From DEPARTMENT OF CLINICAL NEUROSCIENCE
Karolinska Institutet, Stockholm, Sweden

MODELING GENETIC SUSCEPTIBILITY TO MULTIPLE SCLEROSIS

Helga Westerlind

Karolinska Institutet

Stockholm 2014

All previously published papers were reproduced with permission from the publisher.
Cover art created by Andreas Gillberg.
Published by Karolinska Institutet.
Printed by US-AB

ISBN 978-91-7549-567-5

Karolinska Institutet

Department of Clinical Neuroscience

Modeling genetic susceptibility to Multiple Sclerosis

AKADEMISK AVHANDLING
som för avläggande av medicine doktorsexamen vid Karolinska Institutet offentligen försvaras i föreläsningssal Rockefeller.

Måndagen den 9:e juni, 2014, kl 09.00

av
Helga Westerlind
Civilingenjör

Principal Supervisor:
Professor Jan Hillert
Karolinska Institutet
Department of Clinical Neuroscience

Co-supervisors:
Professor Timo Koski
Kungliga Tekniska Högskolan
Department of Mathematics

PhD Ryan Ramanujam
Karolinska Institutet
Department of Clinical Neuroscience

Docent Ingrid Kockum
Karolinska Institutet
Department of Clinical Neuroscience

PhD Izaura Lima bomfim
Karolinska Institutet
Department of Clinical Neuroscience

Opponent:
Assistant Professor Eli Stahl
The Mount Sinai School of Medicine
Department of Psychiatry

Examination Board:
Professor Niklas Dahl
Uppsala Universitet
Department of Immunology

Professor Erik Ingelsson
Uppsala University Hospital
Department of Medical Sciences

Professor Henrik Grönberg
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Stockholm 2014

ABSTRACT

The main aim of this thesis was to investigate genetic and environmental factors and their role in the etiology of Multiple Sclerosis (MS) by using comprehensive registry data or novel computationally intense methods. To date, over 100 genes associated with MS have been identified, but how they interact in the risk for the disease is not yet fully understood. The presence of high prevalence clusters has led researchers to believe that there might be as yet unidentified rare variant involved in the disease etiology. In Paper I, we attempted to search for these rare variants by using a population based linkage approach, estimating haplotypes shared between individuals inherited by descent from some common ancestor. One significant hit was found on chromosome 19, but due to methodological problems the result should be interpreted with caution.

MS is commonly attributed high familial risks, decreasing with relatedness, which indicates a large genetic component involved in the disease etiology. In Paper II, nationwide registry data was used to reinvestigate the familial risks and estimate the proportion of genetics and environment contributing to disease etiology. The relative risks estimated were lower than usually reported, with a sibling relative risk of 7.1 and no significant differences between the sexes. The heritability was estimated to be 64% and the environmental 36% with a non-significant shared environmental component of 1%.

In Paper III, the women-to-men ratio for MS in Sweden was reinvestigated. MS is a disease more common in women than men, and an increase in the women-to-men ratio has been reported in several countries. However, a report from Sweden did not show this increase in women and Paper III extended this report using data from nationwide registers. An increase among women compared to men was identified, and when comparing against the previous study, an inclusion bias, presumably caused by a higher mortality rate among the oldest men, was identified.

One framework used to model complex diseases such as MS is the sufficent cause model, also known as Rothman's pie model. This model hypothesizes that a disease can be caused by several mechanisms, or pies, each consisting of a set of different factors and when all factors are present they will inevitably cause disease. Paper IV extends this model into a stochastic version and presents an algorithm that can estimate the probability that an a priori suggested mechanism has caused disease in a certain individual. The algorithm showed high classification accuracy on synthetic data; however it needs further investigation of its properties.

In conclusion, this thesis revise the familial risks for MS to more moderate levels, with no differences between the sexes, and confirms the global trend of an increasing women-to-men ratio. No rare variants contributing to MS on population level were identified. We also present a probabilitic version of Rotman's pie model, showing promising results on synthetic data.

For my dad, Henrik Westerlind.

LIST OF SCIENTIFIC PAPERS

I. **Identity-by-descent mapping in a Scandinavian multiple sclerosis cohort**
Westerlind Helga, Imrell Kerstin, Ramanujam Ryan, Myhr Kjell-Morten, Gulowsen Celius Elisabeth, Harbo Hanne F, Bang Otturai Anette, Hamsten Anders, Hall Per, Alfredsson Lars, Olsson Tomas, Kockum Ingrid, Koski Timo, Hillert Jan, Manuscript

II. **Modest familial risks for multiple sclerosis – a registry based study of the population of Sweden**
Westerlind Helga, Ramanujam Ryan, Uvehag Daniel, Kuja-Halkola Ralf, Boman marcus, Bottai Matteo, Lichtenstein Paul, Hillert Jan, Brain 2014 Mar;137(Pt 3):770-8

III. **New data identify an increasing sex ratio of multiple sclerosis in Sweden**
Westerlind Helga, Boström Inger, Stawiarz Leszek, Landtblom Ann-Marie, Almqvist Catarina, Hillert Jan. Multiple Sclerosis Journal, In press

IV. **The learning for mixtures of multicausal interaction networks**
Westerlind Helga, Jääskinen Väinö, Corander Jukka, Hillert Jan, Koski Timo, Manuscript

CONTENTS

1 Introduction 1
1.1 Multiple sclerosis 1
1.1.1 Risk factors 1
1.2 Genetics 3
1.2.1 Genetic variation 4
1.2.2 Linkage disequilibrium 4
1.2.3 Genetic association studies 4
1.3 Epidemiology 5
1.3.1 Causation 5
1.3.2 Confounding 5
1.3.3 Bias 6
1.3.4 Complex diseases 6
1.3.5 Case-control studies 6
1.4 Models for complex diseases 6
1.4.1 Liability threshold model 7
1.4.2 Sufficient cause model 7
2 Aim of thesis 9
2.1 Overall AIM 9
2.2 Paper I 9
2.3 Paper II 9
2.4 Paper III 9
2.5 Paper IV 9
3 Materials and methods 10
3.1 Materials 10
3.1.1 Genetic study population 10
3.1.2 Registers 11
3.1.3 Synthetic data 13
3.2 Statistical methods 13
3.2.1 P-values 13
3.2.2 Multiple testing 14
3.2.3 Chi square 14
3.2.4 Relative risk 14
3.2.5 Odds ratio 15
3.2.6 Tetrachoric correlations 15
3.2.7 Measures of interaction 15
3.2.8 Identical-by-descent (IBD) 15
3.2.9 Sjögren's unmodified method 16
3.2.10 Linear regression 16
3.2.11 Cox regression 16
3.2.12 Kaplan-Meier 16
3.2.13 Heritability analysis 17
3.2.14 Noisy OR 17
3.2.15 Classification Expectation Maximization 17
4 Study summaries 19
4.1 Paper I 19
4.1.1 Results 20
4.2 Paper II 20
4.2.1 Results 21
4.3 Paper III 22
4.3.1 Results 23
4.4 Paper IV 24
4.4.1 Results 26

5 Discussion 27
5.1 Paper I 27
5.1.1 Findings and implications 29
5.1.2 Conclusions and future perspectives 30
5.2 Paper II 31
5.2.1 Findings and implications 33
5.2.2 Conclusions and future perspectives 33
5.3 Paper III 34
5.3.1 Findings and implications 35
5.3.2 Conclusions and future perspective 35
5.4 Paper IV 36
5.4.1 Findings and implications 37
5.4.2 Conclusions and future perspectives 37
5.5 General conclusions and future perspectives 37
5.5.1 Personal remark 39
6 Acknowledgements 41
7 References 43
8 Appendix 51

LIST OF ABBREVIATIONS

AUC	Area under curve
CEM	Classification expectation maximization
CI	Confidence intervals
cM	centiMorgan
CNS	Central nervous system
DNA	Deoxyribonucleic acid
DZ	Dizygotic
EBV	Epstein-Barr virus
EIMS	Epidemiological indestigation of multiple sclerosis
GWAS	Genome Wide Association Study
HLA	Human Leukocyte Antigen
HMM	Hidden Markov Model
IBD	Identical-by-descent
IBS	Identical-by-state
ICD	International classification of disease
LD	Linkage disequilibrium
LT	Liability threshold
LTSD	Liability threshold with sex dimorphism
MGR	Multi generation registry
MRI	Magnetic resonance imaging
MS	Multiple Sclerosis
MZ	Monozygotic
PAR	Swedish In-patient registry
PBLA	Population based linkage analysis
PIN	Personal identity number
QC	Quality control
RR	Relative risk
SC	Sufficient cause

SCB	Statistics Sweden (Statistiska Centralbyrån)
SMSreg	Swedish Multiple Sclerosis Registry
SNP	Single nucleotide polymorphism
STR	Swedish twin registry
TPR	Total population registry
VAL	Stockholm primary care registry

1 INTRODUCTION

This section will first give a brief introduction to the potential mechanisms underlying the disease etiology of multiple sclerosis (MS) and its clinical manifestation. It will also briefly cover the basic genetic and epidemiological concepts used in this thesis.

