)	3.54E-13	intronic	rs1077667	TNFSF14
19	rs34536443	10463118	0.95046	C	1.28 (1.18-1.40)	1.24E-08	exonic	rs8112449	TYK2
19	rs2288904	10742170	0.7688	G	1.14 (1.09-1.19)	9.57E-10	exonic	-	-

Table 1.3 *continued*

CHR	rsID	BP	RAF	RA	OR (95% CI)	P-value	Function	Published rsID	Published Gene
19	rs1870071	16505106	0.2933	G	1.12 (1.08-1.16)	5.68E-10	intronic	-	-
19	rs11554159	18285944	0.7303	G	1.15 (1.11-1.20)	2.58E-13	exonic	rs874628	MPV17L2
19	rs8107548	49870643	0.2545	G	1.09 (1.05-1.13)	1.98E-06	intronic	rs2303759	DKKL1
20	rs4810485	44747947	0.2473	A	1.08 (1.04-1.12)	1.78E-05	intronic	rs2425752	CD40
20	rs17785991	48438761	0.3483	A	1.09 (1.05-1.13)	6.42E-07	intronic	-	-
20	rs2248359	52791518	0.5986	G	1.07 (1.03-1.10)	9.81E-05	intergenic	rs2248359	CYP24A1
20	rs2256814	62373983	0.1855	A	1.11 (1.07-1.16)	8.34E-07	intronic	-	-
20	rs6062314	62409713	0.91863	A	1.10 (1.03-1.16)	0.003871	intronic	rs6062314	TNFRSF6B
22	rs2283792	22131125	0.5146	C	1.08 (1.05-1.12)	1.14E-06	intronic	rs2283792	MAPK1
22	rs470119	50966914	0.3892	A	1.07 (1.03-1.10)	0.000151	intronic	rs140522	SCO2

CHR: chromosome; BP: base pair position; RAF: risk allele frequency; RA: risk allele; OR: odds ratio; CI: confidence interval; P: p-value
Results as published in the IMSGC ImmunoChip analysis (61)

However, these current studies are optimally designed to detect common variants, and there is increasing interest and data suggesting that rare variants may also play a significant role in complex disease.(63) Other possibilities are to look at gene-gene and gene-environment interactions to further understand the risk of MS. Interaction between HLA alleles in MS has been demonstrated (64) and a genome-wide gene-gene interaction study in trios found significant interactions between several genes.(65) These studies provide evidence of the role of genetic interactions in the risk of MS, and they highlight a need for further studies to evaluate these results and search for other significant interactions. Three published studies have investigated interactions between HLA loci and viruses but, again, additional replication studies are needed.(66-68) As discussed above, there are many known environmental risks, but little is understood of the impact of genetics on these risks.

Published disease course analyses

The heterogeneity in the clinical expression of the disease, along with previous knowledge of risk factors for MS, both environmental and genetic, strongly suggests greater heterogeneity in the risk factors for developing the disease.

Most genetic studies to date have focused on genes conferring susceptibility to MS in general; few studies have looked at how genetic or environmental risks affect the clinical heterogeneity present in MS. There is the possibility that genetics plays a role in the expression of the disease, as well as susceptibility to MS. A family study by Barcellos et al. showed concordance for early clinical manifestations in families, and the data from this study support the hypothesis that non-HLA loci modulate varied clinical expression in individuals.(1) Another study by Hensiek et al. looked at familial effects on clinical disease course; they found concordance for age at onset (AAO) in families and clinical course between siblings. No concordance was seen for overall disease severity.(69) On the environmental risk side, month of birth was associated with bout-onset MS (RRMS, SPMS, PRMS) compared to PPMS in a study by Sadovnick et al.(70) No evidence was seen for association of month of birth to disease course, severity, or AAO.

A few genetic studies on aspects of the MS clinical course have been conducted. A study of HLA with AAO, disease course, and severity (MSSS) in Scandinavian MS patients found association between HLA-DRB1*15 and AAO; this was replicated in an extended Scandinavian cohort.(71) Several analyses of clinical course were evaluated in the 2011 IMSGC and WTCCC paper.(60) AAO was analyzed in 8,715 cases and was also found to be correlated with DRB1*15:01. Each additional allele was associated with a decrease in AAO by 10.6 months on average, consistent with previous reports. (60;72) No strong associations were seen in analyses of disease course (bout-onset MS vs. PPMS) or severity (MSSS).

In all, there are few published studies for MS clinical course. There is data to suggest the HLA region is implicated in AAO of MS in addition to overall susceptibility to the disease. There is no strong evidence for association of genetic loci to other aspects of clinical course, although whether this is due to the small number of studies performed, the approach of the studies, a need for larger sample sizes, or a true lack of genetic risk on these traits remains to be seen.

Electronic Medical Records

The use of electronic medical records in research studies

Genetic studies focus heavily on risk of disease development. Few studies evaluate genetic risk of the varied clinical expression of a disease. This is largely due to the difficulties and expense of collecting detailed longitudinal data on the large number of individuals often required for studies of complex diseases. However, these data are frequently recorded in physician notes or dictated into medical records. While data recorded in medical records is generally less standardized than data collected expressly for research purposes, it is a rich resource that should not be overlooked for complex diseases, especially MS.

Extracting data manually from medical records is tedious, time-consuming work that is prone to human error. The advent of electronic medical record (EMR) systems provides an opportunity to drastically shorten the time required to extract relevant medical information and decrease human error. These data represent a very rich, deep, and largely unexploited source of phenotypic information. The EMR system at Vanderbilt University Medical Center (VUMC) is twenty years old and provides a wealth of information that can be leveraged for research studies. This dataset is described below.

Vanderbilt University Medical Center EMR

VUMC instituted its first EMR system in the early 1990s and has continually upgraded and expanded the system since that time. Not all clinical specialties adopted its use simultaneously, but broad use of the system was seen early on.

Relevant to this study, the Multiple Sclerosis Clinic at VUMC was established in 1994 and serves as both a primary and tertiary center for the evaluation and treatment of MS. This clinic transitioned to computer-based documentation in 1997. There have been five total clinicians at the clinic since its inception, including the three currently at the clinic. The clinic was begun, and continues to be directed, by Dr. Subramaniam Sriram. Currently, two of the clinicians see patients two and a half days a week; the third clinician sees patients five days a week. 4,142 visits occurred in one recent calendar year (July 2011 to June 2012), with the number of visits ranging from 288-379 per month. The majority of patients seen at the clinic have MS; however, the MS Clinic also serves patients with related diseases, such as neuromyelitis optica (NMO) and neurosarcoidosis.

Before new patients are seen, a previous diagnosis of MS or a referral from a neurologist is required. Once admitted to the clinic, patients typically have appointments at the MS Clinic every six months if they are on MS treatments or every twelve months if not. Additional visits to check blood levels or provide treatment, such as natalizumab infusions, are conducted at the clinic, as well. Unless relapses correspond with the scheduled appointment, clinicians rarely see the patients during these episodes due to filled schedules. However, patients are strongly encouraged to call in during suspected relapses for counsel or symptomatic treatment.

At the beginning of each visit, a timed 25 foot walk is conducted and recorded in the EMR. Medications are noted and evaluated for effectiveness. Previous and new symptoms are discussed by the patient and clinician to determine their relation to MS, and possible treatments and lifestyle changes may be discussed. MRIs are not requested by clinicians on any routine basis, but are requested on an as-needed basis to confirm current relapses, detect potential pathologic reasons for specific symptoms, or monitor the efficacy of treatment on reducing disease burden since lesions do not necessarily cause clinically observed symptoms. If a patient appears to have a decrease in relapses, which may indicate an effective treatment, but changes in the MRI show subclinical levels of inflammatory disease activity (28), this suggests the treatment is not effective. MRIs at the MS Clinic are requested roughly every 2-3 years per patient as the physician desires to keep track of the disease and the efficacy of the treatments.

Synthetic Derivative

A de-identified research version of the VUMC EMR system is available for researchers via the Synthetic Derivative (SD).(73;74) The SD contains records derived from the EMR for over 2.3 million unique individuals, including inpatients and outpatients. Over one million of these records contain detailed longitudinal data. The average record is 100 kilobytes in size (roughly 30 pages of text). The SD contains data from multiple sources, include clinical narratives, diagnostic and procedural codes, intake and assessment forms, pathology, ECG, and echocardiogram reports, laboratory values, vital signs, medication orders, and genetic data. Image reports are included, the actual images are not.

All clinical data are updated regularly to the SD to include patients new to VUMC and to append new data to clinical records of existing patients as they continue to access care at VUMC. The SD is 55% female. Race/ethnicity is included as an observer-reported value. (https://starbrite.vanderbilt.edu/biovu/)

To preserve the anonymity of individuals in the SD, identifying information is removed from each record, including names, places, and identifying numbers, and the dates in each person’s record are shifted consistently within a 364-day window in the past. A unique identifier is derived using a one-way hash from the patient's medical record number and is used to label the record in the SD. The original medical record number cannot be derived from the SD identifier. Access is restricted to Vanderbilt

University (VU) individuals and requires Institutional Review Board (IRB) approval and a Data Use Agreement.

BioVU

Another VUMC initiative is the Vanderbilt Biobank (BioVU), a de-identified DNA data bank. It was designed to enable discovery and confirmation of genotype-phenotype correlations and is a repository of DNA samples linked to their respective SD records. DNA is extracted from leftover blood from routine blood draws at VUMC and stored for future use. BioVU is an opt-out system; individuals can choose to be excluded by checking an indicated box on their annual consent form or by calling a phone number posted prominently in phlebotomy areas. (74;75) Blood samples from patients who opt-out are tagged ineligible and are discarded prior to any DNA extraction. Prior to initial collection of BioVU samples in February 2007, several ethics communities, a community advisory board, and the legal department at VU were consulted. 200-400 samples per week currently accrue in BioVU. As of August 12, 2013, there are 170,024 samples from adult and pediatric clinic patients.

DNA samples are scrubbed of patient identifying information and labeled with the same unique identifier as the corresponding SD record. Researchers can select samples of interest by using information contained in the SD and request aliquots of DNA for genetic studies. All genetic data generated for BioVU samples is re-deposited into the BioVU system for other researchers to use, thus expanding the database of information.

Of the currently collected samples, 43% are male and 57% female (**Table 1.4**). The samples reflect the surrounding community, and are 67% from Caucasians and 10% African American. Records have a mean of 4.7 ± 4.6 years of history and most (97%) records have at least one medication indicated. In addition, 99% have one or more procedure codes. (BioVU statistics taken from http://starbrite.vanderbilt.edu/biovu) Initial collection of samples occurred only in adult populations. Pediatric samples were included beginning in March 2010.(76)

Table 1.4 BioVU Demographics

Demographic	%
Race	
Asian	0.9
Black	9.4
Hispanic	1.5
American Indian	0.1
Others	1.2
White	67.2
Unknown	19.6
Gender	
Male	43
Female	57
Age	
>75	6.87
71 to 75	5.25
61 to 70	16.11
51 to 60	18.66
41 to 50	15.08
31 to 40	12.01
21 to 30	11.54
11 to 20	7.6
1 to 10	5.86
<1	0.97

The de-identified nature of BioVU prohibits any re-contacting of individuals. Because of this, no additional information can be acquired than what is stated in the SD. Also, if additional DNA is needed, the researcher must wait until the patient returns for future blood draws.

Information is continually accrued in the EMR and this information is pushed through to the SD at frequent intervals. The information contained in this resource is not static, but continues to grow and provides additional data, longitudinally and depth-wise.

Discussion

MS is a very heterogeneous disease, both in risk factors and clinical expression. The genetic risk of MS has been confirmed with astounding success through various family and genetic risk studies. Multiple loci within and 103 loci outside of the HLA locus have been replicated and confirmed to be associated with MS disease risk. Less understood is the genetic architecture of the clinical expression of the disease. One reason for this is the difficulty of collecting the detailed, longitudinal data needed for in-depth studies. The advent of EMR systems provides an excellent opportunity for mining detailed clinical data. Combined with the growing knowledge of MS risk genetics, EMR data could be leveraged to produce more refined phenotypes to address the genetics of the clinical course of MS.

The following chapters will discuss our investigation into the usefulness of EMR data in genetic studies of MS using the SD at VUMC. Algorithms were designed to select individuals based on MS disease status, as well as to extract details of the clinical course of the patients. Using the data derived from the EMR, we then performed genetic analyses for clinical traits of MS.

CHAPTER II

SAMPLE SELECTION AND PHENOTYPE EXTRACTION

The advent of EMR systems provides an opportunity to drastically shorten the time required to extract relevant medical information and decrease human error. Despite this promise, extracting information from EMRs can be challenging. Typically, multi-modal algorithms must be created by incorporating EMR components such as billing codes, medication data, laboratory values, and natural language processing to achieve high positive predictive values (PPV) to identify disease states.(73;77;78) Identification of more detailed phenotypes, such as those envisioned in “next-generation phenotyping”(79) and drug response phenotypes, is more challenging and is only recently being explored.(80;81)

We worked to identify the opportunities available for data extraction from the SD, a de-identified version of the EMR at VUMC. (The SD is described in detail in Chapter I.) We developed algorithms to identify individuals with and without MS. After selection of these samples, we created additional algorithms to extract aspects of the clinical disease course of affected individuals.[1]

Case sample selection

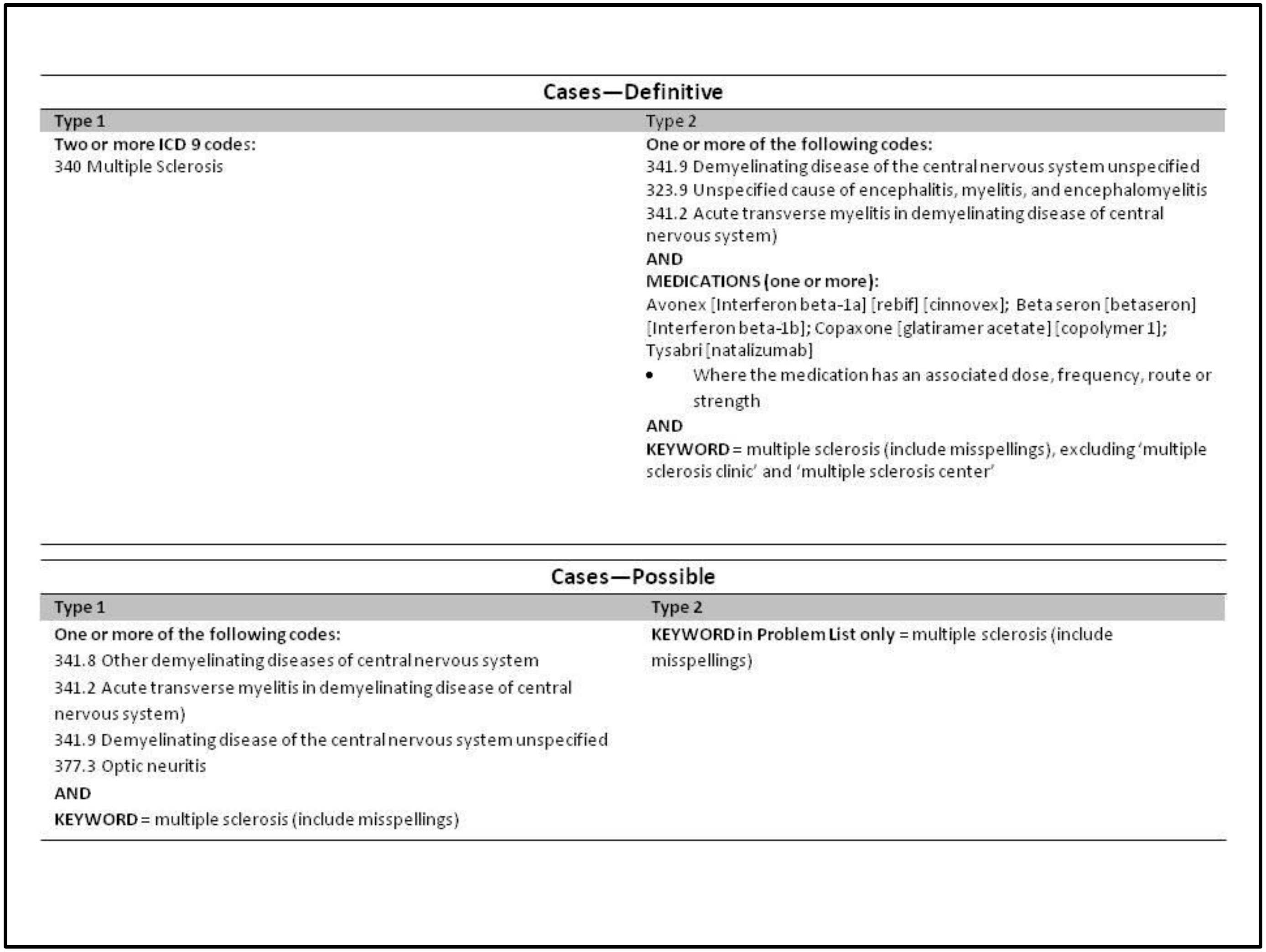
Methods

We utilized four previously published algorithms (73) to identify MS patients from this database; the algorithms focus on International Classification of Diseases (ICD-9)

[1] Sections of text from this chapter were taken from Davis MF, Sriram S, Bush WS, Denny JC, Haines JL. Automated extraction of clinical traits of multiple sclerosis in electronic medical records. J Am Med Inform Assoc 2013 Oct 22.

billing codes, prescribed MS treatments, and keywords located in narrative text. The first two algorithms identify patients who are strongly suspected to have a diagnosis of MS and need no further confirmation. These algorithms are labeled as "definitive type 1" and "definitive type 2." The remaining algorithms identify individuals missed by the "definitive" algorithms but for whom the information in their medical record is suggestive of MS. Further confirmation may be needed for these individuals and the algorithms are labeled "possible type 1" and "possible type 2." In order to improve the specificity and sensitivity of the results, we made minor modifications to three of these algorithms by increasing the number of ICD-9 codes for MS (340) required in the “definitive type 1” algorithm to require two or more instances and including the ICD-9 code for acute transverse myelitis (341.2) to the “definitive type 2” and “possible type 1” algorithms. (Final algorithms are shown in **Figure 2.1**) The algorithms are publicly available on PheKB (http://www.phekb.org/phenotype/multiple-sclerosis-demonstration-project). The “definitive type 1” algorithm identifies individuals with two or more ICD-9 billing codes for MS (340). The “definitive type 2” algorithm requires one or more non-specific billing codes (i.e. 341.9 demyelinating disease of the central nervous system), in addition to one or more medications given to MS patients and a mention of “multiple sclerosis” in the text of the records. Individuals identified by the “possible type 1” algorithms have one or more non-specific billing codes and keyword “multiple sclerosis.” The “possible type 2” algorithm relies solely on the recording of MS in the patient's Problem List (PL), which is a reference list of a patient's diagnoses and medications. The PL is supposed to be updated at every clinic visit, but issues arise with the PL when text from previous PLs is copied without revision. Many diagnoses and medications remain consistent over time, but not all do. The "cut and paste" issues with the PL should be considered whenever text of the PL is used. In this situation, MS is a lifetime diagnosis and copying the text from previous visits should not have introduced any errors in our analyses.

Figure 2.1 Case selection algorithms



Cases—Definitive

Type 1	Type 2
Two or more ICD 9 codes: 340 Multiple Sclerosis	**One or more of the following codes:** 341.9 Demyelinating disease of the central nervous system unspecified 323.9 Unspecified cause of encephalitis, myelitis, and encephalomyelitis 341.2 Acute transverse myelitis in demyelinating disease of central nervous system) **AND** **MEDICATIONS (one or more):** Avonex [Interferon beta-1a] [rebif] [cinnovex]; Beta seron [betaseron] [Interferon beta-1b]; Copaxone [glatiramer acetate] [copolymer 1]; Tysabri [natalizumab] • Where the medication has an associated dose, frequency, route or strength **AND** **KEYWORD** = multiple sclerosis (include misspellings), excluding 'multiple sclerosis clinic' and 'multiple sclerosis center'

Cases—Possible

Type 1	Type 2
One or more of the following codes: 341.8 Other demyelinating diseases of central nervous system 341.2 Acute transverse myelitis in demyelinating disease of central nervous system) 341.9 Demyelinating disease of the central nervous system unspecified 377.3 Optic neuritis **AND** **KEYWORD** = multiple sclerosis (include misspellings)	**KEYWORD in Problem List only** = multiple sclerosis (include misspellings)

Evaluation

We manually reviewed the SD records for 367 individuals from across the four case algorithms to create a gold standard for MS case status. These individuals were selected randomly across each selection algorithm (113 “definitive type 1”, 56 “definitive type 2”, 148 “possible type 1”, 50 “possible type 2”), while ensuring at least 50 evaluated from each dataset. The record of each individual was appraised to determine the clinician's final diagnosis for MS. Initial diagnoses of MS were confirmed by evaluation of the end of the patient's medical record to ensure the diagnosis had not changed. For individuals who did not have an initial diagnosis stated by the clinician, additional evidence, including PLs, medications, and clinician references to MS, were used. Each individual was categorized as diagnosed with MS, possible MS, or no MS, based on clinician impressions.

PPV were calculated as shown in **Equation 2.1**. True positives are individuals who were selected by the algorithm as cases and had an actual diagnosis of MS by a clinician (confirmed by manual review). False positives are patients who were identified by algorithm as having MS, but did not have a clinical diagnosis. PPV for case algorithms were calculated twice, with and without possible cases included as true positives.

$$PPV = \frac{True\ Positives}{True\ Positive + False\ Positives}\ (2.1)$$

Results

5,789 individuals were identified as cases by algorithms, with 4,060 (70%) individuals matching one of the “definitive” criteria (**Table 2.1**).

Table 2.1 Performance of case selection algorithms

Algorithm	**Number of Samples**	**PPV**^{1} (%)	**PPV**^{2} (%)
Definitive Type 1	3975	96	96
Definitive Type 2	85	64	79
Possible Type 1	1315	16	64
Possible Type 2	414	72	86
Total	5789	-	-

^{1}Possible cases counted as false positives; ^{2}Possible cases counted as true positives; PPV: positive predictive value

PPVs ranged from 16-96%. Reported demographics for all cases are listed in **Table 2.2**. Median follow-up time by individual is 4.6 years (range 0-20 years); the average follow-up time is 6.0 years. (**Figure 2.2b**) The average number of ICD-9 codes of 340 for individuals identified by the "definitive type 1" category is 14 (**Figure 2.3**)

Table 2.2 Case dataset demographics

Gender	**# of Individuals**
Female	4484
Male	1305
Age	
Median	54
Range	8-107
Known deceased	508
Ethnicity	
White	3513
Black	440
Asian	11
Hispanic	16
Native American	1
Unknown	1808

Figure 2.2 Follow-up length of individual records

(a) Follow-up time for all individuals (b) Follow-up time for cases (c) Follow-up time for controls

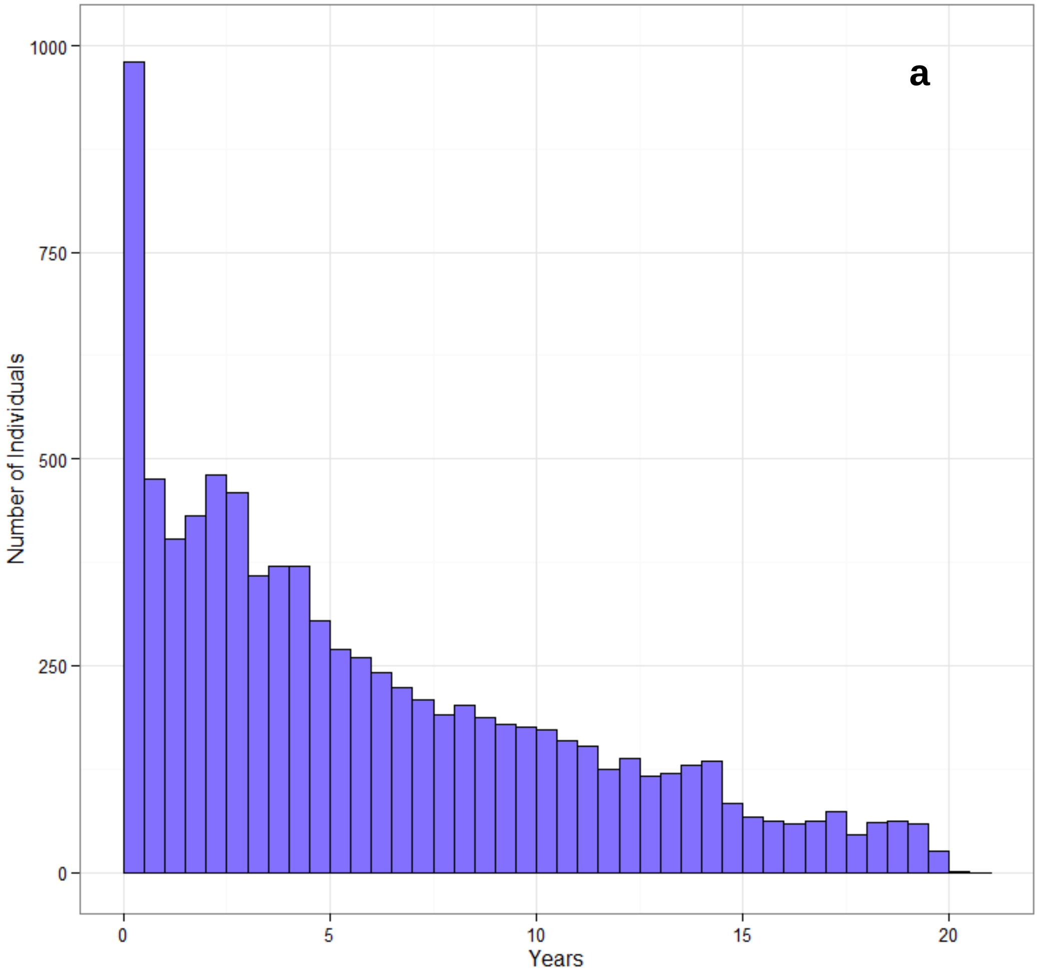


Figure 2.2 Follow-up length of individual records

(a) Follow-up time for all individuals *(b) Follow-up time for cases* (c) Follow-up time for controls

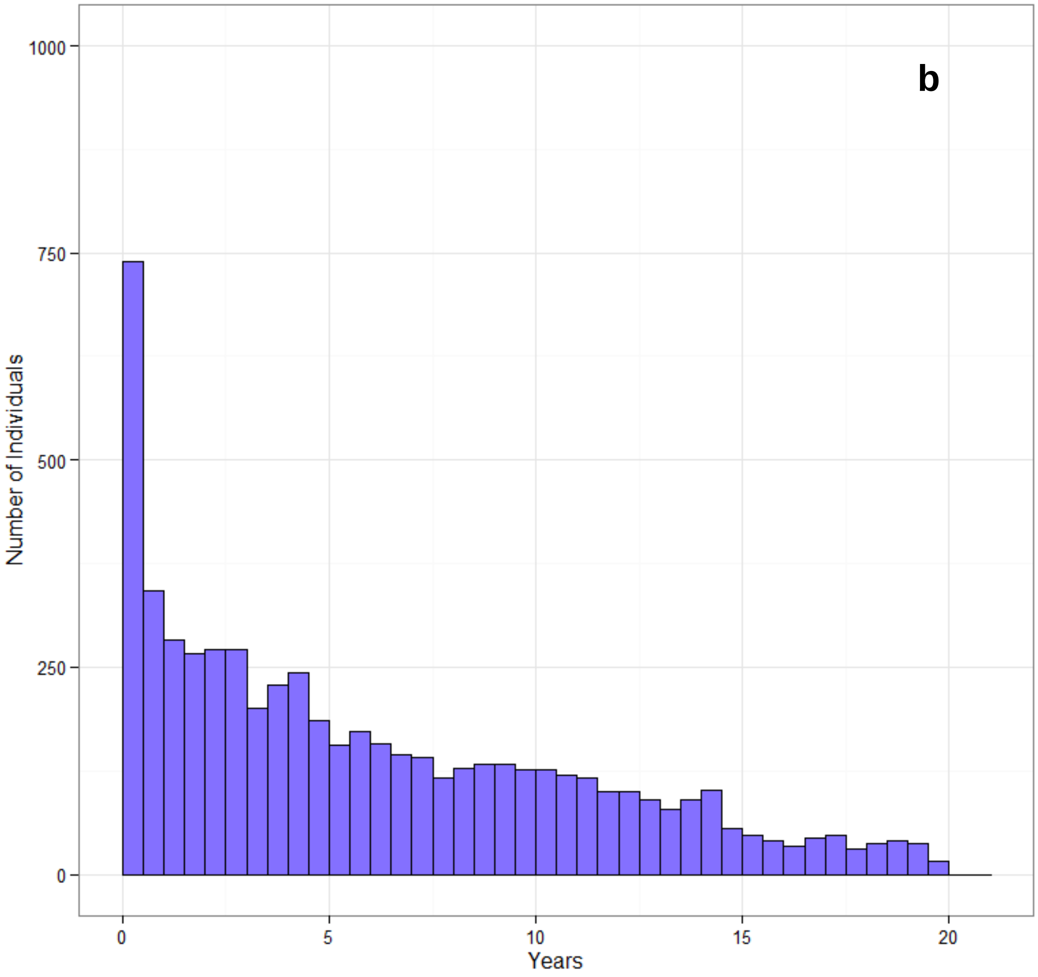


Figure 2.2 Follow-up length of individual records

(a) Follow-up time for all individuals (b) Follow-up time for cases *(c) Follow-up time for controls*