1.1 MULTIPLE SCLEROSIS

MS is a chronic disease of the central nervous system (CNS). Multiple scars, scleroses, form in the brain and spinal cord, caused by a demyelinating event presumably generated by the immune system. An inflammatory component definitely exists, whereas the exact process of these events, such as if inflammation causes demyelination or vice versa, is not yet clarified [1]. Clinically, the activation of an old lesion, or development of a new lesions in CNS, present with neurological symptoms [2]. A magnetic resonance imaging (MRI) of the brain will reveal old and active lesions. Examples of common symptoms are numbness, weakness, fatigue, double vision, pain and gradually diminished walking capacity [2]. To permit a diagnosis of MS according to McDonald's criteria, the lesions or attacks have to be separated in space and time, meaning one attack is not enough to establish the diagnosis, and neither is multiple occurrences of one and the same symptom [3].

The disease most often takes on a relapsing remitting course [1], with bouts that the patient recovers from more or less completely [2]. After some years, around 65% of patients enter what is called secondary progression [1], with a steadily and gradually increasing disability without relapses. It seems like this occurs at an average age of about 44 years, less dependent on time of onset [4]. A smaller proportion of around 20% of patients [1] have a primarily progressive course already from start, gradually worsening with time. The recurring focal inflammation, although largely reversible, is associated with some degree of tissue loss, which in time leads to atrophy of the brain and spinal cord [1]. Even though there is no cure for MS, patients are commonly offered disease modifying treatment with drugs that modify or suppress the immune system and thereby decreases the number of inflammatory lesions and the number of bouts [5]. Interferon beta was the first approved MS treatment, but in recent years an increasing number of new drugs are available [5].

1.1.1 Risk factors

The prevalence of MS varies across the globe. It is commonly accepted that the prevalence increases with the latitude, however, there are studies reporting otherwise [6],[7]. Sweden is a high prevalence country; a study from 2011 reports a prevalence of 188.9 cases per

100,000 [8]. A recent study from Norway reported an even higher prevalence of 203 cases per 100,000, but no latitude gradient within the country [6]. Both incidence rates of MS and its prevalence are reported to increase across the world [7] at least partially attributed to the increased survival of the patients [7]. From Sweden there is no recent updated incidence report.

The women-to-men ratio for MS has in several populations increased throughout the 20th century, with reports of up to 3 times as high risk for a woman to get the disease [7],[9]. This rapid increase of women with MS suggests that some environmental factor is at play and involved in the pathogenesis, possibility interacting with a genetic factor, as a single genetic factor would need longer time to show such a large effect.

Studies of familial risks have reported high concordance rates for MS relatives, reducing with decreasing relatedness. For monozygotic (MZ) twins, figures as high as 25% has been reported [10], but a recent meta analysis revised these figures to more moderate risks with an age adjusted risk for MZ twins of 18.44%, and a relative risk of 116.69 [11]. These high estimates and the decrease of risk with lesser relatedness indicate a large genetic contribution to the etiology.

1.1.1.1 Genetic risk factors

MS is an etiologically complex disease with polygenic inheritance [12]. The first identified risk factor for MS was HLA DRB1*15:01 established in the 1970s [13]. Human Leukocyte Antigens or HLA, are antigen presenting heterodimer globulins, presenting foreign antigens on the cell surface to cells of the immune system. Associations with HLA alleles are typical for autoimmune diseases indicating that peptide presentation and specific peptide recognition by the immune system is central, and MS is thus believed to be such a disease by the larger part of the scientific community. Even so, some argue that the autoimmunity involved in MS may be a secondary and not necessarily a primary event [14].

For some decades, HLA DRB1*15:01 was the only identified genetic association in MS, until in 2000 an independent protective association to HLA*A02:01 was discovered [15]. HLA*A:02 is a Class I antigen presenting molecule, foremost presenting peptides originating from within the cell. How this HLA Class I-association can cause a protective effect is not yet fully understood.

In 2007, the IL7R gene became the first non-HLA gene associated with MS [16]. Introduction of DNA microarray chip typing techniques, which enabled rapid and cheaper genotyping across the genome, as well as large international collaborations collecting sample sets big enough for sufficient power to investigate smaller effects, contributed to today's list of over 100 genes associated with MS [17].

Many of the genes with an association to MS have been reported as important for the immune system, suggesting either antigen presenting and/or immune dys-regulation involved in the disease pathogenesis [18] . However, both when selecting candidate genes and candidate pathways for further investigation, there has amongst researchers been a certain bias towards choosing mechanisms involved in particularly T-cell regulation in the study design [18].

1.1.1.2 Environmental risk factors

Through epidemiological studies, a number of environmental risk factors associated to MS have been identified. Large efforts have been made to collect lifestyle information, such as sending out lifestyle questionnaires to large cohorts of patient and controls. This has resulted in that there today is a steadily increasing list of MS risk factors, such as smoking [19], lack of sun exposure [20], low vitamin D levels [21] and body mass index [22]. In the case of smoking, an interaction with the HLA genes, increasing the risk for MS, has also been reported [23].

Among environmental factors, not only lifestyle, but also viruses, have been reported as associated to MS, primarily Epstein-Barr virus (EBV) [24]. There is however no clear evidence for a virus directly involved in triggering the disease, and the role and association for EBV is debated [25],[26].

Excluding HLA, a common denominator for all genetic and environmental risk factors for MS is that their individual risks are small, but they are frequent within the population. However it is expected as the search, in most studies, has been for associations in common variants.

1.2 GENETICS

This section will in brief highlight some concepts of genetics and genetic variation relevant for this thesis.

1.2.1 Genetic variation

Certain base pair positions in the genome vary commonly within the human population and may therefore be used as genetic markers referred to as *single nucleotide polymorphisms* (SNPs). The particular variant of a certain polymorphism, e.g. of a SNP is referred to as an *allele*, whereas a certain place at the genome is called a *locus*.

SNP genotyping is readily automated and SNPs are thus typically used in genome wide association studies (GWAS) where thousands of patients and controls are compared for hundreds of thousand markers. Only a small minority of SNPs currently associated to complex disease are located in exons, some other in introns but the majority not even within genes. Thus, such associations are hard to interpret for involvement in disease. There is hope that techniques such as *exome sequencing* may identify variants in the exomes that are more directly involved in disease, and possibly rare variants [27].

1.2.2 Linkage disequilibrium

When the *germ cells* are formed, *recombination* between the maternally and the paternally inherited chromosomes is a frequent event. The unit for recombination is Morgan (M), which measures the distance between two markers where we can expect one recombination per generation. More often the higher resolution measure centiMorgan (cM) is used. Recombinations do not occur at random, and certain locations in the genome have very high rates of recombination, these are called *recombination hotspots* [28].

A sequence of alleles located together on a chromosomal segment is referred to as a *haplotype*. The correlation between two positions or markers is called *linkage disequilibrium* (LD). LD differs between populations due to historic evolutionary events [29].

1.2.3 Genetic association studies

An association study examines frequencies of the marker/-s of interest between cases and controls. With the introduction of chip typing technology, the prices for genotyping dropped and instead of studying association in a single candidate gene, it became possible to type for several hundred thousand to over 1 million SNPs across the genome. The result from such an investigation will be a measure of the association between the marker and the trait at study.

A significant association does not say much about the direction of a possible causal relationship between the marker and the outcome, but as genotype precedes disease, it is generally assumed that an associated genetic marker reflects a mechanism more or less directly involved in the pathogenesis of the disorder in study (so called indirect association). An association is therefore not enough to establish a causal relationship, as there also might be chance, bias or confounding giving rise to an association. Therefore it is important to account for this in the statistical analysis or in the study design. LD structure and non-homogeneous populations can cause spurious association, and either ensuring that the cases and controls are ethnically homogeneous or correction for population stratification must be performed in the analysis [30].

1.3 EPIDEMIOLOGY

Epidemiology is the study of how a certain disease appears within a population. This section will in brief cover concepts of epidemiology relevant for the thesis.

1.3.1 Causation

When studying the epidemiology of complex diseases, the ultimate goal is often to understand its causes, to allow possible prevention and cure. But causation and causality is a hard task, as proving causal involvement is difficult within an observational study, as often is the case in medical sciences. In fact, mathematics is the only science that really proves a strict causal relationship. Other sciences mostly have to make do with probabilities and evidence speaking in favor or against a certain hypothesis.

The principle behind causation is that is if A causes B, B will inevitably occur if A occurs, and A has to occur before B. Bradford-Hill set up 9 criteria that can be used as a checklist when studying causality [31]. The criteria are: strength of association, consistency, specificity, temporality, biologic gradient, plausibility, coherence, experimental evidence and analogy.

1.3.2 Confounding

If X is associated to Y, but an unknown, unmeasured variable Z has an effect on both X and Y, Z is confounding. If Z is not included in the model, it will bias the estimated strength of the effect X has on Y. As a confounding variable might be an unknown, unmeasured variable, it might be hard in reality to include all confounding factors in the model [32].

1.3.3 Bias

A systematic error that is not random is called a *bias* [33]. Bias can be unintentionally introduced for example in the inclusion phase or in the analysis phase. One example of inclusion bias is that concordant twin pairs are more likely to participate in a study advertising for volunteers [34]. This needs to be accounted for and included when interpreting the results.

1.3.4 Complex diseases

The classical Mendelian inheritance patterns, dominant and recessive, are present only for a small minority of disease states even among the mostly genetic diseases. In a dominant inheritance pattern, one erroneous copy is enough to cause disease. In a recessive case, both copies of the gene need to be "bad" for disease to develop, and the individual with one functional copy stay healthy. This pattern of inheritance is valid in diseases as well as other traits, such as eye color.

Frequently, one gene is not enough to explain the trait, and for diseases, several or many genes are involved in combination with environmental factors. Such disorders are referred to as complex disorders, and MS falls within this category.

1.3.5 Case-control studies

In a case/control study, individuals are included in the study based on being a case or not, as opposed to cohort studies, where all persons within a group are included regardless of their outcome. In a traditional study design, one control is matched per case, but in some scenarios more controls per case can be beneficial.

In a case/control design, the causative event will by definition have occurred for all cases before they are enrolled in the study and thus, claims of causation can be hard to make.

1.4 MODELS FOR COMPLEX DISEASES

When working with complex disease, it is useful to hypothesize about underlying model causing the disease. The work in this thesis is based on two different principle models: the liability threshold model (LT) and the sufficient cause model (SC).

1.4.1 Liability threshold model

In the LT model, an individual is assumed to have a certain load of liability for a certain disease. This load is assumed to be normally distributed within the population. If an individual has a load above a certain threshold, disease will inevitably occur. Figure 1a shows a theoretical distribution of liability within a population with liability load on the x-axis and frequency in the population on the y-axis. Carter introduced a version of this model designated as the liability threshold model with sex dimorphisms (LTSD), as an explanation for why pyloric stenosis was more often transmitted to the offspring from the lesser prevalent sex [35]. The threshold was assumed to be different between men and women, requiring women, the lesser prevalent sex, to have a higher threshold to develop disease (Figure 1b). As men required less genetic risk load to pass the threshold, the sons of the affected mothers, would more often get the disease.

1.4.2 Sufficient cause model

The sufficient cause model, also known as Rothman's pie model, is another model for complex disease [33]. Here it is hypothesized that a certain set of risk factors will inevitably trigger the disease if all of those risk factors are present in one individual. If not, disease will not occur. A set of risk factors causing the disease can be referred to as a pie, and one disease can have several pies that can potentially be causative.

Let the disease Y be caused by a the set of risk factors of either {A, B, C} or {C,D,E,F} or {A,B,F}. These sets could also be referred to as mechanism. Figure 2 shows a graphical representation of the model. A more formal way of describing the SC model would then be

$$Y = ABC \vee CDEF \vee ABF$$

This is the logical OR gate, and as such a deterministic model. OR should in this context be read as the coordinating conjunction "or" and is capitalized due to tradition in the fields of logics and electronics. It should not be confused with the abbreviation for odds ratio (section 3.2.5), and to facilitate for the reader of this thesis and avoid confusion, OR will in this thesis always refer to the logical expression. Paper IV uses the noisy-OR gate, a probabilistic version of this model, to classify patients into one of the a priori suggested underlying mechanisms.

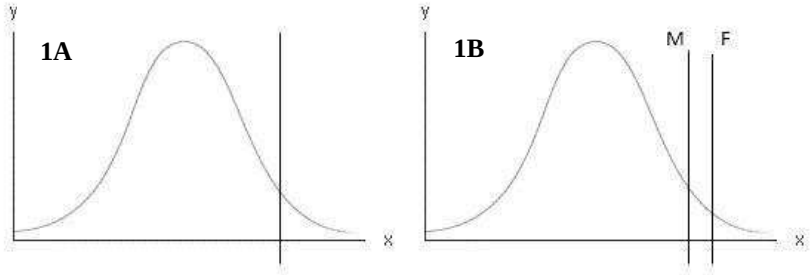

Figure 1: The liability threshold model
1a shows the LT model with liability load on the x-axis and frequency in the population on the y-axis. 1b shows the LTSD model as proposed by Carter, where the lesser prevalent sex (females in pyloric stenosis) would require a higher genetic load than men to get the disease.

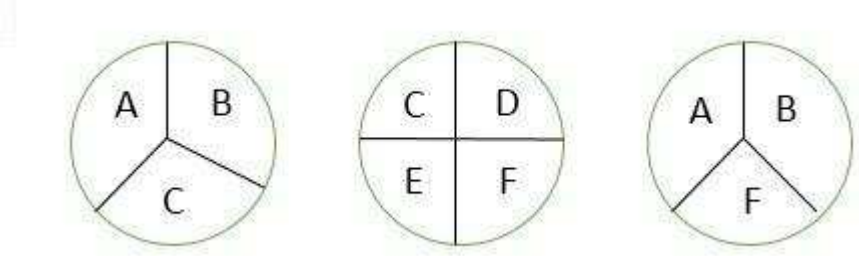

Figure 2: The sufficient cause model
Figure 2 shows the graphical representation of the three mechanisms mentioned in section 1.4.2.

2 AIM OF THESIS

2.1 OVERALL AIM

The overall aim of the thesis is to investigate genetic and environmental factors and their role in the etiology of MS by using comprehensive registry data or novel computationally intense methods.

2.2 PAPER I

The aim of paper I was to try to identify possible rare variants contributing to MS by estimating haplotypes shared identical-by-descent more frequently among case-case pairs reusing GWAS data, hoping these would map to rare disease associated variants.

2.3 PAPER II

The aim of paper II was to reinvestigate the familial recurrence risks in MS by using comprehensive national registers and matched controls.

2.4 PAPER III

The aim of paper III was to reinvestigate the women-to-men ratio for MS in Sweden using the full MS population.

2.5 PAPER IV

The aim of paper IV was to develop a novel method to classify individuals into etiological subgroups of the proposed underlying mechanisms in the SC model.

3 MATERIALS AND METHODS

3.1 MATERIALS

All ethical permissions were obtained from the regional ethics board and patient consent was received according to the Declaration of Helsinki. All registry data were obtained from the Crime database located at the Department of Medical Epidemiology and Biostatistics at KI.

3.1.1 Genetic study population

In Paper I the study population was incident Swedish MS patients included in the project Epidemiological Investigation of MS (EIMS). These were combined with prevalent cases from Denmark and Norway genotyped in the same project (see below). The controls were matched to Swedish individuals participating in the EIMS project, or non-matched healthy controls from the Procardis [36],[37] study, or from Cahres, a large study on breast cancer patients [38]. All patients and controls were of Nordic ancestry.

The cases and the incidence (EIMS) controls were genotyped on the Illumina human quad 660 chip during a large international collaboration [18]. The additional controls were typed on Illumina HumanHap 550 (Cahres controls) and Illumina 1M (Procardis controls), and the genotypes for these individuals were called later with the same algorithm as the cases.