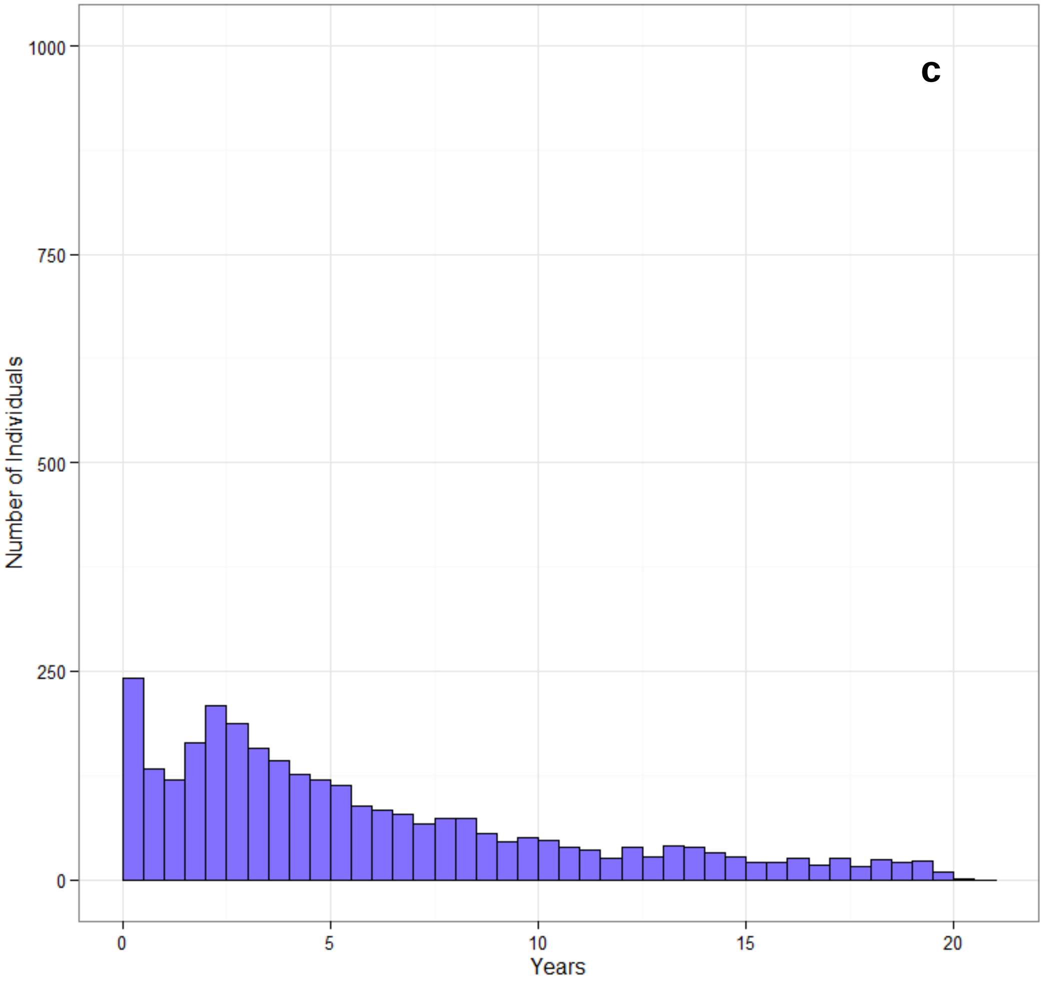
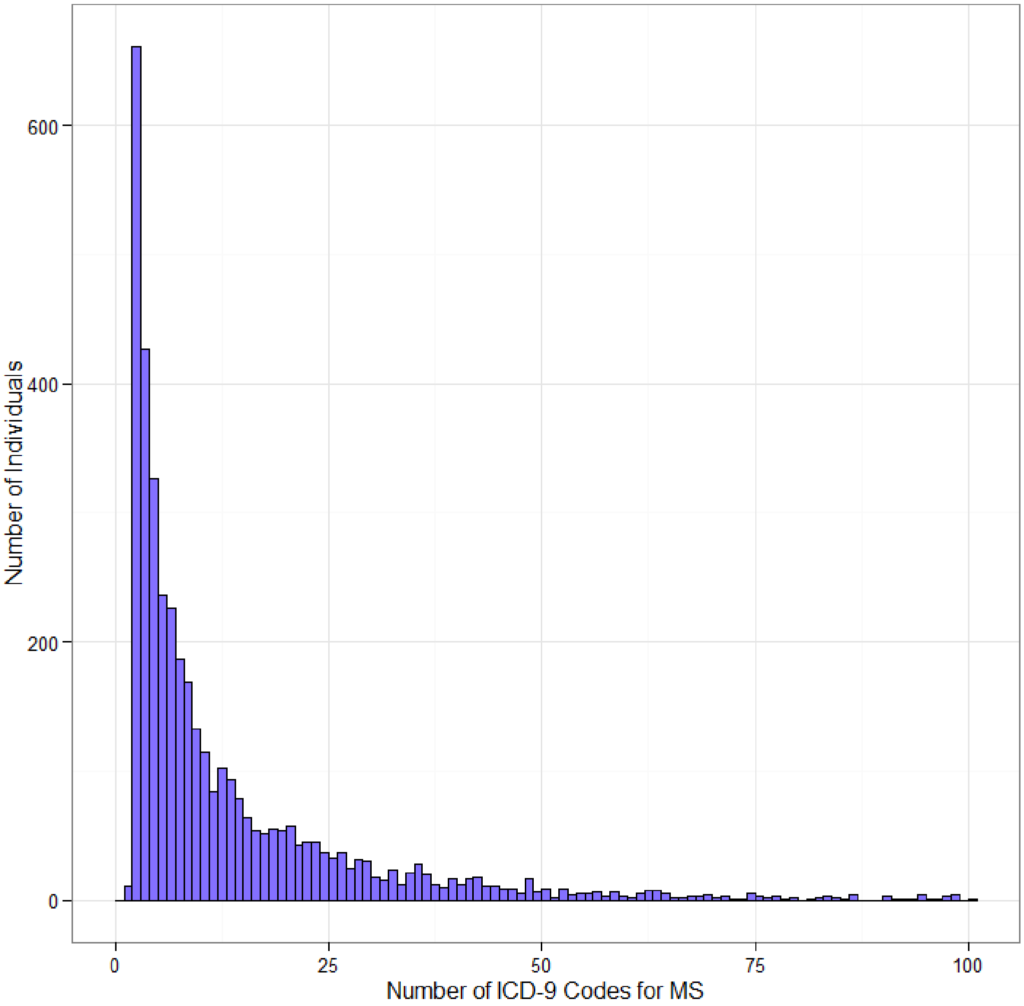


Figure 2.3 Average number of ICD-9 codes (340) for "definitive type 1" individuals



Control sample selection

We created an algorithm to identify individuals to serve as controls. In addition to absence of MS, we desired to select individuals without any evidence of autoimmune diseases. The reasons were two-fold. One, there are many genetic similarities between MS and other autoimmune diseases (as described in Chapter I) and the presence of one autoimmune disease puts an individual at higher risk for developing other autoimmune diseases, such as MS. Two, we wanted any samples we extracted to be “super” controls that could be shared across disease studies. Descriptions of the autoimmune disease exclusions (ICD-9 codes and keywords) are listed in **Appendix A**. 40,000 individuals in BioVU met these criteria. The purpose of identifying control samples was for use in genetic analyses described in Chapters III and IV, and we narrowed down individuals meeting these first criteria by selecting 2,886 individuals that matched with case subjects on age, gender, and BMI. No significant differences were seen in length of follow-up time for individuals selected as cases and controls (t-test, two-sample assuming equal variances, p=0.14; **Figure 2.2a,c**).

Algorithms to extract detailed clinical traits

Methods

In our goal to extract disease course data of MS, we evaluated all aspects of the patients' medical records. We evaluated the first 60 SD records to develop a training set to determine what types of detailed clinical information related to the disease course of MS were frequently available and how they were expressed in the clinical notes. Detailed manual review of all text of the SD was carried out for these 60 cases by the author of this work. Notes were kept on all data of MS disease course that were observed, including disease onset, EDSS, family history, symptoms, and treatments.

Laboratory values for ANA and oligoclonal bands were pulled from structured fields, as well, for the first 899 subjects identified. ANA and oligoclonal band values were found in only a small subset of the records in the laboratory results field, and ANA was found very rarely in the clinical text. After discussion with clinicians at the MS Clinic, we concluded that this is likely because the tests were performed prior to patients being seen at VUMC. As the ANA is used primarily to detect lupus and is required for diagnosis of MS, we did not direct further efforts to extraction. Oligoclonal band results were often discussed in the narrative text.

Relapse information was documented in many of the records, but the information was not located in any particular areas. Relapses were discussed in clinic visits, in communications (phone and messaging), and referral letters. The level of detail describing the relapses was not consistent. For some patients, dates of relapses, the types of symptoms, and length of relapses were meticulously recorded. In other instances, broad references were made to suspected relapses. Information regarding relapses is available in the medical records; however, due to the complexity of this trait we opted to focus on more defined traits for our initial extraction algorithms.

We identified 8 attributes for focus in this study: clinical subtype, presence of oligoclonal bands, year of diagnosis, EDSS score, timed 25 foot walk, year and origin of first neurological symptom, and MS medications. Oligoclonal bands and timed 25 foot walk scores have structured fields and are also reported in the text of the records. All other clinical traits are available only in the texts of the records.

Algorithms to extract clinical data from EMR text were implemented using Perl to access and search the records, which were stored in a MySQL database. Patient records were divided into each clinic note, PL, communication, etc. Algorithms were initially developed using 899 records as a training dataset and then evaluated using a test set of 4,890 records. Our goal was to extract data explicitly stated in the medical records; we did not infer information (e.g. the clinical subtype) from descriptions in the text. Descriptions of the algorithms are below. Pseudocode for all algorithms, including the regular expressions used, is recorded in **Appendix B**.

Clinical subtype

The four clinical subtypes of MS are: relapsing remitting, secondary progressive, primary progressive and relapsing progressive.(82) Subtypes and 100 surrounding characters were extracted from clinic notes, letters and PL that mentioned MS, then recorded in a text file. Regular expressions matching the subtypes and abbreviations were used. Subtypes preceded or followed by words suggesting the clinician was not certain, such as ‘questionable’ or ‘possible’, were excluded by use of regular expressions. Since an individual may be classified with different subtypes over the course of their illness, all distinct subtypes mentioned for each individual were kept, along with the date of the clinic note in which the information was found.

Oligoclonal bands

Over 85% of patients with MS have antibodies present in the cerebrospinal fluid (CSF) and not in the serum.(83) These are referred to as oligoclonal bands and identifying these bands can aid clinicians in diagnosis of MS. Since such testing is often performed by referring providers (and not repeated at referral centers, such as VUMC), it is important to search the clinical documentation in addition to laboratory results. We identified clinic notes, letters, and PLs mentioning oligoclonal bands and extracted 200 characters surrounding the word ‘oligoclonal.’ The result was recorded as positive (i.e. the clinician stated the test was positive or two or more bands were present) or negative (i.e. the clinician stated the result was negative or no bands were present) using regular expressions. No result was reported if one band was observed (inconclusive result). In the event that a person had both a negative and a positive result reported, the algorithm ignored the data and no conclusive result was recorded.

Year of diagnosis

MS is a clinically defined disease and the diagnostic criteria have evolved over the last 30 years.(24-26) Hence, the diagnosis of MS made by the clinician on a particular patient was based on the set of criteria that were relevant and operative at the time of the diagnosis. We extracted the year of diagnosis as recorded by the clinician, regardless of the diagnostic criteria used. Clinic notes and letters in the EMR were examined to identify mentions of the words ‘diagnosis’ and ‘multiple sclerosis.’ We identified exact, e.g. ‘1975’, and relative, e.g. ‘three years ago’, dates that occurred within 70 characters of ‘diagnosis.’

To determine the most likely diagnosis year, we first looked at exact references and recorded the most frequent year as the diagnosis year in our database. If no year of diagnosis was recorded in an exact reference, we analyzed relative references in the same manner after conversion to a specific year. Identifying the most frequently reported year removed many typographical errors that were initially observed.

Measures of progression of disease disability

The EDSS (31) and timed 25 foot walk (84) are two measures used to monitor progression of MS disability. Both could be recorded in structured fields within the EMR in a manner similar to laboratory values. At VUMC, EDSS does not have a structured field but is often mentioned in clinic notes. The MS Clinic created a structured field for the timed 25 foot walk in 2008; however, scores have been collected and recorded in the text since 1999. We created algorithms to extract both of these measures from the narrative text in the absence of structured fields.

The EDSS has a range from 0 (no disability due to MS) to 10 (death due to MS), in increments of 0.5.(31) Scores 0.5-3.5 are determined by physical exam of the physician on various disability across functional systems of the body. Scores 4-5.5 require determining how far a patient can walk unaided, up to 500 meters. Scores 6-9 are relatively easy to score as they depend upon the walking aid a person employs. Six is a cane; seven is a walker, eight and above is bedridden. Half-point increments allow for the extent of reliance. The algorithm to extract these values from the text searched for ‘EDSS’ in notes, PL, and communications. Values (0-10) reported within 50 characters after ‘EDSS’ were extracted, and the closest number within this context was recorded as an EDSS score.

To capitalize on the longitudinal aspect of timed 25 foot walks before structured values were available in 2008, we selected notes, then lines of text, from the clinical notes that mentioned ‘timed walk’, ‘25 feet’, or ‘25 foot.’ Times were extracted and recorded in seconds. The final output of this algorithm also noted if a walking aid (e.g., cane) was mentioned.

Year and origin of first neurological symptom

Since the clinical diagnosis of MS requires the presence of two lesions disseminated in space and time, patients are rarely diagnosed at the first presentation of neurological symptoms. However, the initial presentation of neurological symptoms of the disease may be important for research purposes and appears to aggregate in families (both the age and type of first neurological symptom).(1) While there are many references to symptoms in the narrative text, a complete neurological history must be investigated to be confident of identifying the first neurological symptom. We noticed that such a history was often reported in letters written from physicians at the MS Clinic to referring physicians. Information on initial presentation was available in other notes in the some patients' records, but this information was less structured and it certain cases it was difficult to determine if the symptoms described were in fact the first neurological symptoms. Referral letters appeared to be the easiest was to identify comprehensive neurological histories and we restricted our algorithms to these letters. If we were unable to extract the date and types of symptoms from these notes, the probability of extracting reliable information from other sources would be very small.

Symptoms are highly variable when considered in detail, and we considered how best to classify first neurological symptoms. The most detailed would be to group patients with the same manifestations (e.g. right leg paresthesias, diplopia), but this would lead to many groups of small sample sizes. It would also require full details of the symptoms to be recorded. Many patients may remember numbness in one leg, but not which side, which would hamper this type of classification. We elected to categorize first symptoms based on CNS origination of the symptoms, brain stem, spinal cord, and optic nerve, as the data needed for this type of classification was available and likely to reflect underlying pathology.

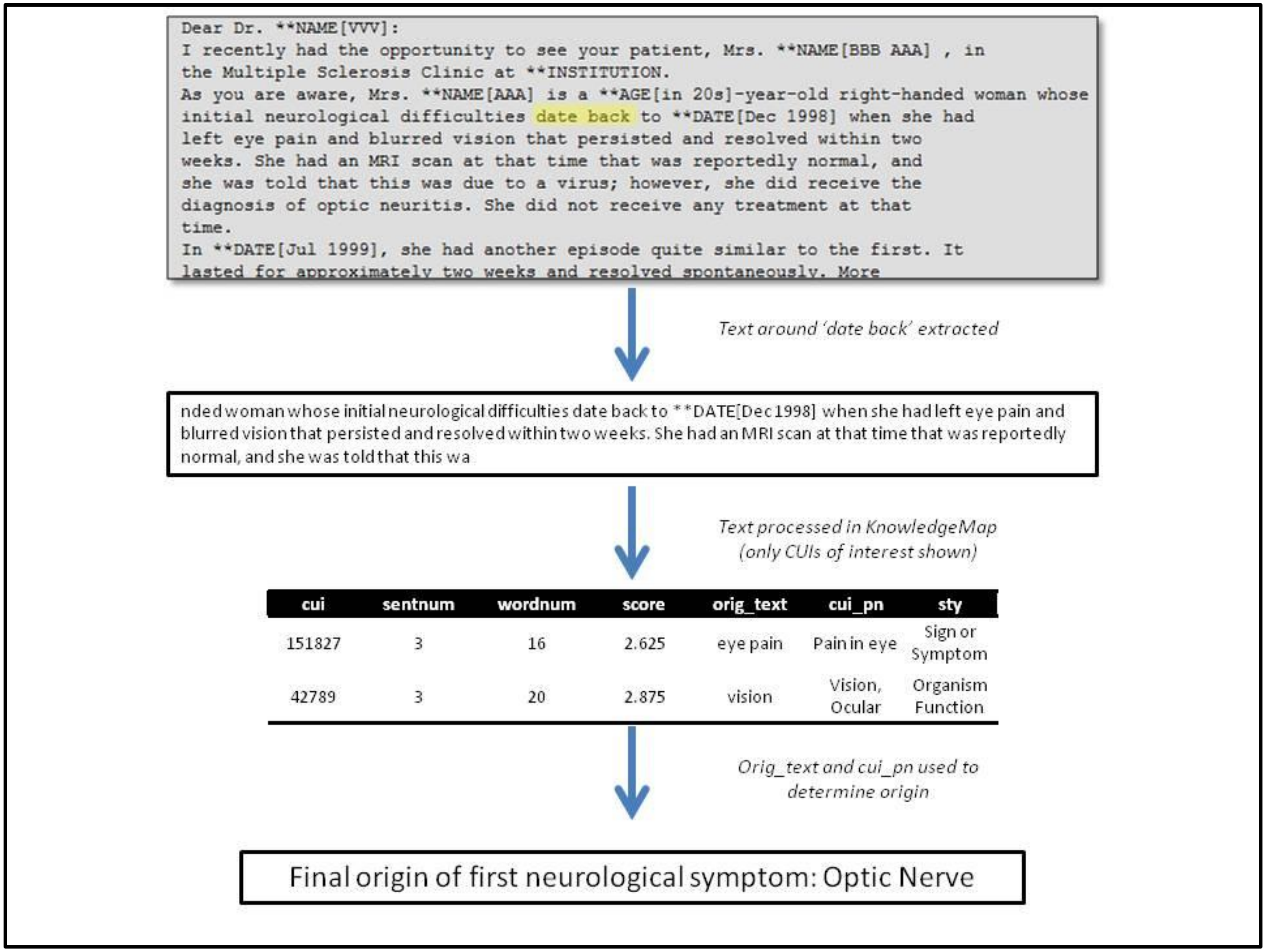
The algorithm to identify year of initial neurological symptom selected 100 characters around phrases referencing the beginning of the disease course, i.e. ‘dating back’ and ‘began’. Specific dates were extracted from these phrases, either exact or relative. If more than one symptom year was identified, the earliest year was kept.

To identify the type of first neurological symptom, 250 characters surrounding phrases that referenced the beginning of the disease course were extracted and passed through the KnowledgeMap Concept Identifier,(85;86) which is a general purpose natural language processing system supporting negation and word-sense disambiguation, similar to MetaMap.(87) Concept unique identifiers (CUIs) representing neurological symptoms were selected as the output of interest, as identified using Unified Medical Language System semantic types (see **Appendix B**). We then used text keywords and CUIs to group the symptoms into CNS site of origin (brain stem, optic nerve, or spinal cord) using a list of MS related neurological symptoms we compiled. Symptoms that did not fall in one of these categories were marked as 'other.' If more than one origin was identified, all were recorded and the origin was marked polysymptomatic. **Figure 2.4** provides a schematic of this algorithm.

Medications

Medications administered for the treatment of MS are fairly specific to this disease. MS medications are often discussed in a clinic visit with the patient and the patient is sent home with pamphlets to determine which medication they wish to start. Although VUMC has electronic prescribing tools, many outpatient prescriptions are only documented in the free text of clinical notes, clinical messaging systems, or PL, and this has been especially true of the MS Clinic. Discussion of MS medications in narrative text could be because the patient is on the medication, the patient failed the medication due to the continued progression of MS or excessive side effects, the clinician is considering the medication for the patient in the future, or the patient came into the clinic with questions regarding a specific treatment. To retrieve medications the patients were actually taking, we focused our efforts on extracting MS related medications from PL only. Medications are recorded in the PL only if a patient is on the medication. As some text from previous clinic visits is copied indiscriminately into new PLs, the PL is not a

Figure 2.4 Schematic of the algorithm to extract origin of first neurological symptom



```
Dear Dr. **NAME[VVV]:
I recently had the opportunity to see your patient, Mrs. **NAME[BBB AAA] , in
the Multiple Sclerosis Clinic at **INSTITUTION.
As you are aware, Mrs. **NAME[AAA] is a **AGE[in 20s]-year-old right-handed woman whose
initial neurological difficulties date back to **DATE[Dec 1998] when she had
left eye pain and blurred vision that persisted and resolved within two
weeks. She had an MRI scan at that time that was reportedly normal, and
she was told that this was due to a virus; however, she did receive the
diagnosis of optic neuritis. She did not receive any treatment at that
time.
In **DATE[Jul 1999], she had another episode quite similar to the first. It
lasted for approximately two weeks and resolved spontaneously. More
```

nded woman whose initial neurological difficulties date back to **DATE[Dec 1998] when she had left eye pain and blurred vision that persisted and resolved within two weeks. She had an MRI scan at that time that was reportedly normal, and she was told that this wa

cui	sentnum	wordnum	score	orig_text	cui_pn	sty
151827	3	16	2.625	eye pain	Pain in eye	Sign or Symptom
42789	3	20	2.875	vision	Vision, Ocular	Organism Function

Final origin of first neurological symptom: Optic Nerve reliable source of the length of time a patient is on a medication. In this algorithm, our goal was to determine if a patient was ever on a medication and the difficulties in teasing apart the dates a patient was on the medication were not relevant.

Extracted medications include interferon beta-1a, interferon beta-1b, glatiramer acetate, fingolimod, natalizumab, mitoxantrone, and teriflunomide. Text matching, using brand and generic names, was performed over the PL text to create a list of medications the patient had taken. Electronic prescribing tools automatically update the PL, so this method should also capture electronic prescriptions with near-perfect fidelity.

Evaluation

One hundred records were selected randomly from the test set for a blinded evaluation of the clinical trait algorithms. These records were reviewed manually for all clinical characteristics extracted by algorithms to define a gold standard. The reviewer recorded the information that the treating clinician(s) appeared the most confident in by the end of the record (i.e., year of diagnosis). The first 10 records were reviewed independently by Dr. Subramaniam Sriram, a board-certified neurologist and founder and chief of the MS Clinic, and the author of this work. Any discrepancies adjudicated by a second board-certified clinician (Dr. Joshua Denny), blinded to the source of discrepancy. Given high initial concordance, the author performed the manual abstraction of the following 90 records. Manual abstraction of the eight clinical traits took an average of 12.6 minutes per individual record, with a range of 1-40 minutes.

Clinical trait data derived from manual abstraction was compared to data extracted via the algorithms designed in this study. Recall, precision, specificity, and F-measure were calculated for all traits.

Results

Our algorithms extracted information for each clinical trait of interest in 903 (16%) to 3,523 (61%) individuals (**Table 2.3**). Specificity, precision, recall, and F-measure were calculated for each algorithm.(78) These measures use true and false positives, true and false negatives, and gold standard positives and negatives. True positives are values for a clinical trait found by algorithm that are an accurate representation of the trait in the text. A false positive is a value for a clinical trait found by algorithm that is not an accurate representation of the trait in the text (either no value is in the text or an incorrect value was extracted). A true negative is when no value is found by the algorithm and no value for the clinical trait exists in the text. A false negative is when the correct value that exists in the text was not found by the algorithm. Gold standard positives are values for clinical traits in the text confirmed by manual review. Gold standard negatives in these calculations are individuals for whom no value for the clinical trait was found in the text by manual review. Specificity (**Equation 2.2**) is a measure to evaluate if an algorithm can correctly identify when the trait of interest is absent. Precision, or PPV, (**Equation 2.1**) measures how believable the results identified by algorithm are, and recall (**Equation 2.3**) measures how much of the data in the text was missed by the algorithm. F-measure (**Equation 2.4**) evaluates the overall accuracy of the algorithm on a scale of 0 to 1 (1 being the best score).

$$Specificity = \frac{True\ negatives}{Gold\ standard\ negtaives}\ (2.2)$$

$$Recall = \frac{True\ positives}{Gold\ standard\ positives}\ (2.3)$$

$$F - measure = \frac{2 \times Recall \times Precision}{Recall + Precision}\ (2.4)$$

Specificities for all algorithms were high, with seven of eight algorithms achieving specificity greater than 90% (**Table 2.4)**. Precision ranged from 87% to 99%. For clinical subtype and timed 25 foot walk, recall was at least 90%. However, recalls for year of diagnosis and origin of first symptom were 33% and 23%, respectively. F-measure for all traits except year of diagnosis and origin of first symptom was above 70%.

Table 2.3 Number of individuals identified by each of the eight clinical trait algorithms

Clinical trait	**Individuals,** *n* (%)
Clinical subtype	3140 (54)
Oligoclonal bands	1043 (18)
Year of diagnosis	1053 (18)
EDSS	903 (16)
Timed 25-foot walk	3523 (16)
Year of first symptom	2301 (40)
Origin of first symptom	1288 (22)
MS Medications	2586 (45)

Table 2.4 Statistics of algorithms compared to blinded manual review of 100 charts for all

Clinical trait	**Gold standard positives, *n****	**Correctly identified, *n****	**Recall, %**	**Precision, %**	**Specificity, %**	**F-measure, %**
Clinical MS subtype	61	60	98	88	81	93
Oligoclonal bands	28	20	71	87	97	78
Year of diagnosis	51	17	33	89	100	49
EDSS	75	61	81	94	100	87
Timed 25-foot walk	120	99	83	99	100	90
MS Medications	99	63	64	95	93	76
Year of first symptom	56	24	43	100	100	60
Origin of first symptom	62	14	23	88	100	36

EDSS: Expanded Disability Status Scale

**n* refers to how many instances were recorded, not number of individuals. For EDSS, clinical subtype, timed 25-foot walk, medications, and origin of first symptom, this could be more than one per individual

Table 2.5 Statistics of algorithms after additional modifications

Clinical trait	**Gold standard positives, *n****	**Correctly identified, *n****	**Recall, %**	**Precision, %**	**Specificity, %**	**F-measure, %**
Timed 25 foot walk	120	108	90	99	100	94
Year of first symptom	56	31	55	97	100	70
Origin of first symptom	62	21	34	88	93	49

**n* refers to how many instances were recorded, not number of individuals. For timed 25-foot walk and origin of first symptom, this could be more than one per individual

After comparison to the gold standard was complete, we identified the need for minor changes in the algorithms for timed 25 foot walk, year of first symptom, and origin of first symptom, which significantly increased recall compared to the original algorithms at a nominal p-value of 0.05 (p=0.02, 0.03, 0.02, respectively; **Table 2.5**). During compilation into the database, some spaces and new lines were removed. We allowed for such changes by making spaces optional in regular expressions for timed walks and year and origin of first symptom. Additionally, we identified another note title that represented letters to referring physicians and included the year and origin of first symptom. The F-measure for the algorithm of origin of first symptom also significantly increased (p=0.02).