3.1.1.1 Quality control

Although the genotypic data used in Paper I had undergone a previous quality control (QC) before publication of the initial GWAS analysis [18], to avoid problems when adding the external controls an additional QC, was performed using the program PLINK [39]. The parameters were set as: minor allele frequency: 0.05, Hardy Weinberg equilibrium 1e-6 and a missingness per individual of 0.07.

3.1.1.2 Outlier removal

To ensure population homogeneity of the genetic analysis in Paper I, the software Eigenstrat's smartPCA algorithm [30] was run, and the 6 most significant principal components were used to cluster the individuals with a nearest neighbor method, requiring 10 neighbors with a maximum Euclidian distance of 0.15. The analysis script is freely available for both R [40] and MATLAB [41] on http://www.kirc.se.

3.1.2 Registers

The Crime database contains anonymized data from several nationwide registers linked on the personal identity number (PIN) by Statistics Sweden (Statistiska centralbyrån (SCB)). These registry data were used in Papers II and III.

3.1.2.1 SMSreg

The Swedish MS registry (SMSreg) was established in 2000, although local efforts had been ongoing for some years previously. To date, it is the nationwide registry used by most MS specialists in Sweden, containing about 14,000 MS active patients. Having entered patient data, SMSreg provides the clinician with a partly graphic, partly tabular overview of central clinical data such as demographics, dates of onset and diagnosis, diagnostic findings, disease course, disability according to the expanded disability status scale ongoing immunomodulatory treatments and some laboratory results. MS clinicians use SMSreg voluntarily, supposedly because they find is useful as a decision support tool. SMSreg was used both for patient identification and age at onset estimations in Papers II and III.

3.1.2.2 Swedish in-patient registry (PAR)

In 1968 a pilot study of the Swedish In-patient registry (PAR) was started. This was taken nationwide in 1970, and the data can be considered complete from 1989 [42]. Specialist and out-patient care were added in 2001. For disease identification, International Classification of Disease (ICD)-codes are used.

3.1.2.3 The Stockholm primary care registry (VAL)

The Stockholm Primary Care registry (VAL) was used in Paper II. It contains data on outpatient visits to health care in Stockholm since 2001. VAL was only used in Paper II.

3.1.2.4 Swedish twin registry (STR)

The Swedish twin registry (STR) contains most of the twin births in Sweden since 1890, and is one of the most complete twin registries in the world. Zygosity is determined by DNA and/or questionnaire [43]. To the date of the study it contained over 190,000 Swedish twins. STR was used in Paper II for identification of twins and zygosity.

3.1.2.5 Total population registry (TPR)

The Total Population Registry (TPR) contained slightly below 15 million individuals with a Swedish PIN at the time of the studies. The PIN was introduced in 1961 for everyone born in Sweden during 1932 or later and residing for a longer period of time. Birth year, month of birth and sex was obtained from TPR in Paper II. In Paper III, country of birth was also used in the analysis.

3.1.2.6 Cause of death registry

The Cause of Death registry contains date and cause of death (registered with an ICD-code) for deaths in Sweden. For Papers II and III, only year and month of the deaths were used. In Paper II, it was used for age adjusted risks and as the time to censoring in the relative risk calculation. In Paper III, it was used for the mortality analysis in the oldest patient group.

3.1.2.7 Multi generation registry (MGR)

The Multi generation registry (MGR) holds information on parents, and possible adoptive parents for more than 9 million index people with a Swedish PIN. MGR was used to identify relatives in Paper II.

3.1.2.8 Identification of patients

An individual was classified as an MS patient, if they were either in SMSreg or in PAR with an ICD code for MS, ICD 8 (340), ICD 9 (340), ICD 10 (G35).

3.1.2.9 Selection of controls

To compare the risk for the relatives of MS patients against an accurate background risk, controls were selected at random. For every MS-relative pair, up to ten control-relative pairs matched on year of birth, gender and the relative's relation to the index-patient were randomly selected. The controls were required to be alive at the time of the MS patient's age at onset.

3.1.2.10 Estimation of age at onset

In Paper II an age at onset estimate was used for basic descriptive statistics and as time to event or censoring in the survival analysis. For the individuals identified through SMSreg, the age-at-onset as estimated by a neurologist was used. For the patients identified through

PAR, the age at the first entry in PAR was used. Although this is not the actual age at onset but age at first hospitalization or visit to specialist care and was on average was 14 years later than the age at onset in SMSreg it is used in the article and throughout the rest of this thesis, unless stated otherwise.

3.1.3 Synthetic data

For Paper IV, synthetically generated data was used to learn the characteristics and test the properties of the model. The synthetic data was generated by first assigning every individual a classification, D_j, with a probability corresponding to the frequency of that mechanism. Based on that classification, the covariates were randomized so that $\Pr(x_i = 1) = 1 - \varepsilon$, where ε is a small number, if x_i was part of the assigned mechanism , and $\Pr(x_i=1) <= 0.5$ otherwise.

Two mechanisms were used, and their frequencies were randomized between 0.2 and 0.8. The parameters ψ_{ij} for each mechanism were randomized between 0.1 and 0.5 to get fairly equal proportions of cases and controls.

The outcome of the individual was calculated by determining the value of F where $F = 1 - e^{\sum_{i \in D_j} \psi_{ij} x_i}$. If $F \leq r$, where $r \in (0,1)$, Y was set to 1, otherwise 0.

3.2 STATISTICAL METHODS

This section will cover the statistical methods and models used in the different studies.

3.2.1 P-values

A *p-value* is the observed significance level of data under a null hypothesis (H_0) [44]. If the p-value is below a certain threshold, H_0 is rejected in favor of H_1, the alternative hypothesis. The threshold most often used for significance testing in medical research is 0.05. This means we can accept falsely rejecting H_0 one time out of 20. This threshold has become the gold standard for medical research. P-values were used in studies I-III to assess significance of the observed statistic.

3.2.2 Multiple testing

When performing multiple dependent tests, it is desired to adjust the significance threshold of 0.05, even though more tests than one are performed. One common way to correct for this is to apply Bonferroni correction. This is done by either multiplying the observed p-value with the number of tests performed, or dividing the threshold with the number of tests. Papers II and III corrected for multiple testing using Bonferroni correction. In Paper I, significance and threshold for significance were assessed by using a permutation analysis. The permutation analysis was run on every 10^{th} marker, using five million randomly generated case/control statuses or until it was ensured that the result was not significant. The genome wide threshold was calculated by using the significance level of every thousand marker and taking the 5^{th} percentile of this distribution as threshold for significance. There are others ways of determining this significance cut off for identity-by-descent (IBD) mapping, however they are less accurate [45]. The permutation analysis was part of an analysis pipeline created by Browning and Thompson [45].

3.2.3 Chi square

In Paper III, a chi square test was performed to assess sampling differences between SMSreg and the nationwide registers. A chi square tests the null hypothesis that there are no differences between the observed and expected number of observations from two or more random variables by calculating the test statistic *Q* where

$$Q_{obs} = \sum_{j=1}^{r} \frac{(x_j - np_j)^2}{np_j}$$

If the observed number of observations is close to the expected, Q_{obs} will approximate a $\chi^2(r-1)$ distribution, where *r* is the number of classes. To assess significance Q_{obs} is compared against the $\chi^2(r-1)$ distribution and if it is greater than the value for the desired significance threshold, H_0 is rejected [44].

3.2.4 Relative risk

A measure of the risk for the disease for the group exposed to a factor is the *relative risk* (RR). It is calculated using the risk for disease, and thus not applicable to the case/control design, as the enrichment for cases in this design will give rise to distorted frequencies. RRs were estimated in paper II using a Cox regression to assess the relative risks for the relatives to the MS patients.

3.2.5 Odds ratio

When using a case/control design the odds ratio will be an approximation of the RR under the assumption that it is a rare disease, or a sampling strategy such as incidence-density sampling is used [46]. The odds ratio is calculated as the odds for the exposed group versus the odds for the disease in the unexposed group. An odds ratio can be approximated for example using a contingency table or a logistic regression. Paper IV discusses measures of interaction (see more in section 3.2.7) using odds ratios.

3.2.6 Tetrachoric correlations

Tetrachoric correlation can be used as an estimate of heritability [47] . It is a measure of correlation between two variables, both assumed to have underlying normal distributions, but observed as binary variables due to an underlying threshold dichotomizing the variables. This is the case we are assuming when we use the LT model and therefore tetrachoric correlations were estimated in Paper II.

3.2.7 Measures of interaction

Under the assumption of the SC model, interaction can be defined. Rothman divided these into biological and statistical interaction. Biological interaction was defined as departure from an additive effect for the combined risk factors and statistical interaction was defined as the interaction term in a logistic regression significantly different from 0. A statistical interaction would not necessarily imply a biological interaction was present [33]. As Paper IV is built on the SC model, measures of interactions corresponding to the ones proposed by Rothman are provided for the noisy-OR model.

3.2.8 Identical-by-descent (IBD)

Haplotypes estimated to be shared IBD meaning, two individuals shared a segment that was inherited from a common, distant ancestor, were identified with the software Beagle's refined IBD method [48]. Refined IBD uses a dictionary approach to identify the segments, and to estimate if these are shared IBD or not, a probabilistic assessment is made. Refined IBD has been shown to have better accuracy and correctly identifying a larger number of segments for outbred populations compared to other methods [48],[49], and was therefore the method of choice in Project I.

3.2.9 Sjögren's unmodified method

In Paper II, absolute risks were calculated for the relatives. As all relatives to MS patients hadn't lived through the entire risk period, age corrected risks, adjusting for the time at risk passed, were calculated by using Sjögren's unmodified method [50]. By weighing all individuals against an age-at-onset distribution calculated beforehand, the denominator is decreased proportionally to the time at risk passed for all individuals. The age-at-onset distribution based on data from SMSreg was used for this. The confidence intervals were calculated using the total sum of the weights as denominator.

3.2.10 Linear regression

To investigate the significance of slope of the line in Paper III, a linear regression was used with the ratio for the birth cohort as the outcome and the birth cohort as predictor. To adjust for the differences in age at hospitalization, the mean age at first hospitalization for the full cohort was included as a covariate in the model.

3.2.11 Cox regression

The relative risks for the relatives in Paper II were estimated with Cox proportional hazards models.

The Cox proportional hazards model consists of two parts: an unspecified baseline hazard function, $h_o(t)$, and the covariates expressed as a linear equation:

$$h_i(t) = h_0(t)\exp(\beta_1 x_{i1} + \cdots + \beta_k x_{ik}).$$

This will give the hazard for individual *i* at time *t*. The resulting regression coefficients can be used as relative risks [51],[52].

All factors matched for in the selection process of the controls and control relatives were included as covariates in the model, as was the age at diagnosis of the MS patient. As a case could occur more than once, for example having more than one sibling, a robust sandwich estimator was used to estimate the standard error.

3.2.12 Kaplan-Meier

The Kaplan-Meier estimator is a non-parametric maximum-likelihood estimation of the survival function estimating the proportion of individuals surviving past a time *t* [51],[52]

$$\hat{S}(t_j) = \hat{S}(t_{j-1})\, x \Pr(T < t_j | T \geq t_j).$$

Kaplan-Meier was used in Paper III to investigate differences in mortality rate between men and women in the oldest birth cohort. The analysis was conducted in SAS version 9.2 using the PROC LIFETEST statement.

3.2.13 Heritability analysis

In the model for heritability used in Paper II, the variance of the trait is hypothesized to be fully explained by the environment and the genetics. The environment is divided into a shared environmental component, and a non-shared component. By using data from MZ and dizygotic (DZ) twin pairs, an equation system is set up. It is known how much genetics are shared for the pairs, and it can be assumed the twins are mostly reared together and thus share the environment,. For this the program OpenMx was used [53] in the statistical program R.

3.2.14 Noisy OR

In the SC model, a full set of mechanisms will inevitably cause the disease. The noisy OR-grid is the probabilistic version of this model, introducing the probability that the event will take place due to a certain mechanism. This model was developed by Pearl in 1988 [54]. Thus the outcome of the model in our setting is the probability that a certain mechanisms triggered disease in a certain individual.

The formalization of the SC model as written in section 1.4.2 will in the probabilistic framework look like:

$$\Pr(Y = y | \underline{\psi}, \underline{\alpha}, \underline{x}) = \sum_{j=1}^{M} \alpha_j p(y | \underline{\psi_j}, j, \underline{x}).$$

Where α is the probability of a certain mechanism, ψ is the factors within that mechanism, and x represents the covariates for individual i.

3.2.15 Classification Expectation Maximization

A Classification Expectation Maximization (CEM) technique was used to iteratively find the maximized likelihood for the noisy-OR gate. The algorithm consists of three steps: *E*-

step, where the current posterior probability is calculated for all observations and all mechanisms. The *C-step* when every individual is assigned to a new mechanism based on the current posterior probability. And finally the *M-step*, in which a maximum likelihood estimation using the updated parameters from the two previous steps is performed.

The likelihood for the CEM algorithm consists of two parts, $L(\alpha)$ and $L(\psi)$. The full derivation of these formulas can be seen in Paper IV.

4 STUDY SUMMARIES

Here, the studies are summarized in short and the results from respective studies are given.

4.1 PAPER I

Even though multi-case families with MS are rare, high prevalence clusters have been reported in for example Värmland [55] and Överkalix [56] in Sweden and Bothnia in Finland [57]. This has lead researchers to believe that some rare variant might be acting on the susceptibility to the disease and in some cases associations have been found [58], but neither of the findings reported from these clusters have been replicated in larger cohorts.

In Paper I it was hypothesized that possible rare variants associated to disease could be identified by performing a linkage analysis on the population of Scandinavia. Although Scandinavia is not a founder population, the individuals tend to cluster together in principal component analysis [18], and would as such be suitable for a population based linkage (PBLA) analysis.

The first computer program introduced for PBLA analysis was PLINK's segmental sharing algorithm [39]. PLINK uses a hidden Markov model (HMM) [59] for pairwise IBD estimation based on the identical-by-state (IBS) status. The underlying HMM methodology requires initial pruning of the genetic data to avoid dependencies violating the Markov assumption. As LD in the genome exist not only between markers following each other sequentially, but also between markers with larger distances in between them, it is questionable if the Markov assumption in reality can be met.

PLINK was at the start of this project the only published algorithm for IBD detection and was initially used. Whether the violation of the Markov assumption caused a problem or not, we cannot tell, but what caused some concern was the cM distances used by the program. As we initially had no data on cM measures, estimates were used and small changes in the cM estimates seemed to cause drastic changes in the output, making significant peaks appear and disappear. During the time of the analysis, more methods for IBD detection had been published, such as Germline [60] and fastIBD [49]. In the benchmarks conducted by Browning and Browning, PLINK had substantially less statistical power than these new algorithms. It thus seemed like fastIBD would be a more suitable method for our project. Browning and Thompson had also published a pipeline for IBD mapping [45]. This pipeline included the analysis from start with detection of

segments with the fastIBD algorithm, to significance testing with a permutation analysis. It was therefore decided to change to this method.

The analysis time turned out to be exhaustive, with hardware requirements that were not easy to meet and months spent on the permutation analysis. After personal communication with the authors of Beagle at the American Society for Human Genetics congress in 2012, it turned out they now had improved their method further into refined IBD. Refined IBD had more statistical power for detection of the shorter segments [48], and as we had not a traditional isolate, but a more outbred population, it was decided to change methods to refined IBD but keep the analysis pipeline developed by Browning and Thompson.

A script converting the data from refined IBD to fastIBD was developed. This script can be found on http://www.kirc.se.

4.1.1 Results

After QC and outlier removal, 3,953 MS patients and control with genotypic data were used.

Significant peaks were found at the very telomeric ends of chromosomes 1, 7, 15 and 19. The authors confirmed this to be a known artifact of the method, and a filter removing the markers with the 10 lowest percent of IBD coverage was introduced. After this filter, only one marker not in a telomeric position was left. This marker was located on chromosome 19 in the gene GNA11. No significant hits were found in the HLA region.

4.2 PAPER II

In a series of paper based on MS patients in Canada by Ebers and Sadovnick and co-workers, high recurrence risks for family members to MS patients were reported [10],[61]–[63]. The risk was highest for MZ twins, decreasing with declining relatedness. This pattern would indicate a large genetic contribution to the disease. Furthermore, an increased risk on the maternal side has been found in the studies, indicating a parent-of-origin effect involved in the etiology [64]–[66].