Discussion

We identified a large number of individuals with MS and detailed clinical information with minimal cost and time requirements. 40,000 individuals, when restricted to individuals in BioVU, met our control requirements, providing an extremely large dataset for further use with any autoimmune disease dataset. Both the MS case algorithms and the algorithms to extract detailed MS information performed well, with precision for the clinical trait algorithms between 87-100%. We are unaware of any other published dataset of MS patients of this size that has such detailed clinical information. This dataset provides a rich resource for better understanding MS and also shows that extraction of detailed disease states and markers of prognosis in patients with chronic disease is possible and may yield a powerful tool in chronic disease research.

Upon review, there were more false positives in the “possible type 1” category than desired. While only 16% had a diagnosis of MS, 64% were at least being considered for MS to varying degrees. Of the individuals who did not fall in either of these categories, many were seen in the MS Clinic for other diseases, such as neurosarcoidosis. Depending on the purpose of the study, individuals identified by this algorithm should be used with caution. If it is used, we recommend a manual review of all records to confirm the diagnosis. No other trends were seen for false positives identified by the other algorithms.

While many studies have identified individuals serving as cases and controls for disease status from EMRs (73;77;88;89), this is one of the first studies to focus on specific clinical traits of a disease by text mining of the EMR. A few other studies have used text mining approaches to extract blood pressures, pacemaker implantations, and left ventricular ejection fractions as a marker of heart failure. (90-92) We have shown that detailed clinical information valuable to research studies is recorded in the medical records of individuals with MS and that this information can be extracted in a highly reliable manner. Such methods could potentially be applied across multiple EMRs, such as envisioned by the eMERGE network (93) and SHRINE. (94)

We aimed for high precision creating a reliable database of information, rather than focusing on high recall, although the resulting recall of many algorithms was indeed high. The ability to create highly specific algorithms for these clinical traits is due to a number of factors, many attributable to the nature of the disease studied. First, the diagnostic criteria of MS are straightforward and, if followed exactly, a diagnosis of MS is fairly certain, especially when this diagnosis is verified by a neurologist. Second, treatments for MS are rarely used in other diseases. Third, because VUMC has a MS Clinic, with only five total clinicians staffing it since its opening in 1997, a large number of clinic notes focused on MS disease course for each individual and much less variability in the style and content of clinic notes than may be found in other disease clinics with larger staffs. It should be noted, however, that not all individuals whose records we analyzed were enrolled in the MS clinic or even seen by a VUMC neurologist. Some patients may have been seen at VUMC for other reasons and treated for MS elsewhere.

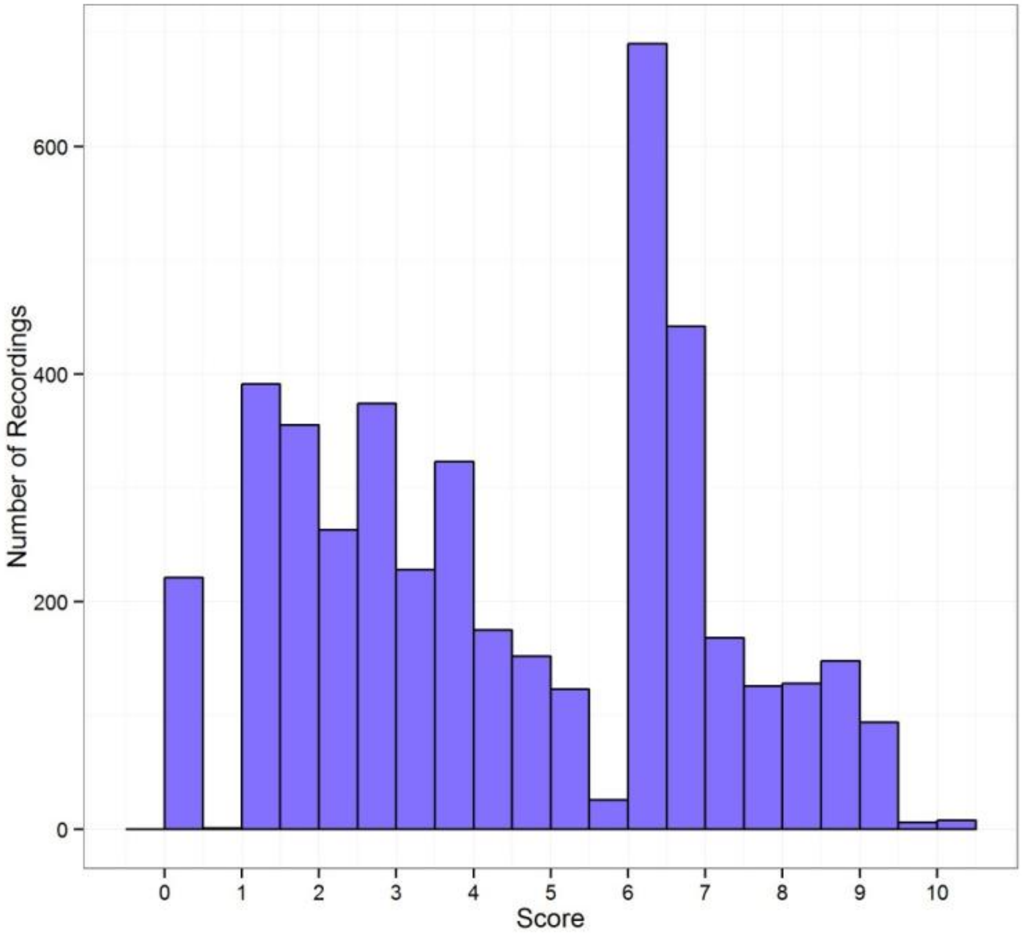
We found an average of 2.9 clinical traits per record. (**Table 2.6**) We extracted all eight clinical traits evaluated for only four individuals. All four individuals were identified by the "definitive type 1" algorithm. 715 individuals had no clinical traits extracted. It is interesting to note that these individuals were identified by all algorithms (259 "definitive type 1", 1 "definitive type 2", 264 "possible type 1", 191 "possible type 2"). The algorithm by which the individual is identified does not seem to indicate a lack of detailed clinical information.

Table 2.6 Number of clinical traits extracted per individuals

Clinical traits	**Individuals, *n***
0	715
1	867
2	881
3	1049
4	1002
5	810
6	348
7	113
8	4

We were initially surprised by the small percentage of patients (16%) with recorded EDSS scores. However, after consulting with clinicians at the MS Clinic, we discovered EDSS is rarely calculated due to the difficulty of scoring the midpoint scores in a clinic setting, which require ambulatory patients to walk up to 500 meters. It is likely due to these difficulties that we have found not all patients have a reported EDSS score and very few mid-scale scores are reported. (**Figure 2.5**) A spike is observed in the

Figure 2.5 Distribution of extracted EDSS scores

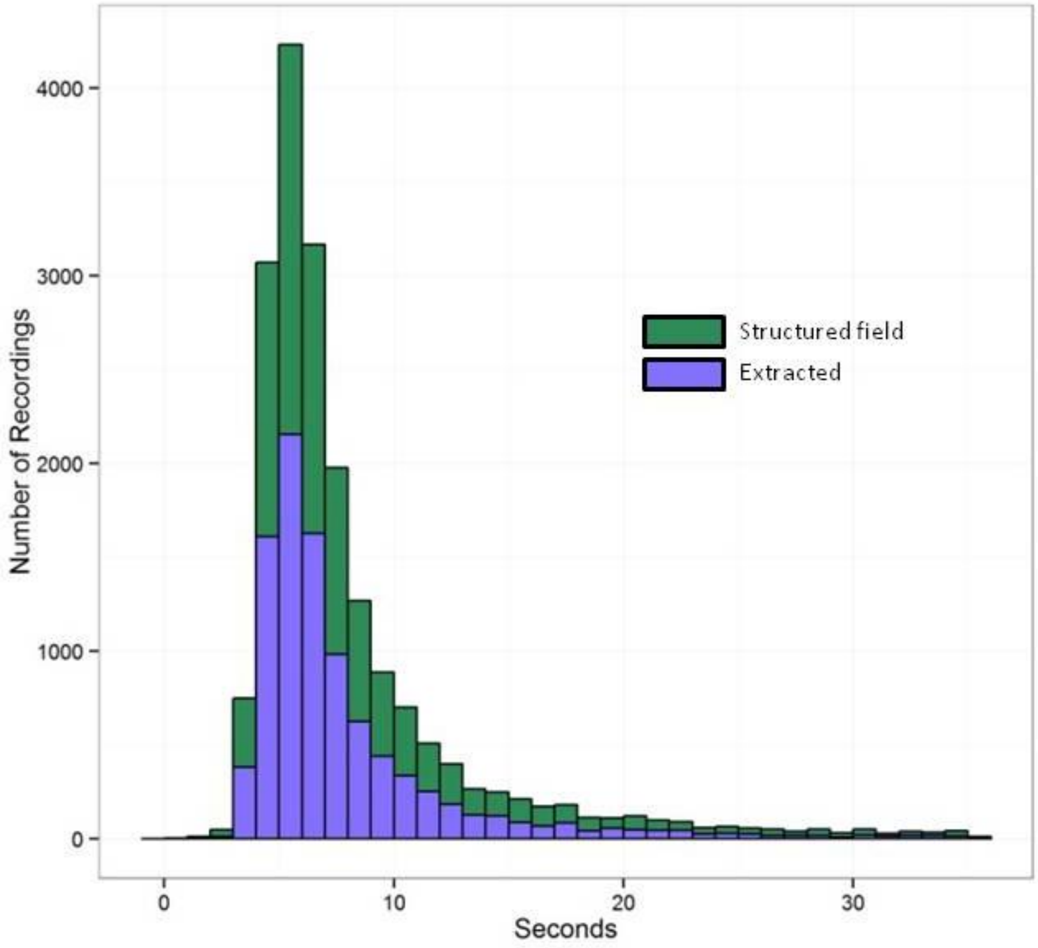


number of reported EDSS scores at six; this score depends upon the use of a cane, which makes it very easy to establish. Conversation with the clinicians and staff actually creating the notes in the EMR was important as we became familiar with this dataset and determined what would be the most productive traits on which to focus.

Beginning in 2008, timed 25 foot walks were also included in a structured field in the EMR at the MS Clinic after being recorded in a paper chart. We compared the structured field scores (9,733 scores) to those extracted from the text of the records (9,982 scores) and found no significant difference between the two datasets (t-test, two-sample assuming equal variances, p=0.25), further validating NLP methods as a secondary means of data extraction. The distributions of scores appear very similar. (**Figure 2.6**) Of the overlapping dates between the two datasets, 2,445 scores were common in both and 2,220 (91%) of these were within one second of each other. Discrepancies less than one second often occurred from a rounded time being reported in either the structured field or clinic note. Many other discrepancies appeared to be typographical errors (i.e. 8.5 seconds instead of 5.5 seconds). Of the timed walks that occurred during overlapping dates, 638 were found only in the extracted dataset, and 5,836 were recorded only in the structured field.

The addition of a structured field for the timed 25 foot walk in 2008 provided a unique opportunity to determine how much data we are missing when we do not have access to the original record (in this case, a paper chart) and have to rely on mention in the clinical narrative. The accuracy of the extracted scores compared to those in the structured field was very high, giving credence to the belief that the scores recorded by the clinician in the text are a useful surrogate for the original record. However, we are aware of the large collection of scores that were not recorded in the clinical note—70% of scores recorded in the structured fields were not identified by our algorithm searching

Figure 2.6 Distributions of timed 25 foot walk scores as found in the structured fields and extracted from the text of the clinical records

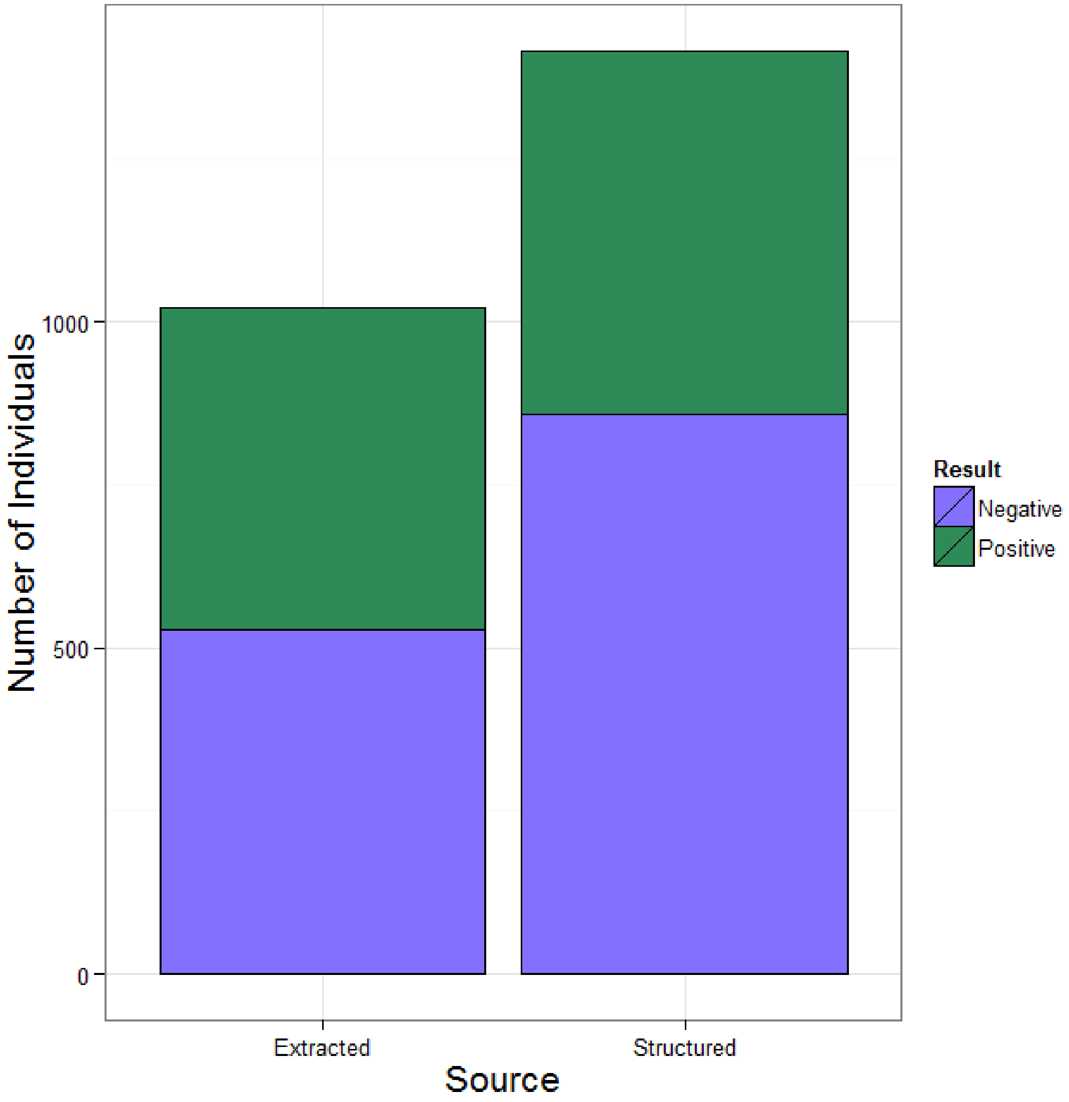


the narrative text, and the high recall compared to manual abstraction leads us to believe this is because the scores were not mentioned in the text as opposed to failure of the script. It is interesting to note that a smaller subset of scores were recorded by the clinician in the text of the clinic note but evidently failed to be recorded in the structured field.

Structured fields in the EMR would be the most accurate way to extract data. Unfortunately these fields do not always contain the desired information due to the nature of the data or the EMR, and NLP provides an opportunity to recapture this data. One drawback to using EMR-derived data for laboratory values is that if the test was not performed at the primary institution (e.g. VUMC), it will not be reported in a structured field, as discussed earlier in this chapter. For example, the test for oligoclonal bands is most commonly ordered when trying to make a diagnosis of MS. Indeed, only 24% of cases had a value for oligoclonal bands in the relevant structured fields. Because this is a common test performed when diagnosing MS, the result is often echoed in the narrative text. We capitalized on clinic note references to extract this information in an additional group of individuals. 481 individuals had a result in the structured laboratory field and in the narrative text. The concordance between these two datasets was 97.1%, indicating a high reliability in the laboratory results recorded by the clinician in the text. By using the oligoclonal band result reported in the text in the absence of a structured field value, the sample size for this clinical trait increased by 38%.

The proportion of positive results for oligoclonal bands was similar between structured field and extracted results. (**Figure 2.7**) While the proportions are lower than published literature, which suggests over 85% of individuals have oligoclonal bands(83), this may be due to the timing of the tests. At the MS Clinic, tests for oligoclonal bands are rarely performed once a diagnosis has been established, so these results likely

Figure 2.7 Comparison of oligoclonal band results by source



reflect oligoclonal band presence upon diagnosis. It is known that this status can change and oligoclonal bands may be detected later in disease course. If the tests were performed at other points in the disease course, we would expect to see results more similar to the published percentage.

The counts of individuals with each subtype of MS and origin of first symptom that we extracted are listed in **Table 2.7** and **Table 2.8**.

Table 2.7 Counts of individuals for each origin of first neurological symptom

CNS Origin	Number of Individuals
BS	407
SC	632
ON	161
BS, SC	63
BS, ON	10
SC, ON	16
BS, SC, ON	1
Total	1290

Table 2.8 Counts of individuals by subtype

Subtype	Number of Individuals
RRMS	2380
SPMS	1066
PPMS	471
RPMS	56
Total	3973

Initially, we used MedEx(95) to extract medications. This proved to be a poor choice for our dataset. MedEx was designed to increase confidence in a medication being currently taken by requiring the presence of dosage and route information. The majority of MS medications are given in one dose and one type of administration and this information is often not noted in the clinic record. Thus, it is hard to differentiate, without further text processing, if a medication is being taken or being discussed for another reason. Because of these difficulties, we focused on extraction of medications from PLs, which contain active lists of medications for each patient. By doing this, we gained greater confidence in determining which medications a person had been on. However, PLs are not always updated, resulting in a lower recall rate than desired. The numbers of individuals on each medication are listed in **Table 2.9**.

Table 2.9 Counts of individuals for each type of medication

Medication	Individuals
interferon beta-1a	1440
interferon beta-1b	541
interferon NOS*	29
glatiramer acetate	890
fingolimod	35
natalizumab	291
mitoxantrone	128

NOS: not otherwise specified
*Includes individuals with mention of “interferon” but no indication of which type

The algorithms we have written are not overly intricate, yet have yielded an extensive amount of clinical data on a large population. The scripts described in this chapter searched for specific references by the clinician about clinical traits. They did not use the text to infer information, such as diagnosis year or clinical subtype, both of which could have been done to enhance recall. Specifically, we had very low recall in our algorithm to extract diagnosis year. Upon review of instances of algorithm failure, many times we missed when a patient was diagnosed in the course of the record, as it is rare that a clinician would record the current year, instead stating, “I believe Mr. [NAME] fully meets the criteria for a diagnosis of MS” or simply listing “multiple sclerosis” as the final impression of the clinic visit. Algorithms targeting current diagnoses would greatly improve the recall of this clinical trait.

Through the work described in this chapter, we have shown that EMR databases are a rich resource of information of the detailed clinical course of multiple sclerosis. Much of this information is extractable from clinic notes by simple algorithms, with high specificity, precision, and recall.

CHAPTER III

GENETIC DATA COLLECTION AND PREPARATION

After selection of cases and controls via the algorithms discussed in Chapter II, genotyping of samples with DNA occurred concurrently with the extraction of clinical traits. Of the 5,789 samples extracted by the case algorithms in Chapter II, 1,221 were part of BioVU and had DNA available for genotyping. The ImmunoChip, described below, was selected as the genotyping platform for this project because of the enrichment of loci associated with autoimmune diseases, including MS. We are exploring the genetics of clinical traits that have been studied very little, as evidenced by the small number of published studies described previously. Due to this, focusing on loci previously implicated in autoimmune diseases appeared prudent for genetic analyses of these traits as compared to an unfocused genome-wide evaluation. While a survey of these loci using the ImmunoChip is still a broad analysis covering a large portion of the genome, it allows us to focus on candidate loci. After genotyping and SNP calling, stringent quality control (QC) measures were employed to preserve the highest quality data possible.

ImmunoChip

The ImmunoChip consortium was formed in late 2009 from independent groups studying the genetics of autoimmune diseases. The explicit purpose of forming this collaboration was to build a custom genotyping array to fine-map genes implicated in their respective diseases, test for replication of significant hits, and investigate the shared genetics of autoimmune diseases. This chip is entitled the ImmunoChip and was created as a custom bead-array by Illumina. Several studies for various diseases using the ImmunoChip have been published. (96-101)

The 196,524 SNPs on the ImmunoChip were chosen in a variety of ways. In the first three categories listed below, each disease group was allotted a certain number of SNPs that could be on the chip. The final category is a catch-all for the remainder of SNPs on the chip.

1) Fine-mapping of loci

Loci chosen by the individual disease groups were required to contain a SNP that reached genome-wide significance for their disease ($p < 5 \times 10^{-8}$) and was replicated in another study. All available SNPs at that time near the locus that passed quality control procedures at Illumina were included on the chip for fine-mapping. SNPs were selected using the available HapMap project (NCBI136/hg18) and 1000 Genomes data (February 2010 release), primarily generated from individuals of European descent.(102-104) This was not considered a significant limitation as most autoimmune diseases (and thus the datasets) are more prevalent in European Caucasians.(82)

2) High-density coverage of loci, but not mapped to completion

Loci had to contain a SNP that reached a replicated genome-wide significance threshold, but replication of the association was not required. Selection of SNPs relied heavily on what SNPs were then available in the HapMap project, and additional SNPs were selected to capture the greatest amount of variation possible (i.e. tagging SNPs).

3) Wildcards

Each group was allotted SNPs that they could choose to put on the chip for any reason. For MS, the IMSGC asked for suggestions from investigators involved in this group, then compiled the combined list to submit to the larger consortium. If the SNPs could be added to the chip, they were, without any further criteria. 2,109 SNPs were originally submitted to the IMSGC from various investigators; this list was then prioritized and submitted to the ImmunoChip consortium to be placed on the chip. 659 of these SNPs are represented on the final ImmunoChip. (**Table 3.1**).

Table 3.1 SNPs requested by the IMSGC for inclusion on the ImmunoChip

Submission Categories	Requestor	Requested	Actual on Chip
Wildcards	IMSGC Total	2109	659
	Sawcer/UK	23	22
	Harbo/Norway	17	17
	Kockum/Sweden	1517	433
	Zuvich/Vanderbilt	120	118
	McCauley/Miami	405	50
	Hemmer/Germany	33	25
Fine-mapping	IMSGC	4801	1228
	WTCCC2	1130	1038
		Total	**2868**

UK: United Kingdom; IMSGC: International Multiple Sclerosis Genetics Consortium; WTCCC2: Wellcome Trust Case Control Consortium 2

4) Miscellaneous

This category includes a conglomerate of SNPs: every known SNP for the major histocompatibility complex (MHC) was selected (6,293 SNPs), because of its large role in autoimmune disease, and a variety of SNPs were chosen by the Wellcome Trust Sanger Institute (WTSI), including every SNP from any disease that reached genome-wide significance in the original WTCCC study (105), whether the disease was autoimmune related or not.

275 loci are fine-mapped on the ImmunoChip (falling into one of the first two selection criteria categories). 21 of these loci (represented by 2,266 SNPs) were

uniquely submitted by the IMSGC and were selected from the most recent meta-analysis and existing GWAS data, published or preliminary.(60;106) These loci are listed in **Table 3.2**, along with the corresponding SNP(s) in the locus that reached genome-wide significance in a previous study.

Many other loci on the chip are of direct interest to MS, such as the MHC, and were submitted by multiple groups. The additional loci and individual SNPs covered on the chip are of interest to MS in that they have been implicated in other autoimmune diseases and are, therefore, plausible candidates for MS.

Table 3.2 Loci fine-mapped to completion on the ImmunoChip (submitted only by the IMSGC)

Gene	**SNP**
CD58	rs2300747
CD6	rs17824933
CLEC16A	rs11865121, rs12708716
CYP27B1	rs703842
ELMO1	rs11984075
EVI5	rs11808092
IFI30	rs874628
IL12A	rs4680534
IL2RA	rs2104286
IL7R	rs6897932
IRF8	rs17445836
KIF21B	rs12122721
MPHOSPH9	rs1790100
PKIA	rs967426
RGS1	rs2760524
STAT3	rs744166
TMEM39A	rs1132200
TNFRSF1A	rs1800693
TRIM40	rs2285797
TYK2	rs34536443
UBASH3B	rs7127978

Sample Selection and Genotyping

Samples were requested from BioVU (using the algorithms described in Chapter II) in three rounds as DNA became available for additional individuals identified by our algorithms. (**Table 3.3**) In the initial round of extraction, samples selected by the "possible type 1" were also manually reviewed. Individuals for whom a diagnosis of MS was not considered were excluded, to create a set of 139 samples identified by this algorithm. These samples all individuals identified by the other algorithms were extracted. Genotyping occurred at Vanderbilt University on the Illumina iScan in three batches over an 18 month period. All available case samples were pulled in the first round; subsequent pulls of DNA for cases were completed from newly acquired blood samples.

Additional controls were also genotyped the second round of genotyping. These samples were identified via the algorithm described in Chapter II and matched with cases for age, BMI, and gender. Individuals were first selected at a 2:1 ratio, then an additional 595 samples were randomly selected from the remaining pool of control samples matched for additional genotyping. Samples that failed genotyping or QC in the first two rounds were re-genotyped in subsequent rounds when enough DNA was available. Samples were plated on 96-well plates with interplate and intraplate duplicates, and HapMap trios of various ethnicities (79 samples, 32 unique individuals). Cases and controls were randomized throughout the plates.

Table 3.3 Samples extracted for each round of genotyping

Algorithm	Round 1	Round 2	Round 3
Case Definitive 1	713	58	165
Case Definitive 2	2	1	2
Case Possible 1	139	28	42
Case Possible 2	45	10	16
Controls Original Matching	1677	167	0
Controls Extended Matching	19	1	0
Controls No BMI Matching	102	26	0
Non-matched Controls	0	595	0

SNP calling

192,402 autosomal SNPs were successfully genotyped in 3,808 experimental samples, which were called in conjunction with the IMSGC ImmunoChip dataset using OptiCall.(107) Samples in the IMSGC dataset were called in three groups at the WTSI. The samples genotyped at VUMC were called among 13,049 samples, including samples genotyped at Boston, France, Germany, Miami, and Virginia.(61)

The selection of a calling algorithm to determine the genotypes of the samples was challenging due to the extensive fine-mapping and abundance of SNPs with low minor allele frequency (MAF). Several algorithms were evaluated, and genotypes for analysis were called at the WTSI in conjunction with other IMSGC samples. At the time the first VUMC samples were genotyped, the three prevailing and available calling algorithms were GenCall, GenoSNP, and Illuminus.(108;109) These three algorithms have been compared with respects to their accuracy.(110) Succinctly, Illuminus performs best when calling common variants, although GenoSNP performs better for rare variants. GenCall is the only approach that allows for calling of the X and Y chromosomes.

Initially, calling was performed using both Illuminus and GenoSNP. SNPs highly concordant between the two calling algorithms were kept and extensive evaluation of the data was used to create a set of rules to determine which calls to keep for discordant SNPs, as well as determining which SNPs should be excluded from downstream analyses. At length, problems with the Illuminus algorithm were discovered. Concurrent to these issues, a new calling algorithm, OptiCall, had been released, which was better suited for this set of SNPs.(107) All IMSGC samples were recalled using OptiCall and these calls were used in the final analysis. For samples genotyped at VUMC, X chromosome SNPs and HapMap samples were called using the Illumina GenCall algorithm to check the gender assignment of all samples and concordance of HapMap samples.

Quality Control

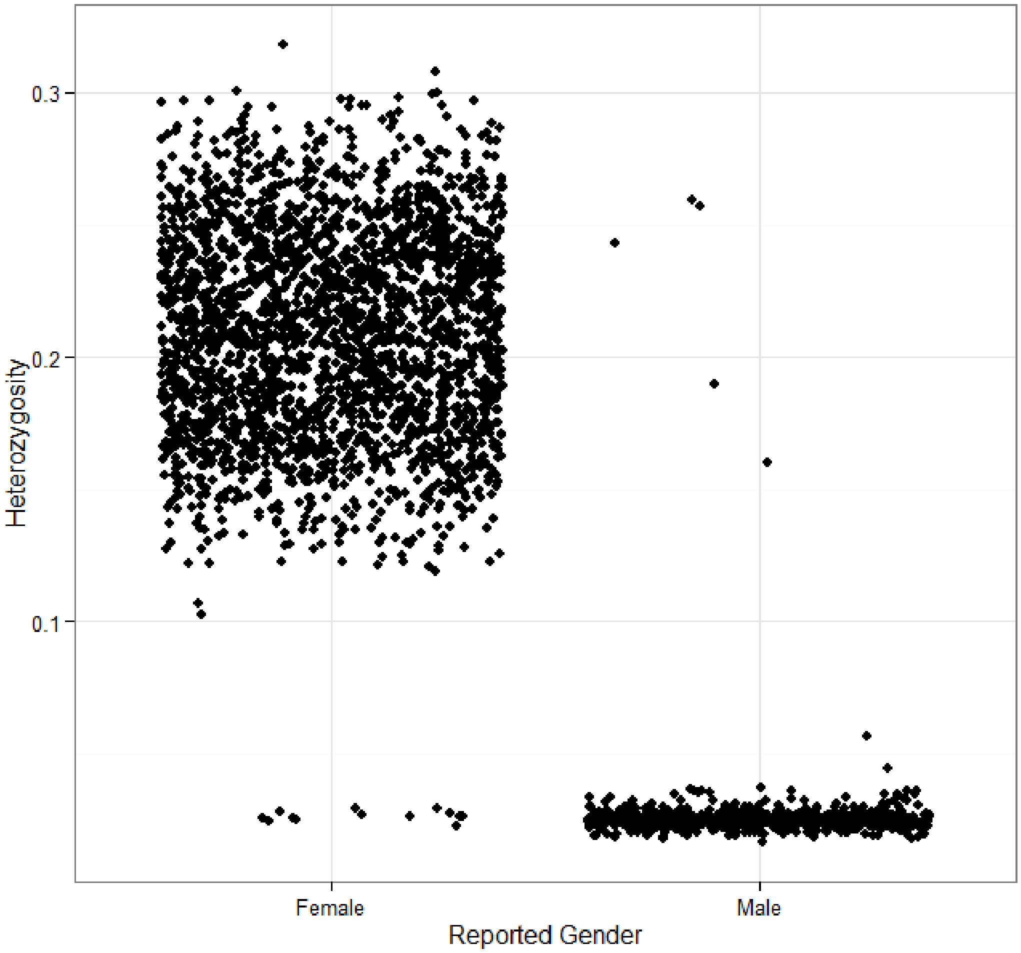
Strict sample and SNP QC measures were employed to ensure highest quality genotyping data was analyzed. 325 samples on poorly performing chips and individuals who later opted-out of BioVU were removed after calling in OptiCall. Of the samples on poorly performing chips, those with sufficient DNA concentration were re-genotyped in later batches. PLINK (111) was used to evaluate concordance, allele frequency, SNP and sample call rates, and relationships (identity by state, IBS). PLATO (112) was used to check reported genders by evaluating X chromosome heterozygosity. Individuals with heterozygosity less than 5% were presumed male and individuals with heterozygosity more than 5% were presumed female.

Concordance between all duplicate HapMap samples was observed (>96%) and used to confirm correct plate orientation and efficient genotyping. Concordance to SNPs also in 1000 Genomes data was observed. The ImmunoChip contains a vast number of low frequency SNPs (67,906 autosomal SNPs with MAF < 0.05 in our raw dataset); these low frequency SNPs are difficult to evaluate in a small dataset. This obstacle was overcome by utilizing the SNP QC of the IMSGC ImmunoChip analysis in the hope of eliminating some problems we would not have been able to identify with our smaller dataset, especially as the ImmunoChip is a custom-designed chip with less stringent QC in the chip design that standard chips. QC in the IMSGC dataset was stratified by country origin of the samples. 35,102 SNPs that failed QC in the US dataset in this analysis were removed prior to additional QC in our dataset.(61) These SNPs were excluded due to duplicate assays (801), cluster failures (151), mendelian errors (4,226), low call rate (7,558), Hardy-Weinberg disequilibrium (3,908), monomorphism (14,658), or differential missingness between cases and controls (3,800).

In the remaining 157,300 SNPs and 3,483 samples, we performed additional QC measures. 164 samples identified as “compromised” blood samples in BioVU due to other conditions, including the presence of plasma cells in the blood, severe combined immunodeficiency (SCID), myeloma, Hodgkin’s lymphoma, lymphoma, leukemia, polycythemia vera, myelofibrosis, transfusions, and other neoplasms, were excluded. Each of these conditions could potentially affect the reliability of the DNA extracted from the blood.

After exclusion of 826 SNPs with a call rate of less than 99% in our dataset, we additionally excluded 46 samples with call rates less than 99%. 19 samples with inconsistent reported and genetic genders based on X chromosome heterozygosity were excluded (13 samples reported female and 6 samples reported male). (**Figure 3.1**) One reported female was found to have a transgender operation recorded in the medical record, but gender in the records for all other samples were consistent throughout.

Figure 3.1 X chromosome heterozygosity for all samples



Pairwise IBS was calculated to identify duplicate samples and ascertain any related individuals. One of the 63 expected duplicate pairs did not show concordance and two pairs of unexpected duplicates were identified. One of the samples in the discordant reported duplicate pair was found to be an unexpected duplicate of another sample. All five samples involved in these errors were excluded. For each of expected duplicate pairs, the sample with the lowest call rate was dropped. At this point, our dataset was frozen and contained 156,474 SNPs and 3,189 samples (1,004 cases, 2,184 controls). Demographics of these samples are shown in **Table 3.4**.

Table 3.4 Demographics for samples in the frozen QC dataset

Demographic	**Cases**, *n*	**Controls**, *n*	**Total***, n*
Gender			
Female	787	1723	2510
Male	216	461	677
Age			
Median	51	52	52
Range	20-92	20-92	20-92
Known deceased	39	120	159
Ethnicity			
White	730	1633	2363
Black	102	205	307
Asian	0	2	2
Hispanic	1	4	5
Native American	0	3	3
Unknown	170	337	507

One of the cases was found to have conflicting information regarding a diagnosis of MS during the manual review evaluation of the case algorithms described in Chapter II. The phenotype of this individual was updated to ambiguous based on this information. The records of all individuals identified by "possible type 1" and "possible type 2" algorithms were reviewed manually to determine if a definite diagnosis of MS by a clinician had been given. Only individuals with a clinical diagnosis were used for subsequent analyses; 130 individuals with only possible or probable diagnoses were excluded. Six sibling pairs, five parent-offspring pairs, and one cousin pair were observed based on IBS. Due to the de-identified nature of BioVU, further confirmation of these relationships was not possible. In each analysis performed for MS risk and clinical traits in Chapter IV, we dropped the individual with the lowest call rate from each relationship pair present. For all analyses, SNPs with MAF < 0.05 were removed in the respective datasets. HWE was calculated and SNPs out of HWE were flagged at $p < 1 \times 10^{-4}$ and evaluated further if they were significant in any analysis.

Various ethnicities are represented in our dataset and a substantial number of our samples (507) have no reported ethnicity. When ignored, population substructure can cause spurious results in analyses. This can be combated two ways: samples can be stratified by ethnicity and analyzed separately or covariates representing the differences in population structure can be included in analyses. Due to the relatively small size of our dataset, we opted to include covariates and utilize all samples in each analysis. To this end, Eigenstrat analysis was run on the frozen dataset.(113) Plots showing the stratification of the samples using the first three principle components are displayed in **Figure 3.2**. The first and second principle components separate whites, blacks, and Hispanics, as identified in BioVU, into three groups (although there are few Hispanic samples in this dataset). The first and third principle components again separate white and black populations. The second and third principle components tease apart the Asian and Hispanic samples in the dataset. All analyses described in Chapter IV were adjusted for the first three principal components.

Figure 3.2 Principle components for all samples
(*a) eigenvalues 1 and 2* (b) eigenvalues 1 and 3 (c) eigenvalues 2 and 3

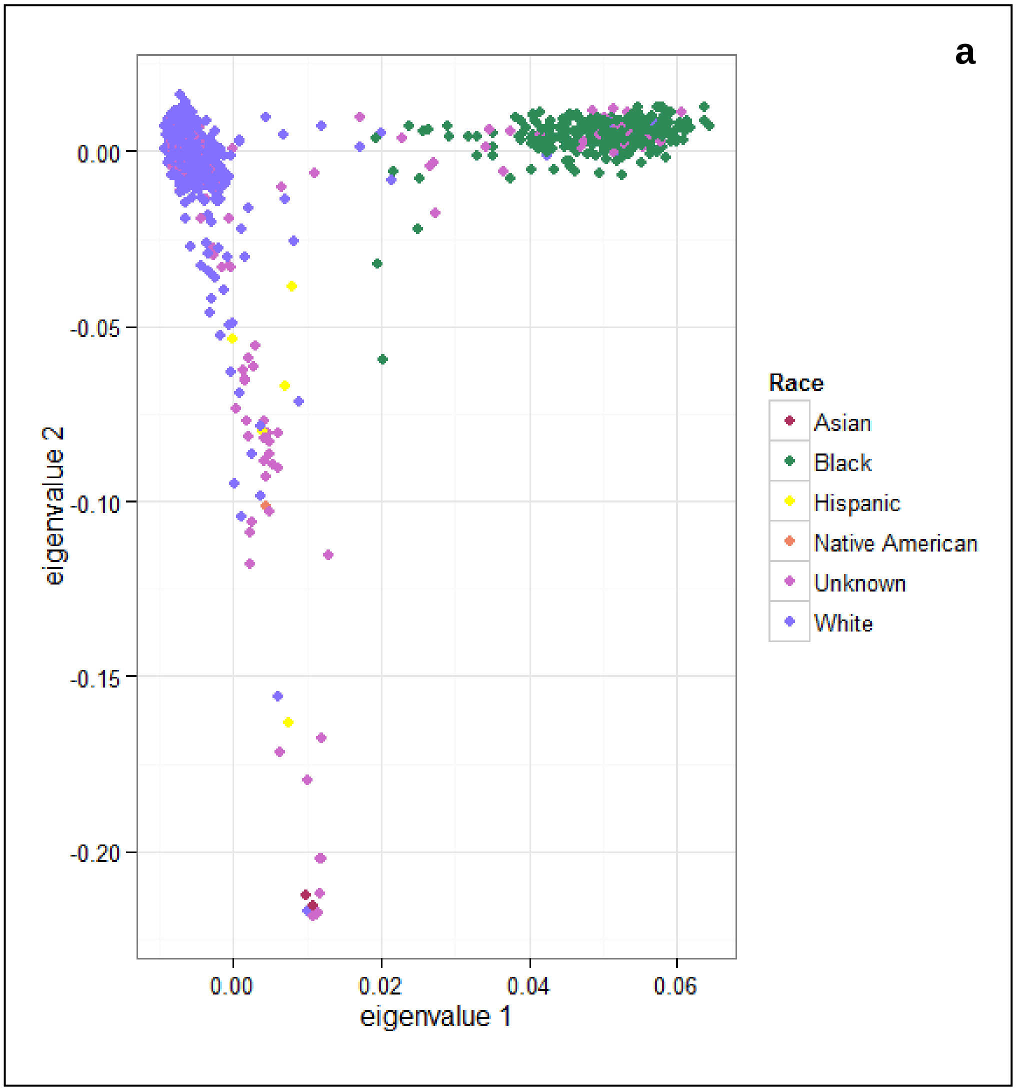


Figure 3.2 Principle components for all samples
(a) eigenvalues 1 and 2 *(b) eigenvalues 1 and 3* (c) eigenvalues 2 and 3

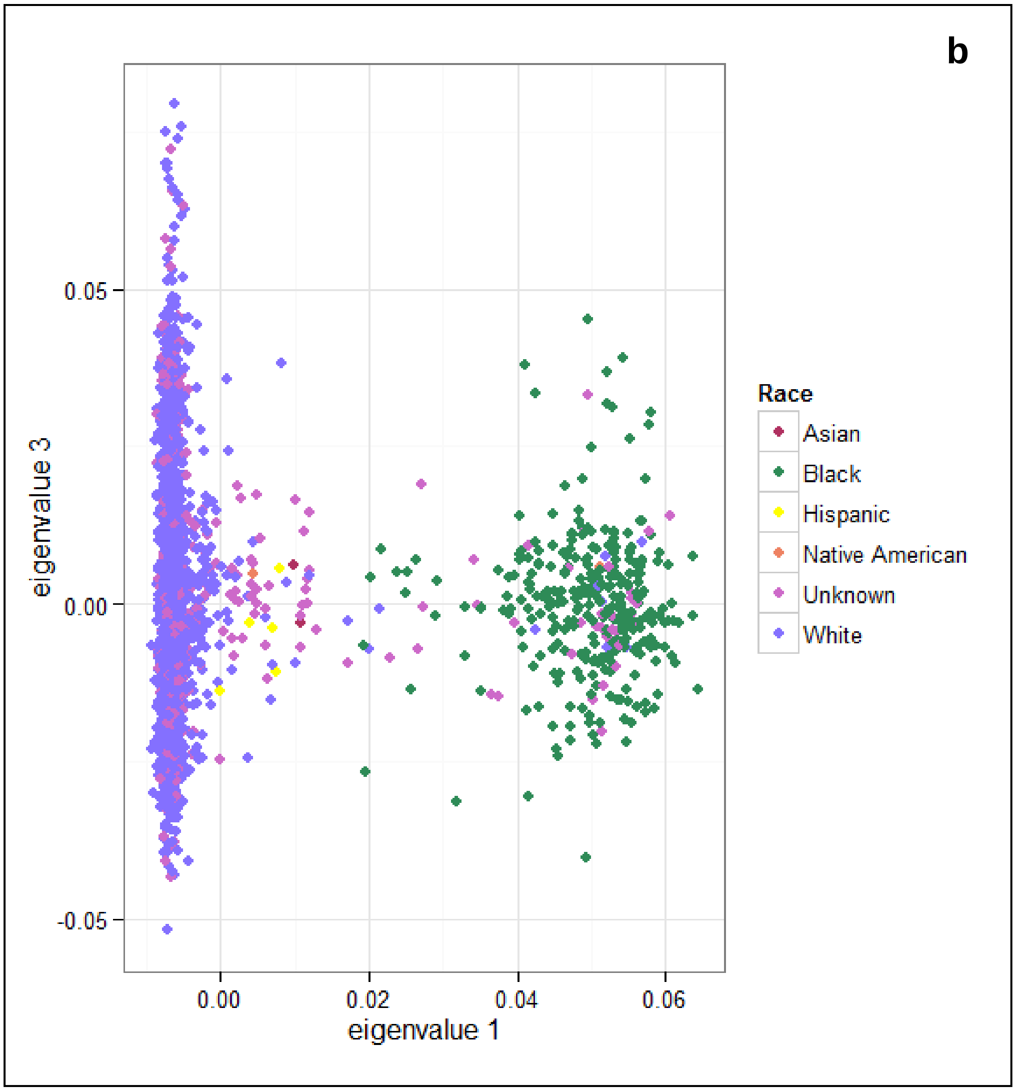
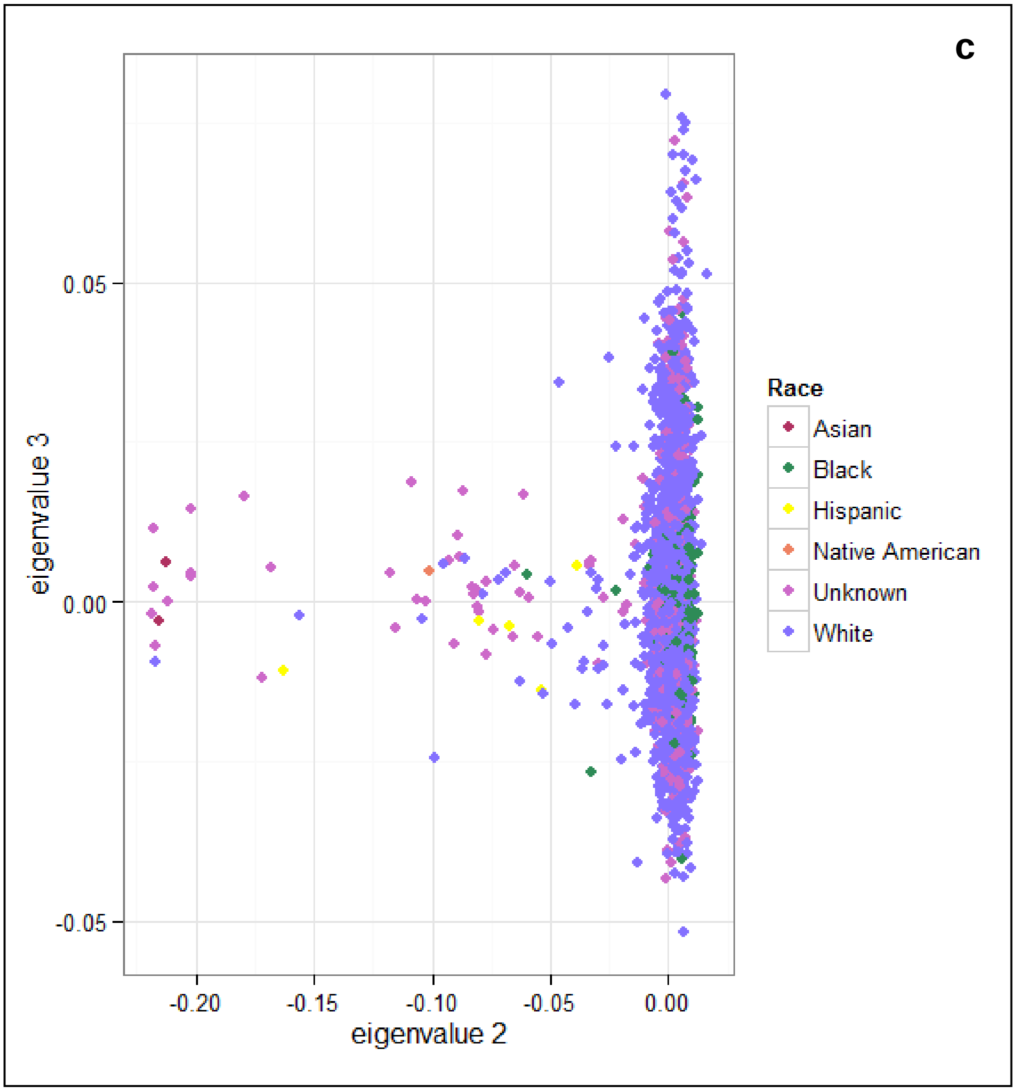


Figure 3.2 Principle components for all samples
(a) eigenvalues 1 and 2 (b) eigenvalues 1 and 3 *(c) eigenvalues 2 and 3*



CHAPTER IV

GENETIC ANALYSES

Through the work described in Chapters II and III, we created two large datasets, one of phenotype data and one of genotype data. After quality control of these data, we combined the phenotype and genotype data to conduct genetic analyses. Regression analyses were performed to test the significance of known MS loci in our dataset and for association to the clinical traits of MS. Genetic risk scores based on the 110 replicated MS-associated SNPs and a representative SNP from the HLA region were calculated for all individuals. Details of the types of analyses and results are discussed in this chapter.

Case-control analysis

Our first analysis was to confirm known MS loci in our dataset. As these samples were selected via algorithms using medical records as opposed to first-hand interviews, we wanted to further validate the samples by confirming that this MS population shares the same genetic landscape as other confirmed datasets. Initial confirmation of MS loci was done in a pilot study of BioVU.(73) However, that study was done with 90 samples, which were all reviewed manually by a clinician. We have a much larger dataset that would be impossible to hand curate. Combined with the fact that many more loci are known to be associated with MS at the present time than at the time of the pilot study, we would expect to have greater ability to confirm associations with many loci.

110 MS-associated SNPs representing 103 regions, excluding the MHC, as published by 2013 are located on the ImmunoChip. We extracted 105 of these SNPs; the remaining five did not pass QC (rs4679081, rs7769192, rs2150702, rs533646, and

rs2744148). Samples identified by the two "possible" case algorithms that did not have a definitive diagnosis of MS upon manual review were not included, as discussed in Chapter III. We made no restrictions upon minor allele frequency at this point, although power calculations, as discussed below, were performed to take into account the likely ability we would have to detect effects with low minor allele frequency. We performed logistic regression analysis for MS disease status for all 105 SNPs in PLINK. (111) Logistic regression was also run on chromosome 6 to investigate the HLA region. The first three principle components as determined by Eigenstrat were included as covariates. (113) The inclusion of principle components allowed us to include all samples in this analysis regardless of ethnicity.

The most significant SNP in the HLA region was rs9271775, with p = 9.73 x 10^{-30}. 45 of the other 105 known SNPs were significant at a one-sided p-value of 0.05. (**Table 4.1**) We used a one-sided p-value for significance because we were looking for not only replication of an effect, but also the direction of the effect (concurrent with the direction of published results for each SNP). The effects of the significant SNPs were in the same direction as in the IMSGC ImmunoChip analysis.(62) 44 of the remaining 60 SNPs were in the correct direction. While these were not significant, the direction of their effects was consistent with previous studies, providing evidence that the underlying MS genetic architecture in our dataset is similar to previously reported datasets. Under a binomial distribution, this is significantly greater than the number of SNPs we would expect to see in the correct direction by chance (p = 1.97 x 10^{-4}).

Table 4.1 Logistic regression results for 110 known MS SNPs

CHR	ImmunoChip ID	rsID	BP	MA	MAF	OR	P	Power	IC OR	IC RAF	IC RA	Allele	Direction
1	imm_1_2515525	rs3748817	2515525	G	0.38	1.03	0.61	0.70	1.14	0.64	A	opp	opp
1	rs3007421	rs3007421	6452776	A	0.16	1.03	0.67	0.40	1.12	0.12	A	same	same
1	rs12087340	rs12087340	85519581	A	0.08	1.16	0.14	0.66	1.22	0.09	A	same	same
1	rs11587876	rs11587876	85687771	G	0.20	0.94	0.37	0.47	1.12	0.79	A	opp	same
1	1kg_1_92748052	rs41286801	92748052	A	0.14	1.21	0.02	0.76	1.20	0.14	A	same	same
1	1kg_1_101013481	rs7552544	101013481	G	0.41	0.90	0.08	0.39	1.08	0.56	A	opp	same
1	rs11581062	rs11581062	101180107	G	0.29	1.05	0.46	0.20	1.05	0.29	G	same	same
1	imm_1_116881689	rs6677309	116881689	C	0.16	0.73	0.00	0.94	1.34	0.88	A	opp	same
1	rs666930	rs666930	120060493	A	0.45	0.90	0.06	0.47	1.09	0.53	G	opp	same
1	rs2050568	rs2050568	156036865	A	0.49	0.99	0.89	0.41	1.08	0.53	G	opp	same
1	imm_1_158978428	rs35967351	158978428	T	0.31	0.98	0.75	0.40	1.09	0.67	A	opp	same
1	imm_1_190808095	rs1359062	190808095	G	0.20	0.99	0.93	0.69	1.18	0.82	C	opp	same
1	imm_1_199141351	rs55838263	199141351	G	0.28	0.97	0.67	0.58	1.12	0.71	A	opp	same
2	rs4665719	rs4665719	24871364	G	0.27	1.10	0.16	0.37	1.09	0.25	G	same	same
2	1kg_2_43214760	rs2163226	43214760	G	0.30	0.86	0.02	0.47	1.10	0.71	A	opp	same
2	imm_2_60948749	rs842639	60948749	G	0.34	1.05	0.42	0.56	1.11	0.65	A	opp	opp
2	imm_2_68440981	rs7595717	68440981	A	0.27	1.21	0.00	0.44	1.10	0.26	A	same	same
2	rs17174870	rs17174870	112381672	A	0.23	0.88	0.07	0.13	1.03	0.76	G	opp	same
2	imm_2_191682680	rs9967792	191682680	A	0.38	0.88	0.04	0.54	1.11	0.62	G	opp	same
2	imm_2_230823698	rs9989735	230823698	C	0.18	1.04	0.64	0.71	1.17	0.18	C	same	same
3	1kg_3_18760589	rs11719975	18760589	C	0.27	1.04	0.57	0.38	1.09	0.27	C	same	same
3	1kg_3_27732022	rs2371108	27732022	A	0.40	1.05	0.38	0.39	1.08	0.38	A	same	same
3	1kg_3_28053575	rs1813375	28053575	A	0.45	1.04	0.56	0.81	1.15	0.47	A	same	same
3	rs9828629	rs9828629	71613036	A	0.36	0.95	0.36	0.37	1.08	0.62	G	opp	same
3	rs2028597	rs2028597	107041527	A	0.08	1.05	0.65	0.11	1.04	0.92	G	opp	opp

Table 4.1 *continued*

CHR	ImmunoChip ID	rsID	BP	MA	MAF	OR	P	Power	IC OR	IC RAF	IC RA	Allele	Direction
3	imm_3_120705146	rs1131265	120705146	G	0.18	0.85	0.03	0.77	1.19	0.80	C	opp	same
3	rs1920296	rs1920296	123026267	A	0.35	0.85	0.01	0.73	1.14	0.64	C	opp	same
3	rs2255214	rs2255214	123253229	A	0.45	0.91	0.09	0.59	1.11	0.52	C	opp	same
3	rs9282641	rs9282641	123279458	A	0.07	0.77	0.02	0.28	1.12	0.92	G	opp	same
3	imm_3_161173806	rs1014486	161173806	G	0.42	0.97	0.62	0.55	1.11	0.43	G	same	opp
4	rs7665090	rs7665090	103770651	A	0.47	0.87	0.02	0.39	1.08	0.52	G	opp	same
4	rs2726518	rs2726518	106392648	A	0.39	0.92	0.16	0.43	1.09	0.55	C	opp	same
5	imm_5_35914913	rs6881706	35914913	A	0.26	0.85	0.02	0.53	1.12	0.72	C	opp	same
5	imm_5_40434853	rs6880778	40434853	A	0.39	0.93	0.19	0.50	1.10	0.60	G	opp	same
5	imm_5_55476487	rs71624119	55476487	A	0.22	0.89	0.10	0.54	1.12	0.76	G	opp	same
5	rs756699	rs756699	133474474	G	0.15	1.01	0.90	0.37	1.12	0.87	A	opp	opp
5	imm_5_141486748	none	141486748	A	0.37	1.00	0.94	0.32	1.07	0.61	C	opp	opp
5	imm_5_158692478	rs2546890	158692478	G	0.49	0.98	0.77	0.26	1.06	0.52	A	opp	same
5	rs4976646	rs4976646	176721176	G	0.38	1.12	0.05	0.67	1.13	0.34	G	same	same
6	rs17119	rs17119	14827475	G	0.22	0.97	0.67	0.39	1.11	0.81	A	opp	same
6	rs9271775	rs9271775	32520558	G	0.23	2.23	0.00	1.00	2.63	0.17	G	same	same
6	rs941816	rs941816	36483282	G	0.19	1.02	0.84	0.52	1.13	0.18	G	same	same
6	imm_6_91033489	rs72928038	91033489	A	0.15	0.96	0.63	0.43	1.11	0.17	A	same	opp
6	imm_6_128320491	rs802734	128320491	G	0.28	1.06	0.35	0.11	1.03	0.69	A	opp	opp
6	rs11154801	rs11154801	135781048	A	0.34	1.12	0.07	0.54	1.11	0.37	A	same	same
6	rs17066096	rs17066096	137494601	G	0.23	1.08	0.25	0.63	1.14	0.23	G	same	same
6	imm_6_138286509	rs67297943	138286509	G	0.22	0.97	0.68	0.48	1.12	0.78	A	opp	same
6	imm_6_159390547	rs212405	159390547	A	0.37	0.90	0.07	0.77	1.15	0.62	T	opp	same
7	rs1843938	rs1843938	3079560	A	0.43	1.02	0.72	0.40	1.08	0.44	A	same	same
7	imm_7_26981513	rs706015	26981513	C	0.19	1.30	0.00	0.56	1.14	0.18	C	same	same