The material gathered in Canada, and other studies on recurrence risks, were mostly based on data collected from the clinics, or after advertising for volunteers. Method for confirmation of MS diagnosis in relatives has varied, ranging from asking the patient [67] to actively seeking and examining the relatives that, based on the patient's description, could potentially show neurological symptoms indicative of MS [10],[68] or a mix [69].

In 2005, a case registry based paper from Denmark presented lower risks for relatives to MS patients [70]. By using registries, results will not have to depend on patients' recollection of possibly affected relatives. A study from 2009 based on Swedish registry data also reported lower risks [71].

With the studies from Sweden and Denmark being exceptions, it seemed like the genetic background risk differed between countries, as countries with a higher prevalence of MS tended to have higher familial risks.

At the Department of Medical Epidemiology and Biostatics at Karolinska Institutet Professor Paul Lichtenstein has developed a method for estimating recurrence risk by using registry data and controls with relatives matched to the patient and the patient's relative. Identifying relatives through registries provides more accurate identification of relatives and their possible diagnosis. By using the registries covering most medical care in Sweden, the possible bias of enriching for women and concordant pairs that could arise from clinics based studies is decreased. In Paper II this method was used to reinvestigate the familial recurrence risks for MS in Sweden.

A heritability analysis, to estimate the proportion of the genetic and environmental contribution to the disease etiology, was also conducted. A previous review of the twin studies made on MS showed estimate in a wide range, and with broad confidence intervals [72].The analysis in project II was based on twins identified through the STR, and to increase statistical power of the analysis, the analysis was expanded with close to 2.5 million full and half siblings.

To address the parent-of-origin effect, stratification by sex for the familial relative risks was performed. We also conducted an analysis of transmission to offspring, also called the Carter effect, which has previously show conflicting results in studies on MS [73],[74].

In total, over 28,000 MS patients were identified, an estimated 96% of the Swedish MS patients, making it the largest and most complete study of familial risks in one single population in MS so far.

4.2.1 Results

The absolute risks were in the same range as a meta analysis published earlier during 2013 [11]. However, the relative risks were lower than previously published. The sibling recurrence risk was found to be as low as 7.13 compared to in the meta analysis published 16.67 [11]. The twin risk was 23.62 compared to 116.69 in the meta analysis, with non-

overlapping confidence intervals [11]. There were no significant difference in transmission to offspring between the sexes, and neither was there any difference in risks between maternal and paternal relatives to the MS patients. The heritability analysis estimated the genetic component to be around 64% (95% confidence interval (CI): 36%-76%), the shared environmental component to 1% (CI: 0%-18%) and the non-shared environment to 35% (CI: 24%-51%).

4.3 PAPER III

MS is more frequent among women than men. Interestingly, an analysis of MS patients in Canada showed that the women-to-men ratio was increasing throughout the 20th century. [9]. The study was based on birth cohorts divided by five year periods from 1931 and until 1981. The result was to some extent replicated in studies from Norway [75] and Denmark [7], but a study from Sweden failed to replicate the increase [76]. The study from Sweden was based on data from SMSreg and showed a high women-to-men ratio for all birth cohorts. The lack of increase in the sex-ratio was surprising, and a possible explanation could be that an environmental factor had been present in Sweden before the other countries. The increase in women would already had taken place and could therefore not be identified in the data.

In Paper III, we aimed to reanalyze the sex-ratio in Sweden by adding the MS patients identified in Paper II, giving us almost twice the study population as the initial study. VAL was excluded to avoid the possibility of a bias due to having a more dense coverage of the capital. Furthermore, everyone born outside Sweden was excluded.

The calculations were based on prevalence proportions. By using prevalence proportions, we could calculate the prevalence of each birth cohort independently of the others and later compare the cohorts with other statistics.

A sensitivity analysis was performed by investigating the sex-ratio in five different subcohorts: MS obtained as primary diagnosis, MS obtained as secondary diagnosis, MS diagnosis given to an in-patient, MS diagnosis given to out-patients and MS diagnosis given during or after 1989, which is the year when PAR is considered to have full coverage [42].

As the increase in women became apparent, we investigated a possible explanation for the difference in results between the studies from Sweden. For both studies, we had access to the number of MS patients included in each birth cohort and compared these with a chi

square test. Focusing on the oldest birth cohort, we used the cause of death registry to obtain age at death and with a Kaplan-Meier analysis investigated possible difference in mortality rate between the sexes.

No comparison of mortality against the general population was performed due to lack of ethical permission and limitations of the project.

The environmental factor proposed to contribute to the increase of MS in women has been speculated to be attributable to western lifestyle. Smoking has been one factor of interest [77], but as we had no access to smoking data, this could not be investigated.

Another suggested environmental factor is the increasing age at which women give birth to the first child. Several studies from Denmark have investigated the effect pregnancy might have on age at onset [78],[79]. A post hoc analysis was therefore conducted by calculating at what age the women with MS gave birth to their first child, and comparing this figure to the general population.

4.3.1 Results

It was found that the sex-ratio in Sweden increased from 1.70 for patients born in the 1930ies to around 2.6 for patients born in later cohorts. Figure 3 shows the result from the sensitivity analysis with a consistent increase for all subgroups.

The mean increase per year was estimated using a linear regression, and to correct for differences between the cohorts in time between age at onset to age at hospitalization, an estimated difference for each cohort based on SMSreg and PAR was included as a covariate. The slope of the line turned out significant with a mean increase of 0.11 units per birth cohort.

When comparing the women and men in the oldest birth cohort identified in SMSreg against PAR, 18% of the women compared to the 10% of the men were in SMSreg (p-value < 0.001). Upon further analysis, it was found that the mean age at death was the same for both sexes, but a larger proportion of the men had died. A Kaplan-Meier analysis confirmed this difference to be significant (p-value < 0.001).

The age for women when giving birth to the first child increased between the birth cohorts, but did not differ from the general population.

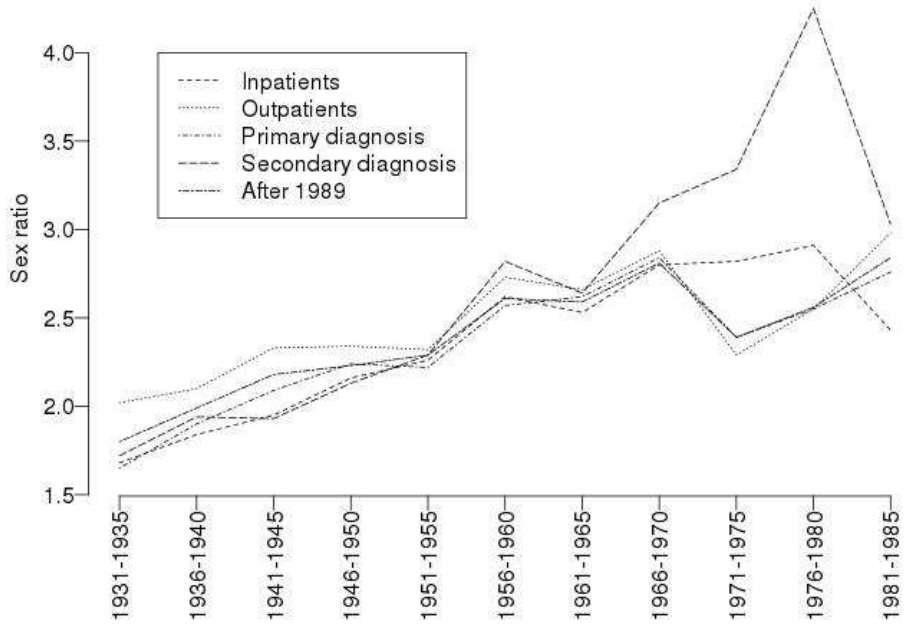

Figure 3: Sensitivity analysis of the sex-ratio in Sweden. The peak for the secondary diagnosis for patients born 1976-1980. Consists of a very small number of patients (245/61).

4.4 PAPER IV

Modeling complex disease with an underlying deterministic model is both intuitive and appealing as the concept of causality is more easily understandable if disease is said to occur if and only if all causal factors are present. In reality, this will introduce difficulties, as all risk factors for diseases like MS are not yet identified. Subgrouping patients based on underlying disease etiology and thereby attempting to explain a complex disease more fully in a subgroup of patients, is a research area that has gotten more attention in recent years. Paper IV is a project attempting to estimate the probability that an individual developed disease due to a particular cause.