Table 4.1 *continued*

CHR	ImmunoChip ID	rsID	BP	MA	MAF	OR	P	Power	IC OR	IC RAF	IC RA	Allele	Direction
7	imm_7_28139264	rs917116	28139264	C	0.28	1.19	0.01	0.48	1.12	0.20	C	same	same
7	1kg_7_37348990	rs60600003	37348990	C	0.09	1.21	0.05	0.51	1.16	0.10	C	same	same
7	imm_7_50296113	rs201847125	50296113	A	0.26	0.95	0.45	0.51	1.11	0.70	G	opp	same
7	rs354033	rs354033	148920397	A	0.24	1.07	0.29	0.13	1.03	0.74	G	opp	opp
8	1kg_8_79738359	rs1021156	79738359	A	0.30	1.12	0.09	0.55	1.12	0.24	A	same	same
8	rs2456449	rs2456449	128262163	G	0.34	1.02	0.72	0.50	1.10	0.36	G	same	same
8	rs4410871	rs4410871	128884211	A	0.26	0.93	0.28	0.55	1.12	0.72	G	opp	same
8	imm_8_129228127	rs759648	129228127	C	0.31	1.19	0.00	0.40	1.09	0.31	C	same	same
10	imm_10_6139051	rs2104286	6139051	G	0.24	0.92	0.21	0.90	1.21	0.72	A	opp	same
10	rs793108	rs793108	31455112	A	0.46	0.99	0.91	0.48	1.09	0.50	A	same	opp
10	rs2688608	rs2688608	75328355	C	0.49	0.89	0.04	0.32	1.07	0.55	A	opp	same
10	imm_10_80718617	rs1782645	80718617	A	0.42	1.10	0.10	0.45	1.09	0.43	A	same	same
10	rs7923837	rs7923837	94471897	A	0.35	0.90	0.09	0.53	1.11	0.61	G	opp	same
11	rs7120737	rs7120737	47658971	G	0.15	1.13	0.11	0.47	1.13	0.15	G	same	same
11	imm_11_60549906	rs34383631	60549906	A	0.41	1.15	0.02	0.57	1.11	0.40	A	same	same
11	rs694739	rs694739	63853809	G	0.34	0.92	0.18	0.35	1.08	0.62	A	opp	same
11	imm_11_118230104	rs9736016	118230104	A	0.39	0.86	0.01	0.49	1.10	0.63	T	opp	same
11	imm_11_118260948	rs523604	118260948	G	0.49	0.99	0.92	0.45	1.09	0.53	A	opp	same
12	imm_12_6310270	rs1800693	6310270	G	0.41	1.16	0.01	0.76	1.14	0.40	G	same	same
12	ccc-12-6373761-G-C	rs12296430	6373761	C	0.18	1.13	0.10	0.57	1.14	0.19	C	same	same
12	imm_12_9796957	rs11052877	9796957	G	0.35	1.08	0.18	0.52	1.10	0.36	G	same	same
12	seq-t1d-12-56468329-A-T	rs201202118	56468329	T	0.29	0.88	0.05	0.67	1.14	0.67	A	opp	same
12	imm_12_122159335	rs7132277	122159335	A	0.17	1.12	0.14	0.39	1.10	0.19	A	same	same
13	1kg_13_98884260	rs4772201	98884260	G	0.19	0.83	0.02	0.46	1.12	0.82	A	opp	same
14	imm_14_68331225	rs2236262	68331225	G	0.46	0.93	0.18	0.36	1.08	0.50	A	opp	same

Table 4.1 *continued*

CHR	ImmunoChip ID	rsID	BP	MA	MAF	OR	P	Power	IC OR	IC RAF	IC RA	Allele	Direction
14	rs4903324	rs4903324	75031264	A	0.19	1.01	0.87	0.36	1.10	0.19	A	same	same
14	1kg_14_87502081	rs74796499	87502081	A	0.05	0.67	0.01	0.59	1.31	0.95	C	opp	same
14	rs12148050	rs12148050	102333541	A	0.41	1.00	1.00	0.35	1.08	0.35	A	same	opp
15	imm_15_76994521	rs59772922	76994521	G	0.23	0.88	0.07	0.37	1.11	0.83	A	opp	same
15	rs8042861	rs8042861	88778337	C	0.50	0.97	0.64	0.41	1.08	0.44	A	opp	same
16	imm_16_11102272	rs12927355	11102272	A	0.30	0.79	0.00	0.93	1.21	0.68	G	opp	same
16	imm_16_11196307	rs4780346	11196307	A	0.23	1.19	0.01	0.37	1.09	0.23	A	same	same
16	imm_16_11343491	rs6498184	11343491	A	0.16	1.05	0.54	0.61	1.15	0.81	G	opp	opp
16	rs7204270	rs7204270	30064464	G	0.48	1.13	0.03	0.46	1.09	0.50	G	same	same
16	imm_16_67243406	rs1886700	67243406	A	0.14	1.11	0.21	0.35	1.11	0.14	A	same	same
16	rs12149527	rs12149527	77668097	A	0.44	1.16	0.01	0.40	1.08	0.47	A	same	same
16	rs7196953	rs7196953	78206895	A	0.31	0.94	0.37	0.33	1.08	0.29	A	same	opp
16	imm_16_84551985	rs35929052	84551985	A	0.10	0.80	0.02	0.42	1.14	0.89	G	opp	same
17	imm_17_35165903	rs12946510	35165903	A	0.44	1.01	0.87	0.36	1.08	0.47	A	same	same
17	imm_17_37784289	rs4796791	37784289	A	0.42	1.09	0.14	0.50	1.10	0.36	A	same	same
17	rs4794058	rs4794058	42952097	A	0.49	1.12	0.05	0.35	1.07	0.50	A	same	same
17	rs8070345	rs8070345	55171539	G	0.49	0.89	0.05	0.77	1.14	0.45	A	opp	same
18	rs7238078	rs7238078	54535172	C	0.24	1.01	0.90	0.20	1.05	0.77	A	opp	opp
19	rs1077667	rs1077667	6619972	A	0.19	0.90	0.16	0.69	1.16	0.79	G	opp	same
19	imm_19_10324118	rs34536443	10324118	G	0.04	0.87	0.32	0.56	1.28	0.95	C	opp	same
19	rs2288904	rs2288904	10603170	A	0.21	1.01	0.88	0.59	1.14	0.77	G	opp	opp
19	rs1870071	rs1870071	16366106	G	0.32	1.16	0.02	0.56	1.12	0.29	G	same	same
19	chr19_18146944	rs11554159	18146944	A	0.26	0.88	0.05	0.70	1.15	0.73	G	opp	same
19	rs8107548	rs8107548	54562455	G	0.26	1.12	0.08	0.39	1.09	0.25	G	same	same
20	imm_20_44181354	rs4810485	44181354	A	0.24	1.12	0.08	0.34	1.08	0.25	A	same	same

Table 4.1 *continued*

CHR	ImmunoChip ID	rsID	BP	MA	MAF	OR	P	Power	IC OR	IC RAF	IC RA	Allele	Direction
20	imm_20_47872168	rs17785991	47872168	A	0.32	1.05	0.45	0.43	1.09	0.35	A	same	same
20	rs2248359	rs2248359	52224925	A	0.43	0.95	0.40	0.31	1.07	0.60	G	opp	same
20	imm_20_61844427	rs2256814	61844427	A	0.18	1.09	0.24	0.44	1.11	0.19	A	same	same
20	rs6062314	rs6062314	61880157	G	0.09	0.93	0.45	0.22	1.10	0.92	A	opp	same
22	rs2283792	rs2283792	20461125	A	0.46	0.90	0.05	0.41	1.08	0.51	C	opp	same
22	rs470119	rs470119	49313780	A	0.42	0.96	0.50	0.29	1.07	0.39	A	same	opp

Power calculations are for one-sided p-values of 0.05; PLINK logistic regression outputs two-sided p-values. Hence, SNPs with p-values less than 0.10 in this table are significant at a one-sided p-value of 0.05.
CHR: chromosome; BP: base pairs; MA: minor allele; MAF: minor allele frequency; OR: odds ratio; P: two-sided p-value produced by PLINK; Power: power to detect the effect in our dataset using the ImmunoChip OR and RAF; IC OR: odds ratio found by the IMSGC ImmunoChip analysis; IC RAF: risk allele frequency in the IMSGC ImmunoChip dataset; IC RA: risk allele in the IMSGC ImmunoChip analysis; Allele: If the alleles tested in our analysis and the IMSGC analysis are the same; Direction: If the direction of effect is the same in our analysis and the IMSGC analysis; opp: opposite allele or direction, depending on column; same: same allele or direction, depending on column.

To determine if we identified as many SNPs as we would expect, we calculated our power to detect each effect based on our sample size using a one-sided alpha value of 0.05 and the odds ratio and risk allele frequencies observed in the IMSGC ImmunoChip analysis. Power calculations were performed in QUANTO.(114;115) Our power to detect each risk allele is listed in **Table 4.1**. Based on these power calculations, we would have expected 50.5 SNPs to be significant at this level in this analysis, close to the number we detected. We did not calculate a p-value, but based on the number of significant SNPs and other SNPs that trended in the correct direction, these data strongly confirm a similar underlying MS genetic architecture of the case individuals selected via algorithm from the EMR to those previously reported.

Genetic risk scores

We created a genetic risk score based on the known MS loci for each sample. To calculate this, we gave each SNP a weight based on the proportion of the total sum of the odds ratios. For each individual, this weight was multiplied by the number of risk alleles the patient carried. The scores for all SNPs were summed to create a genetic risk score. (**Equation 4.1**) This was done with and without inclusion of the most significant SNP in the HLA region in our dataset to ensure that the large effect of the HLA was not masking differences among the SNPs with smaller effects. Scores excluding the HLA are displayed in **Figure 4.1**. There was a highly significant difference between the risk scores for the cases and controls (t-test, two-sample assuming equal variances, p = 2.27 x 10^{-32}).

$$Genetic\ Risk\ Score = \sum_{snp_1}^{snp_n} \frac{OR_{snp_i}}{\sum OR} \times number\ of\ risk\ alleles_{snp_i} \quad (4.1)$$

Figure 4.1 Genetic risk scores excluding the HLA locus

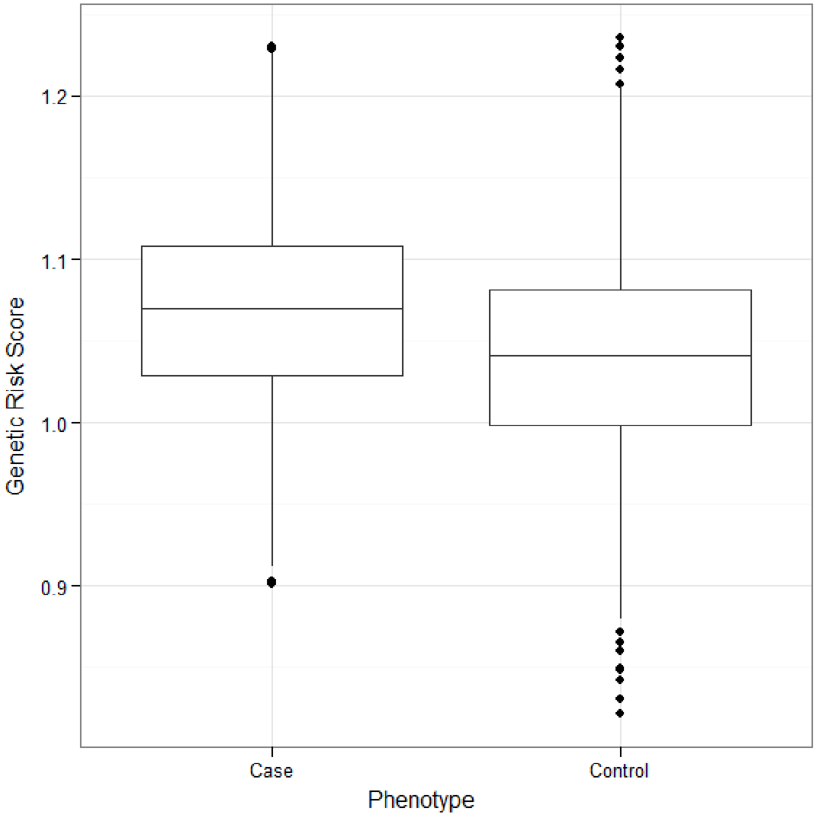
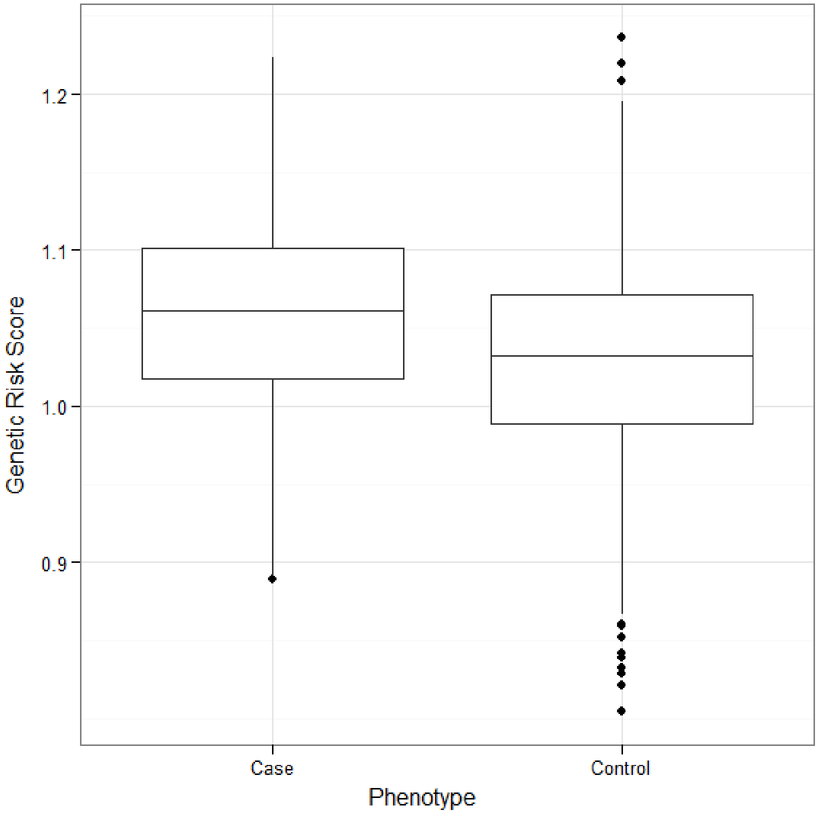


Figure 4.2 Genetic risk scores including the HLA locus



Scores including the most significant HLA SNP are displayed in **Figure 4.2**. Significant differences between the risk scores for the cases and controls were seen (t-test, two-sample assuming equal variance, $p = 1.68 \times 10^{-35}$).

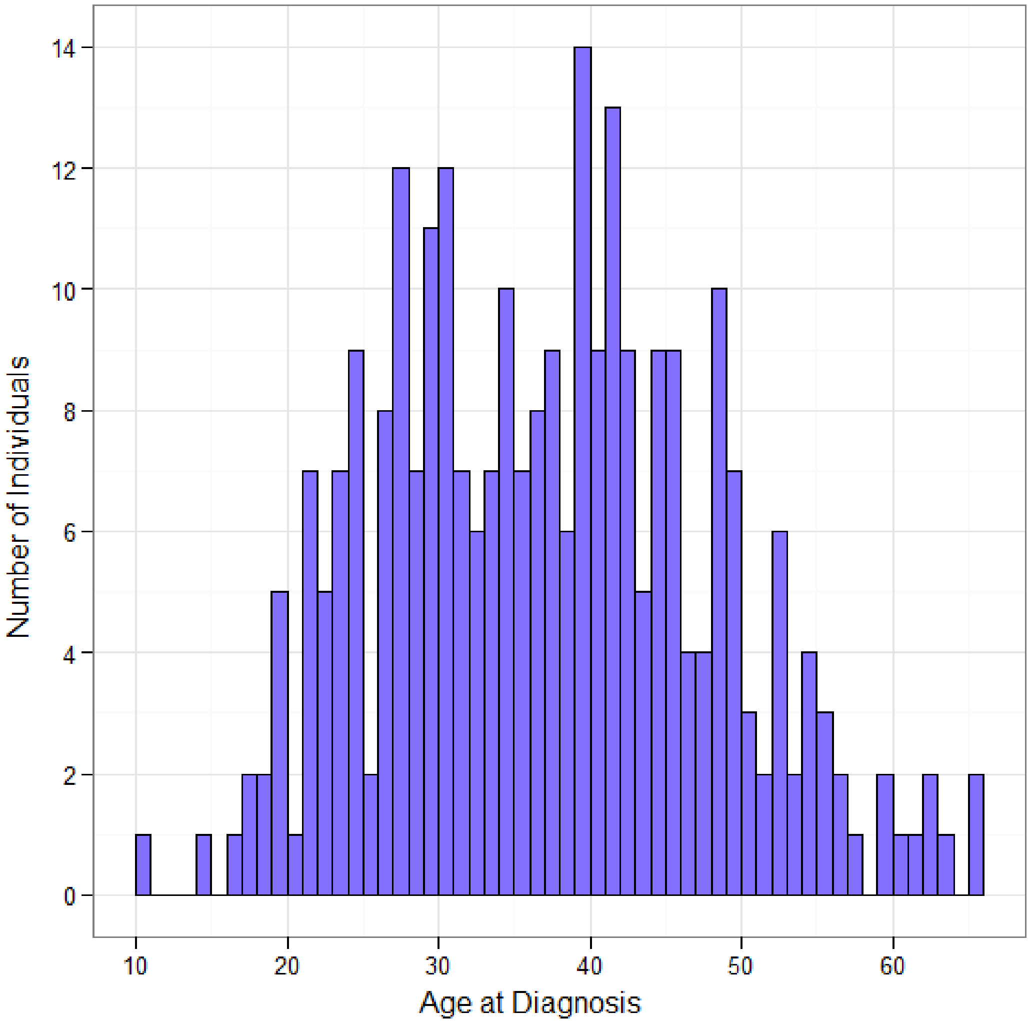
Analyses of clinical traits of disease course

By utilizing the disease course information extracted from the SD as described in Chapter II, we performed regression analyses for seven MS clinical traits. The outcomes evaluated were age at diagnosis, age at first neurological symptom, presence of oligoclonal bands, MSSS, timed 25 foot walk, CNS origin of first neurological symptom, and time to SPMS. The base dataset used for these analyses included all cases from the case-control analysis. Unless otherwise noted, SNPs with p-values less than 1×10^{-4} are shown in the results tables.

Age at diagnosis

The year of diagnosis was extracted via algorithm for 1,061 individuals as described in Chapter II. Birth date for all individuals was extracted. Age at diagnosis was calculated using the year of birth and year of diagnosis for all cases passing QC. 278 individuals had the required information and were used for analysis. The range of ages at diagnosis for these individuals are displayed in **Figure 4.3**. None of these individuals were related based on IBS analysis. 48,744 SNPs with MAF < 0.05 were excluded. Linear regression was performed on the 278 individuals and 107,730 remaining SNPs, with the first three principle components as covariates.

Figure 4.3 Distribution of the ages of diagnosis



The most significant SNPs are displayed in **Table 4.2**. Five of these SNPs cluster in or around *SKAP2*, a gene encoding the src signaling pathway protein. rs73067474 is located intronically; the other SNPs are within 4-7kb upstream. All SNPs in this region are in linkage disequilibrium (LD) ($0.75 < r^2 < 0.90$) are located 300kb upstream from a novel MS locus detected in the IMSGC ImmunoChip analysis (rs706015). Also, SNP rs7804356 in this region (26,858,190bp) has been associated with type 1 diabetes.(116)

Table 4.2 Most significant results for age of diagnosis regression analysis

CHR	ImmunoChip ID	rsID	BP	MA	MAF	BETA	P
7	imm_7_26666501	rs17290008	26666501	A	0.13	-5.85	2.24E-05
7	imm_7_26686349	rs73075509	26686349	A	0.13	-5.85	2.24E-05
7	imm_7_26696265	rs17291131	26696265	A	0.13	-5.85	2.24E-05
7	imm_7_26666964	rs73073796	26666964	A	0.14	-5.60	2.62E-05
2	rs2680836	rs2680836	2.22E+08	A	0.11	5.92	3.08E-05
13	rs12877398	rs12877398	72019306	A	0.12	-5.69	4.69E-05
20	rs388349	rs388349	15322019	A	0.06	7.88	6.99E-05
7	imm_7_26823157	rs73067474	26823157	C	0.12	-5.52	9.58E-05

CHR: chromosome; BP: base pairs; MA: minor allele; MAF: minor allele frequency; P: two-sided p-value

No samples had an age of diagnosis outside of three standard deviations (5.3, 70.0), but analysis after removal of five samples at the extremes of the distribution (18 < age < 58) showed different results (**Table 4.3**). A decrease significant in the SNPs on chromosomes 7 ($p = 3.5 \times 10^{-4}$) and 20 ($p = 1.9 \times 10^{-2}$) suggests these extreme ages were driving their significance. A locus on chromosome 3 not associated previously was significant at $p = 9.42 \times 10^{-6}$. The top four SNPs are in high LD ($r^2 > 0.98$) and are located in *CD80*, a gene involved in T-cell proliferation. rs2228017 is a synonymous variant; the other SNPs are intronic.

Table 4.3 Most significant results for age of diagnosis regression analysis after exclusion of outliers

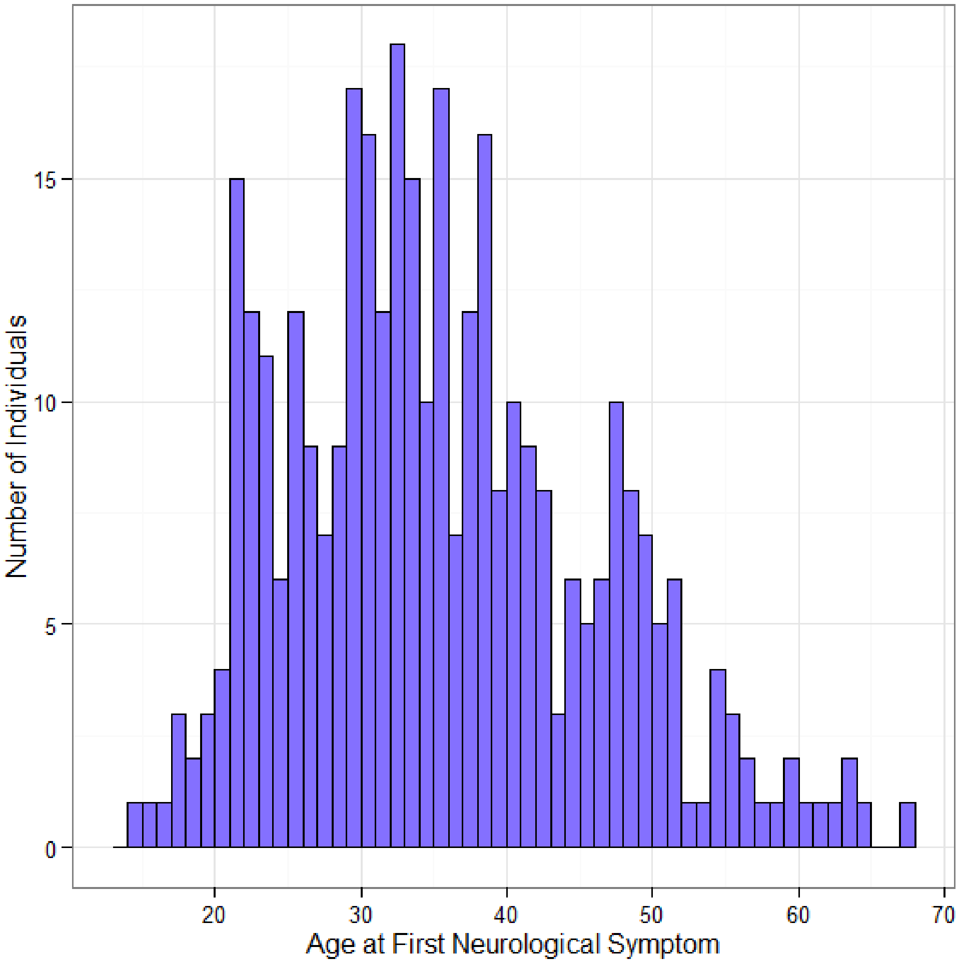
CHR	**ImmunoChip ID**	**rsID**	**BP**	**MA**	**MAF**	**BETA**	**P**
3	imm_3_120739878	rs491407	120739878	G	0.23	-4.55	9.42E-06
3	imm_3_120746370	rs2228017	120746370	A	0.23	-4.55	9.42E-06
3	imm_3_120748830	rs2692620	120748830	G	0.23	-4.55	9.42E-06
3	imm_3_120741492	rs527172	120741492	A	0.23	-4.35	2.35E-05
5	rs7723716	rs7723716	67246186	A	0.31	3.66	5.62E-05
3	rs3913218	rs3913218	189336050	G	0.27	3.63	7.26E-05
5	rs1490790	rs1490790	67243238	G	0.31	3.59	7.75E-05
2	rs2680836	rs2680836	222179704	A	0.11	5.19	8.53E-05
6	rs7767008	rs7767008	28738772	C	0.30	-3.85	8.58E-05
2	rs12621276	rs12621276	37152649	A	0.30	3.58	9.60E-05
5	rs2301010	rs2301010	109160360	G	0.09	-5.74	9.96E-05

CHR: chromosome; BP: base pairs; MA: minor allele; MAF: minor allele frequency; P: two-sided p-value

Age at first neurological symptom

Age at first neurological symptom was calculated using the year of first symptom extracted from the narrative text and the birth year. 349 individuals had genotype data and passed QC. These individuals were not related by IBS analysis. The distribution of these ages can be seen in **Figure 4.4**. The median age was 34 (mean 35 years, range 14-67 years). 49,360 SNPs were excluded for MAF < 0.05, leaving 107,114 SNPs remaining for analysis. Linear regression was performed, with the first three principle components as covariates.

Figure 4.4 Distribution of the ages of first neurological symptom



The most significant results are listed in **Table 4.4**. The SNPs on chromosome 2 are in complete LD in this dataset (r^2=1; D'=1) and are located 550kb downstream from a MS locus discovered in the IMSGC ImmunoChip analysis.(62) rs30749941 on chromosome 10 is 700kb upstream from another novel MS locus. SNP rs1467536 significantly deviates from HWE in the controls of the overall dataset (p = 6.38 x 10^{-9}) but only slight deviation in seen in the individuals in this analysis (p = 0.02).

Table 4.4 Most significant results for age of first neurological symptom regression analysis

CHR	ImmunoChip ID	rsID	BP	MA	MAF	BETA	P
14	rs1467536	rs1467536	100131604	A	0.36	3.69	2.37E-05
10	rs4746003	rs4746003	71208298	A	0.27	-3.58	6.02E-05
10	1kg_10_30749941	rs306582	30749941	C	0.08	5.64	7.96E-05
2	1kg_2_25414511	rs1465764	25414511	T	0.07	6.16	8.79E-05
2	1kg_2_25419411	rs1550117	25419411	A	0.07	6.16	8.79E-05

CHR: chromosome; BP: base pairs; MA: minor allele; MAF: minor allele frequency; P: two-sided p-value

Presence of oligoclonal bands

We performed logistic regression for presence or absence of oligoclonal bands. Results from the structured field and extracted from the narrative text of the SD were used, with preference given to structured field results in cases of discrepancies (see description in Chapter II). Of the 316 genotyped cases that passed QC, 191 were positive for oligoclonal bands and 125 were negative for oligoclonal bands. One apparent parent-offspring pair was identified by pairwise IBS analysis. Both individuals were negative for oligoclonal bands; one subject was removed. 48,420 SNPs with MAF < 0.05 were excluded. Regression analysis of the remaining 108,054 SNPs with the first three principle components was performed. The most significant results of this analysis are displayed in **Table 4.5**. rs6743119 in located in an intron of *PRKCE*, a protein kinase C involved in cellular signaling pathways, including neuron channel activation. rs9271366 is in the HLA region on chromosome 6. This specific SNP has been associated with ulcerative colitis. (117)

Table 4.5 Most significant results for presence of oligoclonal bands

CHR	ImmunoChip ID	rsID	BP	MA	MAF	OR	P
2	rs6743119	rs6743119	45813464	A	0.40	2.13	5.18E-05
10	1kg_10_59518462	rs1759342	59518462	A	0.23	0.47	8.29E-05
6	rs9271366	rs9271366	32694832	G	0.27	2.28	9.30E-05

CHR: chromosome; BP: base pairs; MA: minor allele; MAF: minor allele frequency; OR: odds ratio; P: two-sided p-value

In light of the practices of the MS Clinic in testing for oligoclonal bands primarily at the time of diagnosis of MS, this was essentially an analysis for a genetic association to the presence of oligoclonal bands at time of diagnosis, as opposed to development of oligoclonal bands at any time in the disease course.

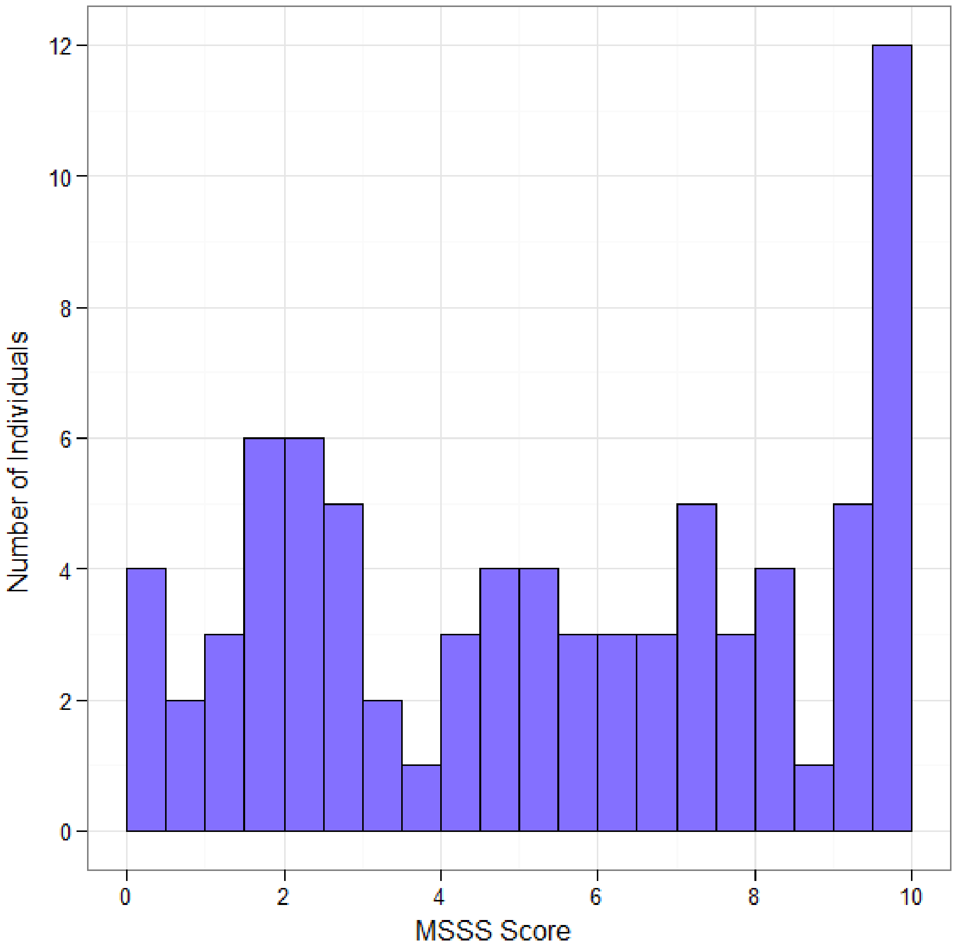
MSSS

The prevailing measure for disease progression is the EDSS. However, this scale fails to take into account the length of disease course, which makes it difficult to compare disease severity between multiple individuals. For example, a person who requires walking assistance such as a cane within five years of diagnosis compared to a person who has relatively little disability and needs no walking assistance for forty years after diagnosis suffer from very different disease courses. To aid in comparison between individuals, the MSSS was created. (32) This scale uses the EDSS scores and length of disease in a group of individuals to normalize the scores. The MSSS goes from 0 to 10, the same as the EDSS, but instead of representing the extent of disability, a score represents how many individuals have the same extent of disability as the person being tested. For example, an MSSS score of 5 indicates that 50% of people who have had MS as long as the individual of interest have lesser disabilities and an MSSS score of 3 indicates that 30% of people who have had MS as long as the individual of interest have lesser disabilities.

We used MSSStest (32) to calculate MSSS for 79 individuals with an EDSS score and year of diagnosis, which was used to calculate length of disease in months. The global MSSS scale included with the program was used to calculate MSSS for our samples. We used this scale instead of creating one from our samples because the global scale dataset (9,981 individuals) was much larger than our dataset and provided a better sampling of EDSS distributions. For patients with more than one EDSS score recorded, the most recent score was used. The distribution of these scores for the 79 individuals is shown in **Figure 4.5**.

Based on these individuals, 47,441 SNPs with MAF < 0.05 were removed prior to analysis. No individuals were genetically related. Linear regression for MSSS was performed for the remaining 109,033 SNPs using the first three principle components as covariates. The most significant results are listed in **Table 4.6**. The two SNPs with the lowest p-values, rs6718188 and rs12613548 (r^2 = 0.78), are 170kb upstream and 7kb downstream, respectively, from *SP3*. *SP3* is a transcription factor and affects the transcription of numerous genes throughout the genome. Another candidate gene based

Figure 4.5 Distribution of MSSS



on these results is *MAPKAPK2*, a kinase involved in inflammatory responses on chromosome 1. rs10863784 is in an intron of this gene. Lastly, rs2401399 is located 40kb downstream of a SNP in *CLEC1*, which is a known loci associated with MS. The LD between the *CLEC1* SNP (rs11052877) and rs2401399 is not strong in this dataset (r^2 = 0.02; D' = 0.67). Also, this SNP is out of HWE in the controls of the larger dataset (p = 5.07 x 10^{-11}), but no deviation from HWE is seen in these 79 samples (p = 0.31).

Table 4.6 Most significant results for MSSS

CHR	ImmunoChip ID	rsID	BP	MA	MAF	BETA	P
2	rs6718188	rs6718188	174469857	A	0.33	-2.47	2.24E-06
2	rs12613548	rs12613548	174544015	A	0.32	-2.22	2.11E-05
6	rs9501747	rs9501747	1590552	A	0.08	-3.53	4.36E-05
12	imm_12_9831209	rs2401399	9831209	A	0.07	-4.04	7.32E-05
1	imm_1_204938407	rs10863784	204938407	G	0.09	3.40	8.15E-05

CHR: chromosome; BP: base pairs; MA: minor allele; MAF: minor allele frequency; P: two-sided p-value

Timed 25 foot walk

The timed 25 foot walk is another measure used to monitor disease progression of MS. Of the 3,523 individuals with timed walks, 2,422 had at least two walks recorded (total 16,124 walks). As we wanted to evaluate overall disease progression, we elected to remove timed walk scores that appeared to be extended due to relapses—scores were removed if they were greater than five seconds longer than timed walks recorded before and after the walk in question. After removing 326 walks as possible relapses, all individuals had at least two walks remaining. We next identified individuals whose timed walks covered a period of at least one year to ensure we had a better view of overall disease progression; 321 individuals were excluded.

519 of the remaining 2,101 individuals had genotype data and passed QC. The slope of the time walk scores in seconds per year was calculated based on the first and last recorded timed walks for each individual. The average slope was an increase of 0.67 seconds per year, with a standard deviation of 5.40 seconds. Seven individuals that fell outside of three standard deviations were excluded.

Timed walks for a subset of nine individuals, representative of all individuals, are shown in **Figure 4.6**. The length of time to walk 25 feet did not increase for all individuals; for many it stayed relatively constant, while for others it actually improved.

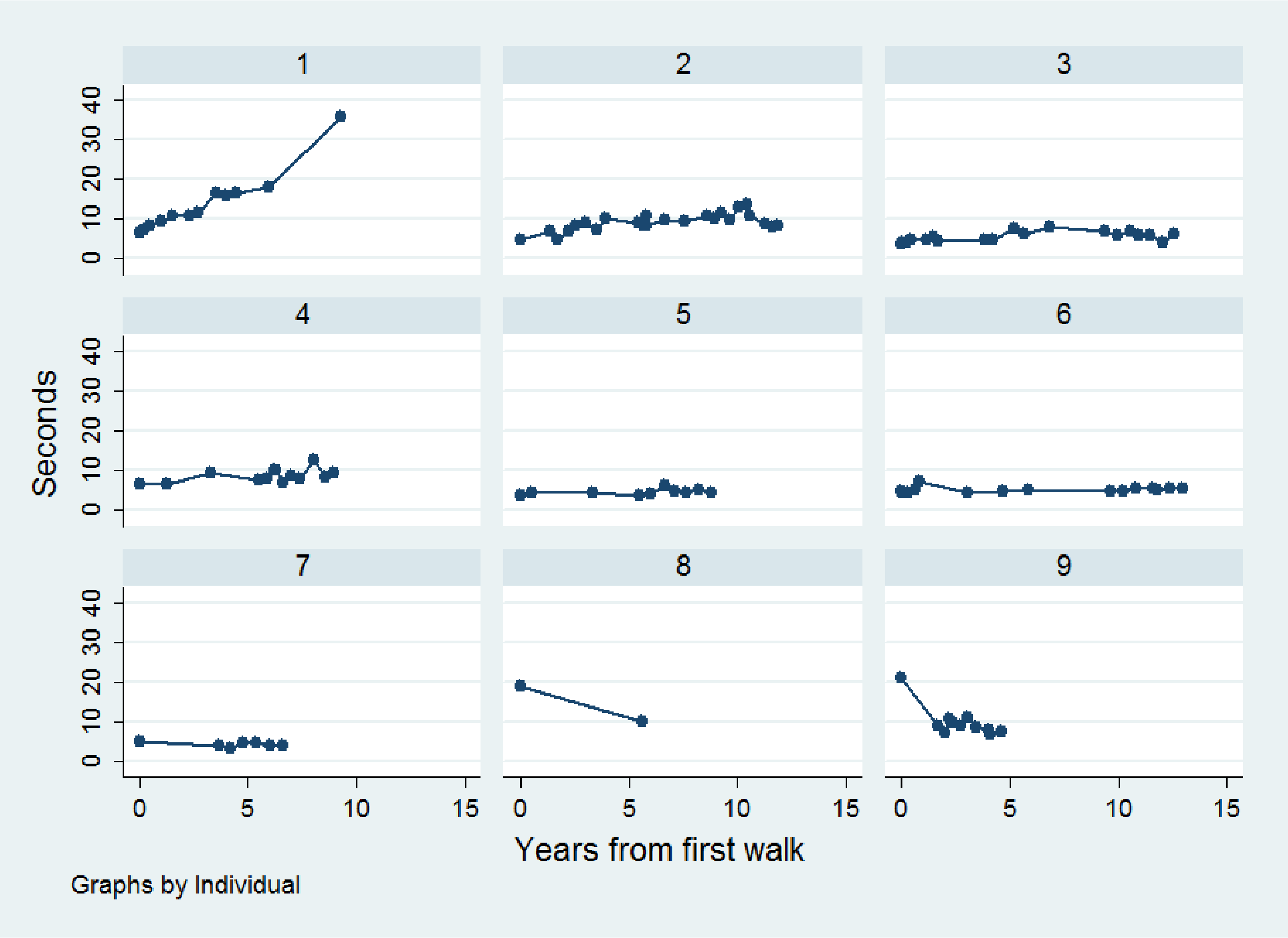
49,729 SNPs with MAF < 0.05 were removed prior to analysis. Linear regression was run on 512 individuals and 106,745 SNPs using slope as the phenotype of interest and the first three principle components as covariates. The most significant SNPs are listed in **Table 4.7**.

Table 4.7 Most significant results for the timed 25 foot walk analysis

CHR	ImmunoChip ID	rsID	BP	MA	MAF	BETA	P
10	rs11201609	rs11201609	87157513	A	0.123	0.7303	2.76E-05
10	rs7068623	rs7068623	87162243	T	0.1309	0.7008	4.08E-05
10	imm_10_49682645	rs11101506	49682645	A	0.05078	0.991	8.08E-05
10	imm_10_49683537	rs11101508	49683537	G	0.05078	0.991	8.08E-05

CHR: chromosome; BP: base pairs; MA: minor allele; MAF: minor allele frequency; P: p-value

Figure 4.6 Timed 25 foot walks for 9 individuals



These SNPs cluster in two areas on chromosome 10. The first is in chromosome 10q23.1, 200kb upstream from *GRID1*. *GRID1* is a glutamate receptor and involved in synaptic plasticity in the CNS. This gene has been associated with anorexia nervosa(118), stearic acid plasma levels(119), pancreatic cancer(120), periodontal microbiota(121), and left ventricular cardiac wall thickness.(122) These phenotypes are not directly related to MS or the timed 25 foot walk, but they do indicate that *GRID1* may affect a variety of phenotypes. The third and fourth SNPs in **Table 4.7** are located at chromosome 10q11.22 in introns in the *WDFY4* gene. No functional information is available for this gene.

CNS origin of first neurological symptom

The CNS origin of first neurological symptom was extracted by algorithm from referral letters written by the clinician and categorized as brain stem, optic nerve, and spinal cord in origin. Individuals whose symptoms fell in multiple categories were designated polysymptomatic origins. For this analysis, we did not include individuals with polysymptomatic origins due to the low frequency of individuals in each possible combination of origins. From the individuals we were able to extract origins of first symptoms, 180 were genotyped, passed QC, and had a single origin of first neurological symptom. The counts of individuals with symptoms of each origin are listed in **Table 4.8**. No pairs of individuals were related according to pairwise IBS analysis. 50,003 SNPs were excluded based on MAF < 0.05 in this dataset.

Polytomous regression was run on 180 individuals and 106,471 remaining SNPs in R using the 'polytomous' package with the first three principal components as covariates. The dataset was broken into 1,000 SNP sets and run in parallel for analysis to decrease computational time. A template script for this analysis is shown in **Appendix C**. The analyses were run using an heuristic model of one versus the rest—individuals with brain stem origin compared to those with optic nerve and spinal cord origins, individuals with optic nerve origin compared to those with brain stem and spinal cord origins, and individuals with spinal cord origins compared to those with optic nerve and brain stem origins. SNPs with the most significant results are shown in **Table 4.9**.

Table 4.8 CNS origin of individuals used for analysis

Origin	**Individuals,** ***n***
Brain stem	50
Optic nerve	30
Spinal cord	100

Table 4.9 Most significant results for polytomous regression of origin of first neurological symptom

CHR	**ImmunoChip ID**	**rsID**	**BP**	**MA**	**MAF**	**BS (P)**	**ON (P)**	**SC (P)**	**Odds**
9	rs4876968	rs4876968	90916697	A	0.11	8.09E-05	1.10E-01	1.37E-02	1.08
5	rs1434660	rs1434660	136512738	A	0.50	2.30E-04	3.97E-01	5.94E-03	0.61
8	rs1446534	rs1446534	118026220	A	0.24	3.72E-04	6.87E-01	3.24E-04	0.94
12	rs7974348	rs7974348	120739096	G	0.19	2.23E-01	5.59E-05	2.40E-02	1.06
12	rs1168671	rs1168671	120714231	G	0.20	4.72E-01	8.55E-05	9.74E-03	1.10
6	1kg_6_30143320	rs9261281	30143320	A	0.09	6.67E-01	9.56E-05	8.54E-03	2.06
1	1kg_1_67594675	rs10749775	67594675	C	0.16	1.21E-02	3.81E-02	1.76E-04	1.41
9	rs1105191	rs1105191	15360575	G	0.43	1.68E-02	2.28E-02	1.78E-04	0.95
5	1kg_5_159793919	rs17057795	159793919	G	0.42	5.40E-03	1.24E-01	2.60E-04	1.77

CHR: chromosome; BP: base pairs; MA: minor allele; MAF: minor allele frequency; BS: brain stem; ON: optic nerve; SC: spinal cord; P: p-value; Odds: odds value corresponding to the p-value highlighted in blue

Several of the most significant results are of particular interest. rs4876968, associated with symptoms of brain stem origins, is located in an intron of *SHC3*, a signaling adapter differentially expressed in the brain and spinal cord. The expression pattern of this gene makes it a likely candidate for the CNS origin of neurological symptoms. rs9261281, associated with symptoms of optic nerve origins, is located in an active promoter region of *PPP1R11*. This gene is located in the MHC class 1 region and is an inhibitor of protein phosphatase. rs10749775, associated with symptoms of spinal cord origin, is intronic to *IL12RB2*, which is involved in Th1 cell differentiation. *IL12RB2* has been associated in other studies with Behçet's disease, which causes inflammation in blood vessels throughout the body(123;124), and primary biliary cirrhosis.(125;126)

Time to SPMS

Individuals initially diagnosed with RRMS may transition to SPMS at any point in their disease course. We analyzed the time in years from diagnosis of RRMS to SPMS for genetic associations using Cox proportional hazards regression.

We identified clinical subtypes of MS based on the clinical text for each individual as described in Chapter II. We also identified year of diagnosis by algorithm. A major point of difficulty was determining when a patient actually transitioned to SPMS. There are no major distinctions between RRMS and SPMS and the change in diagnosis is based upon observations of the clinician. Discussion and monitoring for a possible transition to SPMS can occur over a period of years in the medical record. We focused on the clinical impressions given by the clinician to extract the subtype of MS from the record and excluded all references to subtypes that were not definite (see Chapter II). For each subtype, we extracted the date of the clinic note in which it was recorded. We also needed to confirm that the first instance of an SPMS diagnosis we identified was in fact the actual time of transition to SPMS. Some individuals became patients at VUMC mid-point in their disease course, so the first recording of SPMS may not indicate the first time the diagnosis of SPMS was applicable. To confirm the time of transition to SPMS, we restricted our analysis to include individuals for whom we had identified a previous recording of RRMS. This allowed us to hone in on the actual transition time between the two subtypes of MS. The year of diagnosis and the date of the clinic notes with diagnoses of SPMS provided enough information to calculate the length of time of disease course to SPMS. However, the requirements for each individual to have a diagnosis of RRMS, a diagnosis of SPMS, and a year of diagnosis that we were able to extract limited the individuals for analysis quite stringently.

Individuals with RRMS who had not converted to SPMS still provided valuable information. Those with year of diagnosis were included in the analysis. Cox proportional hazard regression utilized all of these data in a time to event analysis.

228 individuals had the required information, were genotyped, and passed QC. 125 individuals had converted SPMS, 103 had not yet converted to SPMS. 48,371 SNPs with MAF < 0.05 were excluded. The remaining 108,103 SNPs were broken into 1,000 SNP datasets and run in parallel to decrease computational time. Regression was run in R using the 'survival' package with the first three principal components as covariates. An example script for this analysis is shown in **Appendix D**. The most significant SNPs are listed in **Table 4.10**.

Three SNPs, rs776176, rs776175, and rs4761251, are located in introns of *BEST1*, a transmembrane protein. rs10885868 on chromosome 10q25.3 is in an intron of *GFRA1*, which is a receptor for *GDNF* and is involved in neuron survival and differentiation.(127;128) rs6952706 is located 17kb upstream from *IRF5*, a transcription factor that has been associated with other autoimmune diseases, including lupus erythematosus(129;130), inflammatory bowel disease(125;131), systemic sclerosis(132;133), and rheumatoid arthritis.(134) 97 SNPs on the ImmunoChip were located in *IRF5*. After SNP pairwise pruning for r^2, LD was calculated for 9 SNPs in this gene with rs6952706. r^2 values did not indicate a high level of LD between this SNP and *IRF5* (0.00 < r^2 0.10). Finally, rs45509092 on chromosome 16p11.2 is in a small intron of *ITGAX*, a leukocyte-specific integrin. *ITGAX* has also been associated with lupus erythematosus.(135)

Table 4.10 Most significant results for time to SPMS analysis

CHR	ImmunoChip ID	rsID	BP	MA	MAF	Coeff	P
12	rs776176	rs776176	68328241	G	0.33	0.58	1.00E-05
12	rs776175	rs776175	68328706	A	0.34	0.58	1.20E-05
10	rs10885868	rs10885868	1.18E+08	A	0.37	-0.63	1.60E-05
9	rs7032677	rs7032677	76437417	G	0.49	-0.59	3.60E-05
12	rs249167	rs249167	93854404	T	0.25	0.63	3.80E-05
12	rs4761251	rs4761251	68328535	G	0.46	0.57	4.50E-05
1	rs10915314	rs10915314	4976182	C	0.27	0.56	4.80E-05
7	imm_7_128348691	rs6957206	1.28E+08	A	0.10	0.78	7.10E-05
16	imm_16_31278669	rs45509092	31278669	A	0.05	1.03	7.20E-05
1	rs12141935	rs12141935	1.69E+08	C	0.26	0.56	9.00E-05
6	rs2067002	rs2067002	52479378	G	0.21	0.58	9.30E-05

CHR: chromosome; BP: base pairs; MA: minor allele; MAF: minor allele frequency; Coeff: coefficient; P: p-value

Discussion

We replicated 45 SNPs of the known MS loci at a p-value of 0.05, in addition to the effect of the HLA locus. Based on power calculations, we would have expected to see 50.5 SNPs at this level; while our number is slightly lower, it is still in the range of what we would have expected. Additionally, very few of the SNPs (16/105) showed an effect in the opposite direction of that in the published literature and all of these were under the significance level of 0.05. The overwhelming trend in the right direction for these SNPs lends further weight to the argument that this dataset of MS individuals has an underlying genetic architecture similar to that of published MS datasets. The genetic risk scores of the cases were significantly higher than the controls, which further supports the comparability of our dataset to other published datasets. Genetic risk scores included all SNPs, regardless of significance in our dataset. This difference was preserved when the representative HLA SNP was removed.

Analyses were run among the cases for seven clinical traits. While extraction of traits from the SD in Chapter II occurred in a high proportion of individuals, the sample sizes for the genetic analyses were much lower. One reason is because only a portion of the samples in the SD were available in BioVU; DNA collection started in 2007 and is dependent upon routine blood draws. Additional cases could be extracted and used in the future, dependent upon DNA collection. We genotyped all cases with DNA available at the time of the study. Another reason for smaller sample sizes is because several of the clinical traits for genetic analysis, including MSSS and time to SPMS, required multiple traits extracted in Chapter II so we were only able to analyze individuals with both traits. MSSS requires both an EDSS score and age at diagnosis to calculate. Time to SPMS requires year at diagnosis and knowledge of the year of transition to SPMS. The dataset sizes could be increased by conducting a targeted manual review. If a manual review focused on genotyped individuals with one of the two required traits (such as presence of an EDSS score but not of year of diagnosis), it would not be as time intensive as a broad manual review of records. Any targeting should take into account the recall of the algorithms. For instance, the recall of the algorithm to extract subtypes of MS is 98%; manual review to detect additional subtypes is unlikely to be highly productive.

Genetic analyses were conducted for age at diagnosis, age at first neurological symptom, presence of oligoclonal bands, MSSS, timed 25 foot walks, CNS origin of first neurological symptom, and time to SPMS. While no results reached genome-wide significance or a Bonferroni corrected p-value of 1×10^{-6} (based on 100,000 SNPs), several SNPs were significant at 1×10^{-4}. Multiple SNPs are in regions of interest, with candidate genes. The most significant result in any analysis was on chromosome 2 for MSSS ($p = 2.24 \times 10^{-6}$). This SNP is located near the *SP3* genes, which is a transcription factor and involved in many pathways throughout the body. The most interesting result biologically was a SNP on chromosome 9 associated with origination of first neurological symptoms in the brain stem when compared with the spinal cord and optic nerve. This SNP is located in an intron of *SHC3*, which is expressed in different patterns throughout the CNS. Future studies delineating the expression of *SHC3* in various parts of the CNS could yield greater insight into the possible role of this gene in the first neurological symptom displayed by MS patients.

Additional studies with larger datasets would help to further define the significance of these results, especially if the genes identified could be targets of functional studies to determine their impact on MS clinical disease course. The addition of DNA samples in BioVU would greatly increase the power of genetic studies of clinical traits of disease, specifically studies that require the multiple clinical traits for a single analysis.

We saw very few results for clinical traits in or near loci associated with MS risk. This was somewhat surprising as it is a plausible hypothesis that loci which show small effects on risk may actually play a larger role in the overall disease course. We did not find very much evidence to support this hypothesis. However, many of the most significant SNPs were located in or around genes with functions that could be biologically implicated in the clinical trait analyzed.

CHAPTER V

CONCLUSIONS

Multiple sclerosis is a disease with well-established genetic risk loci. The HLA locus plays the largest role in the genetics of MS, accounting for 10-50% of the heritability of MS.(14) In addition to the HLA locus, 110 SNPs in 103 loci have been confirmed as MS-risk loci.(61) While these SNPs have small effects individually, replication has shown that the effects do contribute to the overall risk of disease. The field has progressed rapidly in the last six years, burgeoning with new results. Collaborations, including the IMSGC, have resulted in large-scale studies that have detected loci that would have otherwise gone undetected. Studies of the genetics of MS continue to search for the remaining portion of the genetic effect.

The clinical expression of MS is extremely heterogeneous; age of onset, types of symptoms, disease courses, rates of progression, and disability levels may vary drastically between individuals. Some studies have shown familial aggregation of clinical traits of MS, but little is understood of the genetics of clinical expression. Part of this lack of knowledge is due to the small number of studies that have been performed. The genetics of clinical expression has not been fully explored.

Medical records provide a source of rich phenotypic information. EMRs provide opportunities to mine the information in medical records in an automated manner; this provides an opportunity to utilize data in medical records without the overwhelming amount of time it would take to review medical records manually.

We used four algorithms based on ICD-9 codes, medications, and text keywords to identify 5,789 patients with MS. Three of these algorithms performed with high specificity; one algorithm ("possible type 1") mostly captured patients with the possibility of having MS, but not confirmed diagnoses. We recommend this algorithm be used in research studies only when confirmation of disease status by manual review can be obtained. We also created an algorithm to identify patients to be used as controls for our MS cases. In addition to exclusion of MS, we required patients to have no other autoimmune diseases. This algorithm found thousands of individuals, providing a large pool from which to pull subjects. We narrowed down this dataset by matching to our cases on age, sex, and BMI. All algorithms, except the "possible type 1" algorithm, performed very well in selecting individuals suitable for the subsequent studies we conducted.