As mentioned in section 1.4.2, the SC model is the classical OR gate, and as such a well-known research topic. The noisy-OR gate, which is the stochastic version used in Paper IV,

was proposed by Pearl in 1988 [54]. The noisy-OR gate enables us to estimate the probability that a certain mechanism has caused disease in a certain individual.

In Paper IV, it is assumed that these subgroups are already identified, or at least hypothesized, and they need to be provided in the model. A CEM is run to estimate the parameters, classifying individuals to the *a posteriori* most likely disease causing mechanism.

The log-likelihood consists of two parts: the log-likelihood for the probabilities of the mechanisms (alpha), and the log-likelihood for the parameter estimates of the different mechanisms (psi). The log-likelihood for alpha turns out to be a trivial computation, but for the parameter estimates, Fischer's method of scoring [80] was used. Fischer's method of scoring a one-step iteration to solve the partial derivatives in the Hessian, thereby decreasing the computational time spent in this step.

This project is a collaboration between Professor Jan Hillert's group at Karolinska Institutet, Professor Timo Koski at the department of Mathematics at The Royal Institute of Technology (KTH), and the Bayesian statistics group led by Professor Jukka Corrander at the University of Helsinki. The mathematical theory was developed by Professor Koski, and an initial implementation of the CEM in MATLAB was made by the student Väinö Jääskinen in the Bayesian statistics group. In the Hillert group, the MATLAB algorithm was translated into R, and in collaboration with Professor Koski, developed further and benchmarked.

To test the properties of the model, the accuracy, sensitivity and specificity were investigated using synthetic data. The accuracy was determined by estimating psi and alpha with the CEM algorithm, using new starting values, and calculating the deviation from the value used to generate the data. The starting values for alpha were randomized between 0.2 and 0.8 and for psi between 0.1 and 0.5, and 3,000 individuals were randomly generated 100 times.

The sensitivity and specificity was estimated by calculating the area under curve (AUC) for a validation set of patients using the values from a training set consisting of 20% of a total sample of 10,000 individuals. The AUC was calculated using the package ROCR [81] and taking the mean AUC of 10 repeated subsampling validations.

To investigate convergence of the CEM algorithm, 500 CEM iterations for 100 randomized data sets were run, and the mean value of the log-likelihood was calculated per iteration.

As the properties of the model on more complex data sets were not yet investigated, it was decided to postpone testing on real data until the behavior of the model had been further investigated.

4.4.1 Results

The mean deviation of the resulting estimates from the alpha and psi used to generate the synthetic data, can be seen in Table 1. The AUC was estimated to be 0.89. The mean log-likelihood increased with 0.5 during the 500 iterations.

	Ψ_{11}	Ψ_{12}	Ψ_{13}	Ψ_{21}	Ψ_{22}	Ψ_{23}	α_1	α_2
Mean deviation	0.008	-0.025	0.017	-0.015	-0.013	-0.002	-0.014	0.014
Standard deviation	0.19	0.17	0.17	0.17	0.18	0.19	0.23	0.23

Table 1: Mean and standard deviation of deviation from the estimated values in the CEM benchmark.

5 DISCUSSION

The discussion is divided by paper into three parts: first, a general methodological discussion is given. This discussion will focus mainly on the underlying model/-s and the impact the modeling assumptions might have on the interpretation of the results. Also, general methodological and mathematical considerations are discussed. The second part covers the findings in the studies and the implications these might have for the field in general. The last part of the discussion deals with how the knowledge gained from the studies might be used in future research. As the end of the Thesis I have attempted to draw general conclusions from the studies and how this can be taken further, ending with a personal remark about the future of the field.

5.1 PAPER I

The theoretical underlying model for complex disease used in Paper I is the SC-model. It is assumed that there are several different mechanisms causing MS, and that (at least) one of these disease causing mechanisms contains a rare genetic variant. In theory, inherited haplotypes could tag this variant [29], and by using an entire population, the variant would be sufficiently enriched for.

Traditionally, linkage has been the method of choice to identify haplotypes tagging rare variants. A disease causing variant is hypothesized to be aggregated within a family with several cases. In other words: there's a special pie causing the disease in that particular family, and one of the pieces in that pie is a rare genetic variants. Using the family's pedigree, the inheritance of the haplotypes between the generations can be traced, and the rare variant identified.

However, multi-case families in MS are rare, and success with linkage strategies have been sparse. In Paper I, it was attempted to take the search for rare variants to a population level by using PBLA, also called IBD mapping.

The so called "missing heritability", was another rational to look for rare variants. In MS, GWAS data has been used to estimate the variance explained, reporting that not more than 30% of the variation is explained by all markers [82] . There might of course be different reasons for the missing heritability such as not including the right markers, and indeed some work has shown that including more SNPs explains significantly more of the variance [83]. In this project, it was hypothesized that the missing heritability partly was due to yet unidentified rare variants, possibly of a greater effect.

Some methodological problems were encountered when using PBLA. The significant hits at the very telomeric ends of chromosomes strongly suggested this could be an artifact of the method, which the authors of the method confirmed (personal communication). As suggested by the authors, the markers covered by the ten lowest percent of general IBD detection were removed, and with them most of the significant hits disappeared. The permutation analysis was made before the filtering, and the threshold for significance was set to 5.3e-6 which corresponds to a genome wide significance threshold of 0.05. The authors confirmed the problem to be known for determining the end of the segments, but when applying the suggested filter, many hits located in the beginning of segments also disappeared, and it is reasonable to suspect the methodological concern would apply to starting points as well as to ends.

After filtering, only one significant hit was left that was not in a telomeric position. Upon closer inspection, it seems to be located in a recombination hotspot. Figure 4 (page 30) shows the pattern of the identified segments for the case-case pairs. The first significant signal on chromosome 19, which was removed during the filtering, consisted of the five first very telomeric markers in the permutation analysis. The second signal, also removed during the filtering contained markers 18-21 in the permutation analysis (markers 181-211 in the marker map). These markers seem to be located in the first recombination hotspot. The third signal, consisting of only one marker, marker 43 in the permutation analysis (431 on the map), seems to be in the second recombination hotspot. Could the recombination hotspots in the beginning of chromosome 19 possibly give rise to the association?

There is also a concern regarding the statistical power of the analysis. Although refined IBD detects 10 times the segments fastIBD detects [48], and fastIBD outperforms most other methods [49], refined IBD still only detects about 0.4% of the IBD segments. However, this figure include all segments shared IBD, including the shorter ones. The longer segments investigated, the more power refined IBD has [48].

Apart from mapping disease variants, segments shared IBD are thought to give information about the population structure. There are several projects trying to utilize this [84],[85], but as the algorithm identifying the segments was yet not so robust, it could not be argued to proceed with further investigation of the population structure in Scandinavia. Some theoretical remarks can nevertheless be made: our cut off of 1 cM is motivated by the Pareto distribution: segments of a shorter length are more likely not relevant for the analysis. The 1 cM cut-off gives us an assumption that we are looking at segments theoretically shared from an ancestor 50 generations ago [45]. A higher cut off, such as 3 cM, could be motivated by LD structure. In paper from 1977 by Watterson

and Guess [86] it is concluded that as rare alleles are generally younger, and we can expect the LD to extend further [29],[86].

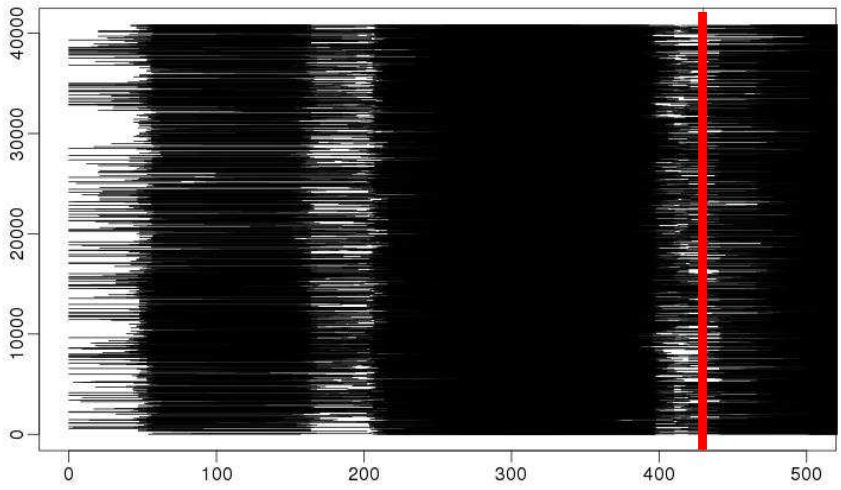

Figure 4 The pattern of IBD segments among the case-case pairs. The vertical line indicates the significant marker. The first significant hit that was filtered out was located to the very left in the beginning of the segments. The 2nd signal was located in the recombination hotspot around markers 170-200.