A manual review of a subset of the records of 60 cases confirmed the depth of MS disease course data available. Age of onset, EDSS, family history, symptoms, and treatment data was collected from the narrative text. Data was available in many records for age of onset, symptoms, and treatment. EDSS was present in a smaller than expected number of individuals. After discussion with clinicians with the MS Clinic, we found this was due to the difficulty in scoring the EDSS in a clinic setting and a greater reliance upon the timed 25 foot walk as a progression measure. Family history, or lack thereof, did not appear to be reliably recorded.

After establishing the existence of clinical disease traits in the medical record, we created algorithms using MySQL and Perl to extract these data in a time efficient manner. Algorithms were created for year and type of first neurological symptom, year of diagnosis, presence of oligoclonal bands, EDSS scores, timed 25 foot walk scores, clinical subtype of MS, and medications. Manual review of 100 records in the test dataset was performed to create a gold standard to calculate statistics evaluating the performance of the algorithms. The algorithms performed very well, with high specificity.

The recall values for year of diagnosis and year and type of first neurological symptom were low. The data extracted was precise, but more than half of data present in the medical records for these traits was missed by the algorithms.

The presence of oligoclonal bands and the timed 25 foot walks provided additional opportunities to evaluate the accuracy of the data described in the narrative text of the clinical records. Oligoclonal bands are a laboratory value. For patients who had the test conducted at VUMC, we were able to compare the results extracted from the narrative text to the results in the structured fields for the laboratory test. The concordance was 97.1%, indicating that the clinician reported value in the text was concurrent with the actual laboratory result; few errors occurred in transcription of the result.

A structured field for the timed 25 foot walk was created for the MS Clinic in 2008, providing another opportunity to check the accuracy of a value transcribed into the narrative text. Of the dates that overlapped between the structured field walks and those extracted from the text, the concordance was 91%. Many of the other scores appeared to be typographical errors. The creation of the structured field allowed us to capture a glimpse of how many timed walks were omitted from the narrative text previous to that date; these scores were recorded in a paper chart and are not available to us. 70% of the timed walks in the structured field were not recorded in the clinic note, suggesting we are missing a vast number of timed walks prior to structured field creation. It is possible that the clinicians began to record the walk time in the clinic note less frequently after creation of the structured field because of the ease in accessing the timed walk. If this was the case, 70% would be an over-inflated estimate of missing scores prior to 2008. We compared the distributions of scores extracted by algorithm from the text and from the structured field. There was no significant difference; if these distributions were different we would be concerned of a bias in which scores were recorded in the text.

In summary, we identified the existence of detailed clinical course information for MS in a de-identified EMR. Several traits were extractable by algorithm with high specificity and precision. While some traits had low recall rates, the data extracted by all algorithms is accurate based on the medical records and could be used for research purposes.

We extracted DNA for all cases identified by algorithm who were part of BioVU. We also extracted DNA for controls for genotyping. All samples were genotyped on the ImmunoChip, a custom genotyping array focused on autoimmune disease loci. The calling of genotypes was difficult because of the presence of rare and common variants on the chip. After trial and error, the OptiCall program was used. Samples were called as part of the IMSGC ImmunoChip data; the larger number of samples aided in more accurate calling of the rare variants. Extensive sample and SNP quality control was performed. After QC, 156,474 SNPs and 3,187 samples (1,003 cases, 2,184 controls) were available for analysis. In this dataset, there are several pairs of cases related based on IBS (six sibling pairs, five parent-offspring pairs, and one cousin pair). As the records are de-identified, we cannot confirm the family structure of these individuals. In any analyses performed, one of each pair of related individuals was removed prior to analysis.

Logistic regression of the known MS loci was performed in cases and controls. 46 of 106 SNPs that passed QC were replicated at $p < 0.05$. 44 of the remaining 60 SNPs trended in the right direction even though they did not reach significant. Genetic risk scores based on these 106 SNPs were significantly higher in cases than controls (p = 27 x 10^{-32}). These results are as we would expect from a standard MS population, confirming the similarity of the genetics of MS with the BioVU population.

Using data collected by algorithms from the SD, we calculated seven phenotypes for genetic analysis: age at diagnosis, age at first neurological symptom, presence on oligoclonal bands, MSSS, timed 25 foot walk, CNS origin of first neurological symptom, and time to SPMS. Linear, logistic, polytomous, and Cox proportional hazard regression analyses, as appropriate, were performed using principle components as covariates to control for population structure. No genome-wide significant results were observed. Two SNPs, for age at first neurological symptom and MSSS, approached a Bonferroni corrected p-value of 1 x 10^{-6}, based on analyses of 100,000 SNPs. rs9935467 was associated with increasing age at first neurological symptom at p = 4.12 x 10^{-6}. It is located on chromosome 16. rs6718188, on chromosome 2, was associated with decreasing MSSS. This SNP is located 170kb upstream of *SP3*, a transcription factor.

An interesting candidate gene for rs4876968, a SNP associated with symptoms originating in the brain stem, is *SHC3*. *SHC3* is a signaling adapter differentially expressed in the brain and spinal cord. Additional studies are needed, but a varied expression pattern of a gene would be plausible given the differences observed in the origin of the symptoms.

A major limiting factor for these analyses was the sample sizes. Five of the seven analyses required data from the three algorithms with low recall (year of diagnosis, year and origin of first neurological symptom). Improvement of recall for these algorithms could greatly increase the sample sizes for these analyses.

This collection of clinical data represents one of the largest databases of detailed, clinical traits available for research of MS. This work demonstrates that detailed clinical information is recorded in the EMR and can be extracted for research purposes with high reliability. The analyses of clinical traits have brought further information to a field with minimal information available. While small in sample sizes, these analyses have provided a groundwork for future studies to expand upon.

Future directions

There are several possible future directions for this project. For the work in the EMR, these include expanding the current algorithms for clinical traits, carrying the algorithms to other institutions, creating new algorithms, and applying the methods for algorithm development to other diseases. For the genetic work, analysis of additional MS samples as they are added to BioVU and collaboration with other research groups to increase sample size and power would aid these analyses.

Several of the algorithms for clinical traits presented in this work could be expanded to increase the sensitivity of explicitly stated values and to calculate values based on data in the text. The data extracted was reliable, but more than half of data present in the medical records for certain traits was missed by the algorithms. The medications script had low recall, largely because we restricted searching to PL. Broadening the algorithm to all clinic notes would likely increase sensitivity. The difficulty is doing this without sacrificing specificity, as seen when we used MedEx. Initial efforts could focus on clinic notes where there is a section for specific medications. If consistent formatting and language could be found, this would add in identifying medications currently taken. An ultimate goal would be to identify not just if a patient was ever on a drug, but to identify the start and stop points as well as the dosage. This information would be useful in and of itself, and it could also be combined with other information, such as relapses and progression, to determine drug failure.

The script for age of diagnosis had low sensitivity for current diagnoses of MS, as discussed in Chapter II. Enhancing this script to identify clinic visits in which a diagnosis of MS is given would greatly increase sensitivity. This will likely require greater use of NLP to determine the confidence of the clinician in the clinical impressions of the visit, using grammar rules to distinguish a diagnosis of MS, a possible diagnosis of MS, and confidence that a diagnosis of MS is incorrect for a patient.

Currently, only explicit statements of EDSS scores are captured by the algorithm. However, it is possible that some scores may be calculated by the data in clinic visits to increase the number of EDSS scores for our samples. The easiest to calculate would be scores six through nine, which are based on the presence of walking aids or restriction to bed. Scores one through four would be difficult but could potentially be inferred from the physical examination. Scores 4 through 5.5, which require an understanding of how far a patient can walk, up to 500 meters, would be the most difficult and we would likely still experience an uneven distribution of scores.

The application of the subject selection and clinical trait algorithms proved to be great tools in creation of a large dataset of MS individuals with longitudinal disease course data at VUMC. Further use of these algorithms would be to apply them to EMR datasets in other institutions. The subject selection algorithms should be easily transferable as there are no parts of the algorithm that are specific to VUMC records. The transferability of the clinical trait algorithms is likely to vary. We expect the most difficult algorithms to transfer would the age and type of first neurological symptom. These algorithms rely on clinician specific wording to identify referral letters, with a history with specific key words in these letters. The general principle could be carried over but evaluation of the clinic notes should be done to evaluate the format of the notes at the intended university or clinic. Presence of oligoclonal bands and timed 25 foot walk algorithms rely on no institution-specific formats. Ascertainment of structured fields at any institution should first be attempted; however, the ease with which we were able to identify these scores suggests NLP-derived algorithms would work well at other institutions if needed. Additional methods of recording the results in the text could be added if deemed necessary. For instance, abbreviations for the time walk, including “ft” and T25FW, were not seen in the records we reviewed but they may be used at other institutions. We know of no specific reasons why the algorithms for age at diagnosis, EDSS, and clinical subtype would not be transferable. The algorithm for medications would depend upon the existence of PL at the institution of interest.

We have shown in this work the feasibility to create algorithms to extract detailed clinical traits. Algorithms to extract additional clinical traits could be considered. Particularly, the identification of relapses in patients with RRMS would be desirable, along with relapse length, frequency, symptoms, and inter-relapse time. Possible approaches are identifying key phrases from the narrative text and communications. Patients do not always come in to the clinic when they are having a relapse, but they are encouraged to at least call in. When calling in, it is recorded in the communications and treatment may be offered by the clinician. The administration of steroids, such as intravenous methylprednisone (IVMP), could be used to indicate relapses. IVMP is generally not a routine medication for patients with RRMS, but is given to rapidly decrease inflammation in a patient with acute symptoms during relapses. Also, relapses could be inferred from a sudden increase in the timed 25 foot walk that is resolved by the next visit. This would require visits surrounding the timed walk in question and may miss patients without prior timed walks and those who are lost to follow-up. Identification of relapses is likely not possible for patients who are not treated for MS at VUMC. Those treated at other institutions may have mention of relapses in clinic visits for other reasons, but a full history of relapses is unlikely to be recorded.

One of the most interesting results from the genetic analyses of clinical traits was for origin of first symptom. A SNP in an intron of *SHC3* was associated with symptoms originating in the brain stem as opposed to the spinal cord and optic nerve. This gene is highly expressed in the brain. One expression study has shown that while it is expressed in various neuronal cell populations, the subcellular location of the protein varies drastically by cell type.(136) Another expression analysis shows that *SHC3* is highly expressed in the frontal cortex, and at lower levels in the brain stem and spinal cord.(137;138) Additional studies to better the expression profile of this gene in various areas of the CNS could outline greater understanding of a possible role it may play in onset of neurological symptoms in MS.

We have honed several techniques to extract detailed disease course data from EMRs. These techniques could be applied to other diseases. MS was an excellent disease to work with because of the high precision in identifying individuals with the disease, several specific clinical traits that are explicit to MS and routinely recorded, frequent follow-up of the disease in clinics is observed, and there is a large number of patients with MS at VUMC for us to focus on due to the MS Clinic. These types of observations should be taken into account when selecting other diseases for which to apply these techniques.

APPENDIX A

CONTROL SELECTION ALGORITHM

Codes: Record does not contain any of the following:

070.2* Viral hepatitis B with hepatic coma

070.3* Viral hepatitis without mention of hepatic coma

070.51 Acute hepatitis C without mention of hepatic comal

070.54 Chronic hepatitis C without hepatic coma

070.7* Unspecified viral hepatitis C

150 Malignant neoplasm of the esophagus

250.* Diabetes Mellitus

255.4 Corticoadrenal insufficiency (Addison's disease)

286.0 Congenital factor VIII disorder

286.1 Congenital factor IX disorder

286.2 Congenital factor XI deficiency

714 Rheumatoid arthritis and other inflammatory polyarthropathies

714.0 Rheumatoid arthritis

714.1 Felty's syndrome

714.2 Other rheumatoid arthritis with visceral or systemic involvement

714.30 Polyarticular juvenile rheumatoid arthritis, chronic or unspecified

714.31 Polyarticular juvenile rheumatoid arthritis, acute

714.32 Pauciarticular juvenile rheumatoid arthritis

714.33 Monoarticular juvenile rheumatoid arthritis

714.9 Unspecified inflammatory polyarthropathy

373.34 Discoid lupus erythematosus of eyelid

695.4 Lupus erythematosus

710.0 Systemic lupus erythematosus

710.1 Systemic sclerosis

710.3 Dermatomyositis

710.4 Polymyositis

710.2 Sjogren's disease

493.* Asthma

530.85 Barrett’s esophagus

555.* Regional enteritis

579.0 Celiac disease

556.* Ulcerative colitis

571.6 Biliary cirrhosis

571.8 Other chronic nonalcoholic liver disease – fatty liver, without mention of alcohol

573.1 Hepatitis in viral diseases classified elsewhere

573.2 Hapatitis in other infectious diseases classified elsewhere

573.3 Hepatitis, unspecified

576.1 Cholangitis

135 Sarcoidosis

696 Psoriasis and similar disorders

696.0 Psoriatic arthropathy

696.1 Other psoriasis and similar disorders excluding psoriatic arthropathy

696.8 Other psoriasis and similar disorders

099.3 Reiter’s disease

719.3 Palindromic rheumatism

720.0 Ankylosing spondylitis

720.8 Other inflammatory spondylopathies

720.81 Inflammatory spondylopathies in diseases classified elsewhere

720.89 Other inflammatory spondylopathies

720.9 Unspecified inflammatory spondylopathy

721.2 Thoracic spondylosis without myelopathy

721.3 Lumbosacral spondylosis without myelopathy

245.2 Hashimoto's thyroiditis

242.0 Toxic diffuse goiter

358.0 myasthenia gravis

358.00 myasthenia gravis without acute exacerbation

358.01 myasthenia gravis with acute exacerbation

775.2 neonatal myasthenia gravis

443.0 Raynaud's syndrome

KEYWORDS: Record does not contain any of the following:

Hepatitis B [hep B]

Hepatitis C [hep C]

Addison*

Hemophilia

rheumatoid [rheum] [reumatoid] [rhumatoid] arthritis [arthritides] [arthriris] [arthristis] [arthritus] [arthrtis] [artritis]

Felty*

juvenile [juv] rheumatoid [rheum] [reumatoid] [rhumatoid] arthritis [arthritides] [arthriris] [arthristis] [arthritus] [arthrtis] [artritis]

juvenile [juv] arthritis [arthritides] [arthriris] [arthristis] [arthritus] [arthrtis] [artritis]

juvenile [juv] RA

juvenile chronic arthritis [arthritides] [arthriris] [arthristis] [arthritus] [arthrtis] [artritis]

Rheumatism

Lupus [SLE] [lupos] [lupis] [lupas]

Crohn* [crones] [crons]

Ulcerative colitis [colitis] [colitus] [UC]

Inflammatory [inflamatory] [inflam] bowel disease [dx] [IBD]

Sarcoidosis [sarcoid] [sarcodosis] [sarcadosis]

psoriasis [soriasis] [psorisis] [psorasis] [psiorasis]

Reiter*

Regional enteritis

Sjogren* [shogren*] [showgrens] [sjorgens] [sjogens]

Asthma

Barrett’s esophagus

Esophageal cancer

Celiac sprue

Biliary cirrhosis [PBC] [primary sclerosing cholangitis]

Non-alcoholic fatty liver [NAFLD] [non alcoholic fatty liver][nonalcoholic fatty liver]

Non-alcoholic steatohepatitis [NASH} [non alcoholic steatohepatitis] [nonalcoholic steatohepatitis]

Dermatomyositis [dermatomyocitis]

Polymyositis [polymyocitis]

Ankylosing [ankleosing] [rheumatoid] [rheum] [reumatoid] spondylitis [spondolitis] [spondylities] [spondilitis]

Thoracic spondylosis

Lumbosacral spondylosis

Hashimoto thyroiditis

Chronic lymphocyctic thyroiditis

Autoimmune thyroid disease [AITD]

Grave* disease [dis] [dx] [dz] [syndrome]

Myasthenia gravis

Raynaud* disease [dis] dx] [dx] [syndrome]

Degenerative joint disease [dis] [dx] [dz]

APPENDIX B

PSEUDOCODE FOR CLINICAL TRAIT ALGORITHMS

We have included the regular expressions used in Perl format.

```
Text  =~ /{regular expression}/{search functions}
```

Subtype

1. Select clinic notes, problem list, letters, and research notes

2. Select clinic notes that contain mention of a subtype of MS

```
/ (?:(?:RR|SP|PP)[-\s]?(?:MS|multiple\ssclerosis))|
  (?:(?:primary|secondary)[-\s]+progressive)|
  (?:relapsing[-\s]remitting)|
  (?:(?:progressive|relapsing)[,-\s\/](?:progressive|relapsing))|
  (?:relapsing\s(?:multiple\ssclerosis|ms\s))/
```

3. Extract 50 characters before and after the subtype mentioned

4. Identify subtype mentioned

5. Identify the subtype as a negative value if the phrase before the subtype matched contains “not” in the phrase

6. Identify the subtype as a negative value if any of the phrase contains other negative phrases

```
/possibl|rule\sout| r\\o/
```

7. Delete subtypes marked as negative values

8. Remove multiple occurrences of subtypes per person

Oligoclonal bands

1. Select clinic notes and letters

2. Select notes that reference oligoclonal bands

```
/oligo/i
```

3. Split notes into sentences

```
/(?:\.\s+)|(?:-(?:\s+)?){2,5}|(?:>\s?<)|\\n|<.?B>/
```

4. For each line, pull out 100 characters on either side of ‘oligoclonal’

```
/.{0,100}oligoclonal.{0,100}/i
```

5. Search the phrase to find if a positive result was indicated and save it with the ID and note date

```
/(?:positive|\bpresent\b)(?:\s)?(?:\b\w{1,20}(?:\b\s)?){0,3}oligoclonal|oligoclonal.{0,15}(?:positive)|in(?:\s)?CSF(?:\s)?only/ix
```

```
/[1-9]\s?oligo|(?:two|three|four|five|six|seven|eight|nine|ten)\s?oligo|\+(?:\s)?oligo/ix
```

6. Search the phrase to find if a negative result was indicated and save it with the ID and note date

```
/no(?:\s)?oligo|not(?:\s)?show.{0,20}oligo|negative.{0,20}oligo|band.{0,20}negative|neg\s?oligo|band(?:s?\s?)neg/ix
```

7. Remove duplicate results (ID and result-positive/negative)

8. Mark individuals with both a positive and negative result as inconclusive

Diagnosis Year

1. Search all notes where the note content matches a form of diagnosis and multiple sclerosis

```
/diagnos(?:e[\sd]|is])/i && $refhead[4] =~ /multiple\ssclerosis|\bms\b/i
```

2. Split the note into phrases

```
/(?:\.\s+)|(?:-(?:\s+)?)+|(?:>\s?<)|\\n/
```

3. If a phrase contains diagnos and multiple sclerosis in the same line, pull out diagnos and the 70 characters after

4. Save the year mentioned in the sub-phrase, noting if an exact year was mentioned or a relative referenced was used (i.e. years ago)

```
 /\d{4}/
/(\d{1,2})\s+years\s+ago/
/(one|two|three|four|five|six|seven|eight|nine|ten)\s+years\s+ago/)
```

5. If # years ago, subtract the years from the date of the note and save as the diagnosis year

6. Once all diagnosis years have been extracted and saved into a separate file:

7. For each individual, count the number of times each year was mentioned with an exact reference

7b. Output the year with the most counts as the final diagnosis year

8. For each individual without an exact year reference, count the number of times each year was mentioned with a relative reference

    8b. Output the year with the most counts as the final diagnosis year

EDSS

1. Select all notes

2. Select notes that contain ‘edss’

```
/edss/i
```

3. Split matching notes into lines

```
/(?:\.\s+)|(?:-(?:\s+)?){2,5}|(?:>\s?<)|\\n/
```

4. In each line, extract ‘edss’ and the following 50 characters

```
/\bedss\b.{0,50}/i
```

5. If the phrase contains digits preceded by non-letters, allowing for missing spaces (was5, is2.5, etc)

```
/(?:[^A-Za-z]|s)([\d][\d]?(?:\.[\d])?)/i
```

    5b. If the score is less than or equal to 10 and contains a decimal

        5bi. Save the number as the EDSS score

    5c. If the score is less than or equal to 10 and does not contain a decimal

        5ci. Add .0 to the number and save it as the EDSS score

6. If the phrase contains a number written out, convert the number to an integer and save as the EDSS score

```
/is\s+(zero|one|two|three|four|five|six|seven|eight|nine|ten)/
```

7. Remove duplicate values (ID, Note_Date, EDSS_score)

Timed 25 foot walk

1. Select all notes

2. Separate note into sentences

```
/\.\s+/
```

3. Select the sentence mentioning 25 foot walk

```
/timed?[\s-]*walk/i || $refhead[4] =~ /25[\s-]*feet/i || $refhead[4] =~ /25[\s-]*foot/i
```

4. Extract the number of minutes mentioned and convert it to 0-9 digits

```
/([0-9][0-9]?)\s*minute/i
/a\s+minute/i
/one\s+minute/i
```

5. Extract the number of seconds mentioned

```
/([0-9][0-9]?.?[0-9]*)\s*sec/i
```

6. Identify if a cane is mentioned without negative indication

```
/cane/i
/without [a|the] cane/i
```

7. Identify if a walked is mentioned

```
/wheelchair/i
```

8. Calculate the total number of seconds using the minutes and seconds

Year of first neurological symptom

1. Select letters and consultation notes

2. Select the first part of notes that contain ‘multiple sclerosis’ and begin with ‘Dear XX’

```
/multiple\s?sclerosis\s?/i
/dear.{0,1000}/i
```

3. If the first 1000 characters of the note contain references to the history of the illness, save the 200 characters following the phrase of interest for further processing

```
/.{0,50}(dat(?:e|ing)\s?back|began)(.{0,200})/i
```

4. If the saved text contains a four digit year, save it as the year of first symptom

```
/\d{4}/i
```

5. If the saved text references to a date shifted to preserve anonymity (format ‘DATE month_abbreviation’), use this month and the date of the note in which it was found to identify if the current year or previous year is when the symptoms began

```
/DATE\[(.{3})/i
```

6. If the saved text references a relative date (i.e. ‘years’ or ‘months ago’), use the current note date to calculate the year symptoms began

```
/\b(\w+)\s?years?/i
/\b(\w+)\s?months?/i
```

7. Export all individuals with dates to a file

Origin of first symptom

1. Select letters and consultation notes

2. Select the first part of notes that contain ‘multiple sclerosis’ and begin with ‘Dear XX’

```
/multiple\s?sclerosis\s?/i
/dear.{0,1000}/i
```

3. If the first 1000 characters of the note contain references to the history of the illness, export the first part of the note into a new file labeled with the person ID and note ID

```
/.{0,50}(dat(?:e|ing)\s?back|began)(.{0,200})/i
```

4. Run KnowledgeMap on all exported files

5. Load KnowledgeMap results in MySQL

6. Match concept unique identifiers (CUI) with concepts

7. Extract the outputs in categories related to MS symptoms
(sty = semantic type of the concept ; cui_pn = concept unique identifier name)

```
Sty ~= /disease or syndrome/i
Sty ~= /sign or symptom/i
Sty ~= /finding/i
Sty ~= /organ or tissue function/i and cui_pn ~= /sensory function/i
```

Match symptoms to their CNS origin, matching preference in the order below

8. Identify symptoms that stem from the optic nerve
(sty = semantic type of the concept; cui_pn = concept unique identifier name)

```
Original text =~ /eye\b/i
cui_pn =~ /optic\sneuritis/i
cui_pn =~ /vis/i and original text =~ /loss|acuity|diminish|decreas/xi
```

9. Identify symptoms originating in the brain stem
(cui_pn = concept unique identifier name)

```
Original text =~ /speech|dysphagia|face|diplopia|nystagmus/ix)
cui_pn =~ /fac|trigeminal\sneuralgia|tic\sdouloureux|tremor/ix
```

10. If original text matches arm or leg, distinguish limb complaints between brain stem and spinal cord

```
Original text =~ /arm|leg/ix
```

10a. If the original text mentions incoordination, identify as brain stem

```
/incoor/i
```

10b. If the original text mentions weakness or numbness, identify as spinal cord

```
/(weak|numb)/i
```

11. Identify other spinal cord symptoms
 (cui_pn = concept unique identifier name)

```
Original text =~ /numb|tingl|band|hug|lhermitte/i
cui_pn =~ /paresth/i
```

12. Identify urinary symptoms
(cui_pn = concept unique identifier name)

```
cui_pn =~ /urin|incontinence|bladder|bowel/ix  and cui_pn !~ /infect/i
```

13. Identify walking/balance difficulties as brain stem origin
(cui_pn = concept unique identifier name)

```
cui_pn =~ /walk|ataxia|balance|dizz/ix
cui_pn =~ /vertigo/ix
Original text =~ /vertigo/i
```

14. Load data into database

15. Match all symptom origins identified to each individual

Medications

1. Select all Problem Lists (PL) that contain ‘medication’

```
/medication/i
```

2. If any MS medications are found in the PL, print them out along with the ID and note date

```
/(?:interferon\sbeta.{0,2}1[ab])|
                avonex|
                rebif|
                betaseron|
                extavia|
                glatiramer\sacetate|
                copaxone|
                fingolimod|
                gilenya|
                natalizumab|
                tysabri|
                mitoxantrone|
                novantrone|
                teriflunomide|
                aubagio/ixg
```

APPENDIX C

R TEMPLATE FOR ORIGIN OF FIRST SYMPTOM ANALYSES

```
library(polytomous)
library(Hmisc)

##Read in origin categories
p = read.table("[file path]/IID_pheno.txt", header=TRUE,sep="\t")
##Read in eigenvalues
e = read.table("[file path]/keepoutliers.pca.evec_first3_withheader.txt",
header=TRUE,sep="\t")

##genotypes were coded in an additive model in PLINK and split into different
  #datasets with 1000 snps each

#read in genotypes and merge with the origins and eigenvalues
s = read.table("[file path]/split_dataset/recodeAXY.raw", header = TRUE, sep = " ")
c = merge(s,e)
f = merge(c,p)

##Polytomous regression analysis
#Cycle through each SNP (columns 7-1006)

for (j in 1:1000)
{
  index = j+6 #SNP column number

  #Save column name into snp_id and rename column to 'SNP'
  snp_id <- names(f)[index]
  names(f)[index] <- "SNP"

  f$origin <- factor(f$origin)

  anal <- f[,c("origin","e1","e2","e3","SNP")]
  anal2 <- na.omit(anal)



  #Run regression
  a = polytomous(origin ~ SNP + e1 + e2 + e3, data=anal2)

  #Revert column to SNP name
  names(f)[index] <- snp_id

  #Save results to a file
  sink(file=paste("[file path]/setXY/summaryXY.",index,"up.txt",sep=""))
  print(snp_id)
  print(a$p.values)
  sink()
}
```

APPENDIX D

R TEMPLATE FOR TIME TO SPMS ANALYSIS

```
library(survival)


##event indicator
p = read.table("time_to_event_SPMS.txt", header = TRUE, sep = "\t")
##eigenvalues
e = read.table("[file path]/keepoutliers.pca.evec_first3_withheader.txt", header=TRUE,sep="\t")


##genotypes were coded in an additive model in PLINK and split into different
  #datasets with 1000 snps each


#read in genotypes and merge with the event indicator and eigenvalues
s = read.table(file=paste("[file path]/split_dataset/recodeAXY.raw",sep=""), header = TRUE, sep = " ")
k = merge(s,p)
f = merge(k,e)


##Cox proportional hazards analysis
  ### CYCLE THROUGH EACH SNP (columns 7-1006)


for (j in 1:1000)
{
        index = j+6
        a = coxph(Surv(f$year_to_event, f$SPMS) ~ f[,index] + f$e1 + f$e2 + f$e3)
```

```
    ##Extract results to external file for further processing
    sink(file=paste("[file path]/setXY/summaryXY.",index,".txt", sep=""))
    print(a)
    sink()
}
```

REFERENCES

(1) Barcellos LF, Oksenberg JR, Green AJ, Bucher P, Rimmler JB, Schmidt S, et al. Genetic basis for clinical expression in multiple sclerosis. Brain 2002 Jan;125(Pt 1):150-8.

(2) Hauser SL, Goodkin DE. Multiple Sclerosis and Other Demyelinating Diseases. In: Fauci AD, Braunwald E, Isselbacher JD, Martin JB, Kasper DL, Hauser SL, et al., editors. Harrison's Principle of Internal Medicine. 14 ed. New York: McGraw Hill; 1998. p. 2409-19.

(3) Sadovnick AD. European Charcot Foundation Lecture: the natural history of multiple sclerosis and gender. J Neurol Sci 2009 Nov 15;286(1-2):1-5.

(4) Visser EM, Wilde K, Wilson JF, Yong KK, Counsell CE. A new prevalence study of multiple sclerosis in Orkney, Shetland and Aberdeen city. J Neurol Neurosurg Psychiatry 2012 Jul;83(7):719-24.

(5) Koutsouraki E, Costa V, Baloyannis S. Epidemiology of multiple sclerosis in Europe: a review. Int Rev Psychiatry 2010;22(1):2-13.

(6) Bencsik K, Rajda C, Fuvesi J, Klivenyi P, Jardanhazy T, Torok M, et al. The prevalence of multiple sclerosis, distribution of clinical forms of the disease and functional status of patients in Csongrad County, Hungary. Eur Neurol 2001;46(4):206-9.

(7) Papathanasopoulos P, Gourzoulidou E, Messinis L, Georgiou V, Leotsinidis M. Prevalence and incidence of multiple sclerosis in western Greece: a 23-year survey. Neuroepidemiology 2008;30(3):167-73.

(8) Pugliatti M, Sotgiu S, Rosati G. The worldwide prevalence of multiple sclerosis. Clin Neurol Neurosurg 2002 Jul;104(3):182-91.

(9) Dean G, Kurtzke JF. On the risk of multiple sclerosis according to age at immigration to South Africa. Br Med J 1971 Sep 25;3(5777):725-9.

(10) Elian M, Nightingale S, Dean G. Multiple sclerosis among United Kingdom-born children of immigrants from the Indian subcontinent, Africa and the West Indies. J Neurol Neurosurg Psychiatry 1990 Oct;53(10):906-11.

(11) Ascherio A, Munger KL. Environmental risk factors for multiple sclerosis. Part II: Noninfectious factors. Ann Neurol 2007 Jun;61(6):504-13.

(12) Ascherio A, Munger KL. Environmental risk factors for multiple sclerosis. Part I: the role of infection. Ann Neurol 2007 Apr;61(4):288-99.

(13) Ebers GC. Environmental factors and multiple sclerosis. Lancet Neurol 2008 Mar;7(3):268-77.

(14) Naito S, Namerow N, Mickey MR, Terasaki PI. Multiple sclerosis: association with HL-A3. Tissue Antigens 1972;2(1):1-4.

(15) Munger KL, Zhang SM, O'Reilly E, Hernan MA, Olek MJ, Willett WC, et al. Vitamin D intake and incidence of multiple sclerosis. Neurology 2004 Jan 13;62(1):60-5.

(16) Munger KL, Levin LI, Hollis BW, Howard NS, Ascherio A. Serum 25-hydroxyvitamin D levels and risk of multiple sclerosis. JAMA 2006 Dec 20;296(23):2832-8.

(17) Willer CJ, Dyment DA, Sadovnick AD, Rothwell PM, Murray TJ, Ebers GC. Timing of birth and risk of multiple sclerosis: population based study. BMJ 2005 Jan 15;330(7483):120.

(18) Fiddes B, Wason J, Kemppinen A, Ban M, Compston A, Sawcer S. Confounding underlies the apparent month of birth effect in multiple sclerosis. Ann Neurol 2013 Jun;73(6):714-20.

(19) O'Gorman C, Lucas R, Taylor B. Environmental risk factors for multiple sclerosis: a review with a focus on molecular mechanisms. Int J Mol Sci 2012;13(9):11718-52.

(20) Handel AE, Williamson AJ, Disanto G, Dobson R, Giovannoni G, Ramagopalan SV. Smoking and multiple sclerosis: an updated meta-analysis. PLoS ONE 2011;6(1):e16149.

(21) Hedstrom AK, Baarnhielm M, Olsson T, Alfredsson L. Tobacco smoking, but not Swedish snuff use, increases the risk of multiple sclerosis. Neurology 2009 Sep 1;73(9):696-701.

(22) Haahr S, Koch-Henriksen N, Moller-Larsen A, Eriksen LS, Andersen HM. Increased risk of multiple sclerosis after late Epstein-Barr virus infection: a historical prospective study. Mult Scler 1995 Jun;1(2):73-7.

(23) Lucas RM, Hughes AM, Lay ML, Ponsonby AL, Dwyer DE, Taylor BV, et al. Epstein-Barr virus and multiple sclerosis. J Neurol Neurosurg Psychiatry 2011 Oct;82(10):1142-8.

(24) McDonald WI, Compston A, Edan G, Goodkin D, Hartung HP, Lublin FD, et al. Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. Ann Neurol 2001 Jul;50(1):121-7.

(25) Poser CM, Paty DW, Scheinberg L, McDonald WI, Davis FA, Ebers GC, et al. New diagnostic criteria for multiple sclerosis: guidelines for research protocols. Ann Neurol 1983 Mar;13(3):227-31.

(26) Polman CH, Reingold SC, Banwell B, Clanet M, Cohen JA, Filippi M, et al. Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. Ann Neurol 2011 Feb;69(2):292-302.

(27) West TW. Transverse myelitis--a review of the presentation, diagnosis, and initial management. Discov Med 2013 Oct;16(88):167-77.

(28) Klawiter EC. Current and new directions in MRI in multiple sclerosis. Continuum (Minneap Minn) 2013 Aug;19(4 Multiple Sclerosis):1058-73.

(29) Collard RC, Koehler RP, Mattson DH. Frequency and significance of antinuclear antibodies in multiple sclerosis. Neurology 1997 Sep;49(3):857-61.

(30) Compston A, Coles A. Multiple sclerosis. Lancet 2002 Apr 6;359(9313):1221-31.

(31) Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology 1983 Nov;33(11):1444-52.

(32) Roxburgh RH, Seaman SR, Masterman T, Hensiek AE, Sawcer SJ, Vukusic S, et al. Multiple Sclerosis Severity Score: using disability and disease duration to rate disease severity. Neurology 2005 Apr 12;64(7):1144-51.

(33) Cutter GR, Baier ML, Rudick RA, Cookfair DL, Fischer JS, Petkau J, et al. Development of a multiple sclerosis functional composite as a clinical trial outcome measure. Brain 1999 May;122 (Pt 5):871-82.

(34) Hilas O, Patel PN, Lam S. Disease modifying agents for multiple sclerosis. Open Neurol J 2010;4:15-24.

(35) Major EO, Douek DC. Risk factors for rare diseases can be risky to define: PML and natalizumab. Neurology 2013 Aug 7.

(36) Ebers GC, Yee IM, Sadovnick AD, Duquette P. Conjugal multiple sclerosis: population-based prevalence and recurrence risks in offspring. Canadian Collaborative Study Group. Ann Neurol 2000 Dec;48(6):927-31.

(37) Robertson NP, O'Riordan JI, Chataway J, Kingsley DP, Miller DH, Clayton D, et al. Offspring recurrence rates and clinical characteristics of conjugal multiple sclerosis. Lancet 1997 May 31;349(9065):1587-90.

(38) Islam T, Gauderman WJ, Cozen W, Hamilton AS, Burnett ME, Mack TM. Differential twin concordance for multiple sclerosis by latitude of birthplace. Ann Neurol 2006 Jul;60(1):56-64.

(39) Willer CJ, Dyment DA, Risch NJ, Sadovnick AD, Ebers GC. Twin concordance and sibling recurrence rates in multiple sclerosis. Proc Natl Acad Sci U S A 2003 Oct 28;100(22):12877-82.

(40) Mumford CJ, Wood NW, Kellar-Wood H, Thorpe JW, Miller DH, Compston DA. The British Isles survey of multiple sclerosis in twins. Neurology 1994 Jan;44(1):11-5.

(41) Hansen T, Skytthe A, Stenager E, Petersen HC, Bronnum-Hansen H, Kyvik KO. Concordance for multiple sclerosis in Danish twins: an update of a nationwide study. Mult Scler 2005 Oct;11(5):504-10.

(42) Hansen T, Skytthe A, Stenager E, Petersen HC, Kyvik KO, Bronnum-Hansen H. Risk for multiple sclerosis in dizygotic and monozygotic twins. Mult Scler 2005 Oct;11(5):500-3.

(43) Multiple sclerosis in 54 twinships: concordance rate is independent of zygosity. French Research Group on Multiple Sclerosis. Ann Neurol 1992 Dec;32(6):724-7.

(44) Ristori G, Cannoni S, Stazi MA, Vanacore N, Cotichini R, Alfo M, et al. Multiple sclerosis in twins from continental Italy and Sardinia: a nationwide study. Ann Neurol 2006 Jan;59(1):27-34.

(45) Kuusisto H, Kaprio J, Kinnunen E, Luukkaala T, Koskenvuo M, Elovaara I. Concordance and heritability of multiple sclerosis in Finland: study on a nationwide series of twins. Eur J Neurol 2008 Oct;15(10):1106-10.

(46) Hawkes CH, Macgregor AJ. Twin studies and the heritability of MS: a conclusion. Mult Scler 2009 Jun;15(6):661-7.

(47) Ebers GC, Sadovnick AD, Risch NJ. A genetic basis for familial aggregation in multiple sclerosis. Canadian Collaborative Study Group [see comments]. Nature 1995 Sep 14;377(6545):150-1.

(48) Dyment DA, Yee IM, Ebers GC, Sadovnick AD. Multiple sclerosis in stepsiblings: recurrence risk and ascertainment. J Neurol Neurosurg Psychiatry 2006 Feb;77(2):258-9.

(49) Sawcer S, Ban M, Maranian M, Yeo TW, Compston A, Kirby A, et al. A high-density screen for linkage in multiple sclerosis. Am J Hum Genet 2005 Sep;77(3):454-67.

(50) Sawcer S. The complex genetics of multiple sclerosis: pitfalls and prospects. Brain 2008 Dec;131(Pt 12):3118-31.

(51) Gregory SG, Schmidt S, Seth P, Oksenberg JR, Hart J, prokop A, et al. Interleukin 7 receptor alpha chain (IL7R) shows allelic and functional association with multiple sclerosis. Nat Genet 2007 Jul 29.

(52) International Multiple Sclerosis Genetics Consortium. Risk Alleles for Multiple Sclerosis Identified by a Genomewide Study. N Engl J Med 2007 Jul 29;357(9):851-62.

(53) International Multiple Sclerosis Genetics Consortium. The expanding genetic overlap between multiple sclerosis and type I diabetes. Genes Immun 2009 Jan;10(1):11-4.

(54) Hafler DA, Compston A, Sawcer S, Lander ES, Daly MJ, De Jager PL, et al. Risk alleles for multiple sclerosis identified by a genomewide study. N Engl J Med 2007 Aug 30;357(9):851-62.

(55) De Jager PL, Jia X, Wang J, de Bakker PI, Ottoboni L, Aggarwal NT, et al. Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. Nat Genet 2009 Jun 14.

(56) Ban M, Goris A, Lorentzen AR, Baker A, Mihalova T, Ingram G, et al. Replication analysis identifies TYK2 as a multiple sclerosis susceptibility factor. Eur J Hum Genet 2009 Mar 18.

(57) Bush WS, Sawcer SJ, De Jager PL, Oksenberg JR, McCauley JL, Pericak-Vance MA, et al. Evidence for polygenic susceptibility to multiple sclerosis--the shape of things to come. Am J Hum Genet 2010 Apr 9;86(4):621-5.

(58) Brynedal B, Duvefelt K, Jonasdottir G, Roos IM, Akesson E, Palmgren J, et al. HLA-A confers an HLA-DRB1 independent influence on the risk of multiple sclerosis. PLoS ONE 2007;2(7):e664.

(59) Goris A, Pauwels I, Dubois B. Progress in multiple sclerosis genetics. Curr Genomics 2012 Dec;13(8):646-63.

(60) Sawcer S, Hellenthal G, Pirinen M, Spencer CC, Patsopoulos NA, Moutsianas L, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 2011 Aug 11;476(7359):214-9.

(61) Beecham AH, Patsopoulos NA, Xifara DK, Davis MF, Kemppinen A, Cotsapas C, et al. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nat Genet 2013 Nov;45(11):1353-60.

(62) Beecham AH, Patsopoulos NA, Xifara DK, Davis MF, Kemppinen A, Cotsapas C, et al. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nat Genet 2013 Nov;45(11):1353-60.

(63) Nelson MR, Wegmann D, Ehm MG, Kessner D, St JP, Verzilli C, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 2012 Jul 6;337(6090):100-4.

(64) Romero-Pinel L, Pujal JM, Martinez-Yelamos S, Gubieras L, Matas E, Bau L, et al. Epistasis between HLA-DRB1 parental alleles in a Spanish cohort with multiple sclerosis. J Neurol Sci 2010 Nov 15;298(1-2):96-100.

(65) Bush WS, McCauley JL, DeJager PL, Dudek SM, Hafler DA, Gibson RA, et al. A knowledge-driven interaction analysis reveals potential neurodegenerative mechanism of multiple sclerosis susceptibility. Genes Immun 2011 Jul;12(5):335-40.

(66) Ramagopalan SV, Dyment DA, Giovannoni G, Sadovnick AD, Ebers GC. HLA-DRB1*15, low infant sibling exposure, and multiple sclerosis gene-environment interaction. Ann Neurol 2010 May;67(5):694-5.

(67) van der Mei IA, Ponsonby AL, Taylor BV, Stankovich J, Dickinson JL, Foote S, et al. Human leukocyte antigen-DR15, low infant sibling exposure and multiple sclerosis: gene-environment interaction. Ann Neurol 2010 Feb;67(2):261-5.

(68) Martinez A, Alvarez-Lafuente R, Mas A, Bartolome M, Garcia-Montojo M, de LH, V, et al. Environment-gene interaction in multiple sclerosis: human herpesvirus 6 and MHC2TA. Hum Immunol 2007 Aug;68(8):685-9.

(69) Hensiek AE, Seaman SR, Barcellos LF, Oturai A, Eraksoi M, Cocco E, et al. Familial effects on the clinical course of multiple sclerosis. Neurology 2007 Jan 30;68(5):376-83.

(70) Sadovnick AD, Duquette P, Herrera B, Yee IM, Ebers GC. A timing-of-birth effect on multiple sclerosis clinical phenotype. Neurology 2007 Jul 3;69(1):60-2.

(71) Smestad C, Brynedal B, Jonasdottir G, Lorentzen AR, Masterman T, Akesson E, et al. The impact of HLA-A and -DRB1 on age at onset, disease course and severity in Scandinavian multiple sclerosis patients. Eur J Neurol 2007 Aug;14(8):835-40.

(72) Hensiek AE, Sawcer SJ, Feakes R, Deans J, Mander A, Akesson E, et al. HLA-DR 15 is associated with female sex and younger age at diagnosis in multiple sclerosis. J Neurol Neurosurg Psychiatry 2002 Feb;72(2):184-7.

(73) Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, Pulley JM, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. Am J Hum Genet 2010 Apr 9;86(4):560-72.

(74) Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther 2008 Sep;84(3):362-9.

(75) Pulley JM, Brace M, Bernard GR, Masys D. Evaluation of the effectiveness of posters to provide information to patients about a DNA database and their opportunity to opt out. Cell Tissue Bank 2007;8(3):233-41.

(76) McGregor TL, Van Driest SL, Brothers KB, Bowton EA, Muglia LJ, Roden DM. Inclusion of pediatric samples in an opt-out biorepository linking DNA to de-identified medical records: pediatric BioVU. Clin Pharmacol Ther 2013 Feb;93(2):204-11.

(77) Kho AN, Pacheco JA, Peissig PL, Rasmussen L, Newton KM, Weston N, et al. Electronic medical records for genetic research: results of the eMERGE consortium. Sci Transl Med 2011 Apr 20;3(79):79re1.

(78) Denny JC. Chapter 13: Mining electronic health records in the genomics era. PLoS Comput Biol 2012;8(12):e1002823.

(79) Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc 2013 Jan 1;20(1):117-21.

(80) Lependu P, Iyer SV, Bauer-Mehren A, Harpaz R, Mortensen JM, Podchiyska T, et al. Pharmacovigilance Using Clinical Notes. Clin Pharmacol Ther 2013 Mar 4.

(81) Delaney JT, Ramirez AH, Bowton E, Pulley JM, Basford MA, Schildcrout JS, et al. Predicting clopidogrel response using DNA samples linked to an electronic health record. Clin Pharmacol Ther 2012 Feb;91(2):257-63.

(82) Compston A, Coles A. Multiple sclerosis. Lancet 2008 Oct 25;372(9648):1502-17.

(83) Dobson R, Ramagopalan S, Davis A, Giovannoni G. Cerebrospinal fluid oligoclonal bands in multiple sclerosis and clinically isolated syndromes: a meta-analysis of prevalence, prognosis and effect of latitude. J Neurol Neurosurg Psychiatry 2013 Feb 21.

(84) Fischer JS, Rudick RA, Cutter GR, Reingold SC. The Multiple Sclerosis Functional Composite Measure (MSFC): an integrated approach to MS clinical outcome assessment. National MS Society Clinical Outcomes Assessment Task Force. Mult Scler 1999 Aug;5(4):244-50.

(85) Denny JC, Smithers JD, Miller RA, Spickard A, III. "Understanding" medical school curriculum content using KnowledgeMap. J Am Med Inform Assoc 2003 Jul;10(4):351-62.

(86) Denny JC, Spickard A, III, Miller RA, Schildcrout J, Darbar D, Rosenbloom ST, et al. Identifying UMLS concepts from ECG Impressions using KnowledgeMap. AMIA Annu Symp Proc 2005;196-200.

(87) Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010 May;17(3):229-36.

(88) Perlis RH, Iosifescu DV, Castro VM, Murphy SN, Gainer VS, Minnier J, et al. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. Psychol Med 2012 Jan;42(1):41-50.

(89) Ananthakrishnan AN, Cai T, Savova G, Cheng SC, Chen P, Perez RG, et al. Improving Case Definition of Crohn's Disease and Ulcerative Colitis in Electronic Medical Records Using Natural Language Processing: A Novel Informatics Approach. Inflamm Bowel Dis 2013 Jun;19(7):1411-20.

(90) Turchin A, Kolatkar NS, Grant RW, Makhni EC, Pendergrass ML, Einbinder JS. Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. J Am Med Inform Assoc 2006 Nov;13(6):691-5.

(91) Rosier A, Burgun A, Mabo P. Using regular expressions to extract information on pacemaker implantation procedures from clinical reports. AMIA Annu Symp Proc 2008;81-5.

(92) Garvin JH, DuVall SL, South BR, Bray BE, Bolton D, Heavirland J, et al. Automated extraction of ejection fraction for quality measurement using regular

expressions in Unstructured Information Management Architecture (UIMA) for heart failure. J Am Med Inform Assoc 2012 Sep;19(5):859-66.

(93) McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics 2011;4:13.

(94) Weber GM, Murphy SN, McMurry AJ, Macfadden D, Nigrin DJ, Churchill S, et al. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc 2009 Sep;16(5):624-30.

(95) Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. J Am Med Inform Assoc 2010 Jan;17(1):19-24.

(96) Hinks A, Cobb J, Marion MC, Prahalad S, Sudman M, Bowes J, et al. Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. Nat Genet 2013 Jun;45(6):664-9.

(97) Eyre S, Bowes J, Diogo D, Lee A, Barton A, Martin P, et al. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. Nat Genet 2012 Dec;44(12):1336-40.

(98) Tsoi LC, Spain SL, Knight J, Ellinghaus E, Stuart PE, Capon F, et al. Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. Nat Genet 2012 Dec;44(12):1341-8.

(99) Liu JZ, Almarri MA, Gaffney DJ, Mells GF, Jostins L, Cordell HJ, et al. Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. Nat Genet 2012 Oct;44(10):1137-41.

(100) Juran BD, Hirschfield GM, Invernizzi P, Atkinson EJ, Li Y, Xie G, et al. Immunochip analyses identify a novel risk locus for primary biliary cirrhosis at 13q14, multiple independent associations at four established risk loci and epistasis between 1p31 and 7q32 risk variants. Hum Mol Genet 2012 Dec 1;21(23):5209-21.

(101) Cooper JD, Simmonds MJ, Walker NM, Burren O, Brand OJ, Guo H, et al. Seven newly identified loci for autoimmune thyroid disease. Hum Mol Genet 2012 Dec 1;21(23):5202-8.

(102) Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, et al. Integrating common and rare genetic variation in diverse human populations. Nature 2010 Sep 2;467(7311):52-8.

(103) Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, et al. A map of human genome variation from population-scale sequencing. Nature 2010 Oct 28;467(7319):1061-73.

(104) Cortes A, Brown MA. Promise and pitfalls of the Immunochip. Arthritis Res Ther 2011;13(1):101.

(105) Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007 Jun 7;447(7145):661-78.

(106) Patsopoulos NA, Esposito F, Reischl J, Lehr S, Bauer D, Heubach J, et al. Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. Ann Neurol 2011 Dec;70(6):897-912.

(107) Shah TS, Liu JZ, Floyd JA, Morris JA, Wirth N, Barrett JC, et al. optiCall: a robust genotype-calling algorithm for rare, low-frequency and common variants. Bioinformatics 2012 Jun 15;28(12):1598-603.

(108) Teo YY, Inouye M, Small KS, Gwilliam R, Deloukas P, Kwiatkowski DP, et al. A genotype calling algorithm for the Illumina BeadArray platform. Bioinformatics 2007 Oct 15;23(20):2741-6.

(109) Giannoulatou E, Yau C, Colella S, Ragoussis J, Holmes CC. GenoSNP: a variational Bayes within-sample SNP genotyping algorithm that does not require a reference population. Bioinformatics 2008 Oct 1;24(19):2209-14.

(110) Ritchie ME, Liu R, Carvalho BS, Irizarry RA. Comparing genotyping algorithms for Illumina's Infinium whole-genome SNP BeadChips. BMC Bioinformatics 2011;12:68.

(111) Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007 Sep;81(3):559-75.

(112) Grady BJ, Torstenson E, Dudek SM, Giles J, Sexton D, Ritchie MD. Finding unique filter sets in PLATO: a precursor to efficient interaction analysis in GWAS data. Pac Symp Biocomput 2010;315-26.

(113) Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006 Aug;38(8):904-9.

(114) Gauderman WJ. Sample size requirements for association studies of gene-gene interaction. Am J Epidemiol 2002 Mar 1;155(5):478-84.

(115) Gauderman WJ. Sample size requirements for matched case-control studies of gene-environment interaction. Stat Med 2002 Jan 15;21(1):35-50.

(116) Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich HA, et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet 2009 Jun;41(6):703-7.

(117) Yang SK, Hong M, Zhao W, Jung Y, Tayebi N, Ye BD, et al. Genome-wide association study of ulcerative colitis in Koreans suggests extensive

overlapping of genetic susceptibility with Caucasians. Inflamm Bowel Dis 2013 Apr;19(5):954-66.

(118) Wade TD, Gordon S, Medland S, Bulik CM, Heath AC, Montgomery GW, et al. Genetic variants associated with disordered eating. Int J Eat Disord 2013 Apr 9.

(119) Wu JH, Lemaitre RN, Manichaikul A, Guan W, Tanaka T, Foy M, et al. Genome-wide association study identifies novel loci associated with concentrations of four plasma phospholipid fatty acids in the de novo lipogenesis pathway: results from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium. Circ Cardiovasc Genet 2013 Apr;6(2):171-83.

(120) Wu C, Kraft P, Stolzenberg-Solomon R, Steplowski E, Brotzman M, Xu M, et al. Genome-wide association study of survival in patients with pancreatic adenocarcinoma. Gut 2012 Nov 24.

(121) Divaris K, Monda KL, North KE, Olshan AF, Lange EM, Moss K, et al. Genome-wide association study of periodontal pathogen colonization. J Dent Res 2012 Jul;91(7 Suppl):21S-8S.

(122) Vasan RS, Glazer NL, Felix JF, Lieb W, Wild PS, Felix SB, et al. Genetic variants associated with cardiac structure and function: a meta-analysis and replication of genome-wide association data. JAMA 2009 Jul 8;302(2):168-78.

(123) Mizuki N, Meguro A, Ota M, Ohno S, Shiota T, Kawagoe T, et al. Genome-wide association studies identify IL23R-IL12RB2 and IL10 as Behcet's disease susceptibility loci. Nat Genet 2010 Aug;42(8):703-6.

(124) Remmers EF, Cosan F, Kirino Y, Ombrello MJ, Abaci N, Satorius C, et al. Genome-wide association study identifies variants in the MHC class I, IL10, and IL23R-IL12RB2 regions associated with Behcet's disease. Nat Genet 2010 Aug;42(8):698-702.

(125) Mells GF, Floyd JA, Morley KI, Cordell HJ, Franklin CS, Shin SY, et al. Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis. Nat Genet 2011 Apr;43(4):329-32.

(126) Liu X, Invernizzi P, Lu Y, Kosoy R, Lu Y, Bianchi I, et al. Genome-wide meta-analyses identify three loci associated with primary biliary cirrhosis. Nat Genet 2010 Aug;42(8):658-60.

(127) Oppenheim RW, Houenou LJ, Parsadanian AS, Prevette D, Snider WD, Shen L. Glial cell line-derived neurotrophic factor and developing mammalian motoneurons: regulation of programmed cell death among motoneuron subtypes. J Neurosci 2000 Jul 1;20(13):5001-11.

(128) Peterziel H, Unsicker K, Krieglstein K. TGFbeta induces GDNF responsiveness in neurons by recruitment of GFRalpha1 to the plasma membrane. J Cell Biol 2002 Oct 14;159(1):157-67.

(129) Yang W, Tang H, Zhang Y, Tang X, Zhang J, Sun L, et al. Meta-analysis followed by replication identifies loci in or near CDKN1B, TET3, CD80, DRAM1, and ARID5B as associated with systemic lupus erythematosus in Asians. Am J Hum Genet 2013 Jan 10;92(1):41-51.

(130) Lee YH, Bae SC, Choi SJ, Ji JD, Song GG. Genome-wide pathway analysis of genome-wide association studies on systemic lupus erythematosus and rheumatoid arthritis. Mol Biol Rep 2012 Dec;39(12):10627-35.

(131) Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 2012 Nov 1;491(7422):119-24.

(132) Gorlova O, Martin JE, Rueda B, Koeleman BP, Ying J, Teruel M, et al. Identification of novel genetic markers associated with clinical phenotypes of systemic sclerosis through a genome-wide association strategy. PLoS Genet 2011 Jul;7(7):e1002178.

(133) Allanore Y, Saad M, Dieude P, Avouac J, Distler JH, Amouyel P, et al. Genome-wide scan identifies TNIP1, PSORS1C1, and RHOB as novel risk loci for systemic sclerosis. PLoS Genet 2011 Jul;7(7):e1002091.

(134) Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP, et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat Genet 2010 Jun;42(6):508-14.

(135) Hom G, Graham RR, Modrek B, Taylor KE, Ortmann W, Garnier S, et al. Association of systemic lupus erythematosus with C8orf13-BLK and ITGAM-ITGAX. N Engl J Med 2008 Feb 28;358(9):900-9.

(136) Ponti G, Conti L, Cataudella T, Zuccato C, Magrassi L, Rossi F, et al. Comparative expression profiles of ShcB and ShcC phosphotyrosine adapter molecules in the adult brain. Neuroscience 2005;133(1):105-15.

(137) Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, et al. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. Genome Biol 2009;10(11):R130.

(138) Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A 2004 Apr 20;101(16):6062-7.