The distribution of the length of the segments shows that it is at is densest slightly above 1 cM. Setting the cut off higher would risk removing too many segments and with them statistical power of the analysis. As this is an analysis of Scandinavia and not of a founder population, it can be argued that it is relevant to include segments of length between 1 and 3 cM. Assuming 20 years between the generations, this would take us back around 1,000 years in time, to a couple of years after the introduction of Christianity and the beginning of medieval times.

5.1.1 Findings and implications

After filtering, only one marker not in a telomeric position reached significance. This marker is located on chromosome 19, a chromosome that has previously showed up in several linkage studies of MS. But as the marker seems to be located in a recombination hot spot, it should be interpreted with caution until it has been replicated in an independent

cohort. Thus, we see no confirmed signs of rare variants contributing to the disease on population level in Scandinavia. This is consistent with the lack of multi case families and the lower level of familial aggregation and lack of results with linkage. Taking the success in the search for common variants using GWAS [18] and further fine mapping of these hits in the Immunochip collaboration [17] into account, this further implies lack of rare variants contributing to MS.

We could thus not add any piece of the puzzle regarding the missing heritability, but that is assuming refined IBD was sufficiently powered to tag a rare variant in the sample, and that the hypothesized rare variant was sufficiently enriched for in the sample.

5.1.2 Conclusions and future perspectives

We cannot confirm the presence of a rare variant contributing to MS, but the methods used contained some problems.

An ongoing international project on exome sequencing will hopefully cast further light upon a possible rare variant involved in disease.

A problem encountered that is not mentioned much here, is exhaustive analysis time. Months were spent on the permutation analysis, which used over 30 GiB of RAM memory, an amount that might not be easily accessible. The permutation analysis was provided as a small Python script, and could as such easily be rewritten using a more effective approach such as parallel processing, less RAM usage and possibly a faster programming language.

The Beagle program is an already compile jar-file, and as such a black box to the user. It could be seen that the refined IBD algorithm was run in threads, but only one processor core was utilized. As the source code was closed, its efficiency could not be improved on by altering the code to use more cores. The use of Java could be questioned for such a computationally intense algorithm, as it lacks the speed and efficient memory use of languages such as C. Beagle 4.0 tries to decrease the RAM used by frequently writing to temporary files. Project I was partly run on a high performance computer center located at the Royal Institute of Technology in Stockholm. The center was initially concerned that the extensive amount of writing to file would put too much strain on the infrastructure. A mounted space of RAM memory was therefore used for the temporary files. This is not a sustainable way of running an analysis and is not feasible for larger data or parameter

sweeping, and more efficient memory handling should be given attention in future versions of Beagle.

5.2 PAPER II

Paper II is mostly using the liability threshold model as underlying model.

In the heritability analysis, the amount of shared genetics and environment are modeled with equations for different relations. Using these, an equation systems are set up, one for each relation where the amount of genetics and environment shared for each relation is specified. Genetics (A), shared environment (C), and non-shared environment (E) should together explain the whole variation.

By using concordance data on MZ and DZ twins, the equation system can be solved and the proportions of A, C and E respectively can be estimated. To further extend the sample and give more power to the analysis, it was decided to add siblings and half-siblings in the analysis. With this came further assumptions about the data, such as: full siblings share household environment, maternal half-siblings are reared together, and paternal half-siblings are reared apart. A report from Sweden supports the assumption of paternal half-siblings being reared apart [87], but this assumption can of course be questioned and discussed. To decrease variance that might be caused due to different upbringing and exposure to different environmental factors, the analysis included only the oldest sibling pair from each family, and constrained the age difference to a maximum of five years between the siblings.

The equations assume an underlying normally distributed liability with a threshold needed to be crossed in order to develop disease, and as sex was included in the model as a covariate, the liability threshold model with sex dimorphisms as proposed by Carter (see section 1.4.1) was used.

In this model of heritability estimations, interactions are not taken into account. This could potentially overestimate the genetic part of the equation if they are AxE-interactions, and the C component if they are AxC interactions [88]. The C component turned out as non-significant. Most environmental factors associated to MS so far are part of the E-component, as factors like smoking and sun exposure are things an individual is exposed to. There are known gene-environment interactions in MS [23], and assuming the above statements of overestimation of components in the presence of interactions are true, this

would imply that the A component could potentially be slightly over estimated and needs to be interpreted with a bit of caution.

The non-significant shared environment estimate is very in line with previous findings and the above statement of overestimating C in presence of AxC-interactions further supports a low contribution of shared environment. However, an Italian group disagreed to this conclusion and sent in a "letter to the editor" regarding Paper II. They disagreed with the conclusion that our findings were well in line with previous studies, and claimed that their study differed with a significant shared environmental component. Upon inspection of their original study [89], it was revealed that there were no significant differences between the studies, as their CI for the shared environment included 0 and most of their CI's not only overlapped, but also fully included our intervals. Our full authors reply to the letter can be found in Appendix A.

The tetrachoric correlations analysis assumes two underlying normally distributed variables with a binary outcome and a threshold cut-off for the dichotomization. In Paper II, the correlations were reported as online supplementary material, and not much discussed in the paper. When looking at the estimated correlations, they confirm the result from the Cox regression: MZ twins have the highest correlation and DZ twins have a non-significant correlation. Furthermore, all first degree relations are significant, while most second degree relations and cousins have confidence intervals including 0. The few CI's for the more distant relations that do not overlap 0 are however, not far from 0. Due to multiple testing issues and no formal testing of their significance, they should be interpreted with caution.

In the study, all individuals with a Swedish personal identity number were included, i.e. including individuals with a non-Nordic origin. This introduces the assumption that the distribution of the genetic liability load is the same for all ethnic populations, and that difference in prevalence between countries would be attributable to an environmental factor acting on the threshold and present to different degrees. A report from Sweden showed a prevalence among immigrants in the same range as the general population in Sweden, but significantly higher than the country of origin [90]. Based on this report, the assumption of equal genetic liability load seems appropriate.

All MS patients were analyzed jointly in this project. Unfortunately we did not have access to clinical course data, but estimation of the familial risks for the different courses types would be of interest to investigate if course type aggregates within families, or if a particular course type would provide a higher familial risk. When analyzing as one cohort,

it is assumed, like the example above with ethnicity, that the genetic liability distribution is equal in all course groups.

Another possible concern would be if wrong fatherhood were frequently reported in the records. In Sweden, if a couple is married, the husband is automatically registered as the father. If the couple is not married, the father has to admit the paternity in court. The automatic registration of paternity could of course introduce uncertainty in the data. The problem with wrong fathers reported is also a problem likely to exist in other data collections.

5.2.1 Findings and implications

Paper II found lower familial relatives risks than previously reported. This implies less strength of the genetic contribution than previously thought.

We could not replicate the previously reported risk differences between maternal and paternal relatives. This may imply a difference between the populations. Another explanation could be that the differences between studies are attributable to methodological differences. Studies interviewing patients in a clinic would tend to give higher rates of females and concordant pairs according to existing literature [34],[91]. As other registry studies of familial risks in MS show less differences between the sexes [70], this could imply a possible bias in some of the previous studies.

Assuming the LTSD model, and taking the lack of significant differences in transmission to child into account, it seems like men do not need to overcome a higher threshold to develop MS. Alternatively the effect could be too small to be shown in the present study.

The non-significant shared environment gives implications on where to further study for environmental risk factors and thus potentially important clues to the MS etiology.

5.2.2 Conclusions and future perspectives

In conclusion, when using registry data, biases that may arise from data collected during patient interviews in the clinic are avoided. By using randomly selected controls, assumptions about the risk for the general population did not have to be made. A lower risk for family members, with equal risks between the sexes, was found. This could imply that risks may have been overestimated in some of the previous studies. The proposed higher genetic load in men could not be confirmed.

The genetic liability load between Swedes and non-Swedes is of interest to investigate further, but perhaps more interesting would be to investigate if familial risks differs between the clinical course types.

The equal risks between the sexes is contradictory to both previous studies of familial risks, and parent-of-origin effects. A parent-of-origin-effects would however imply the SC-model as underlying, and it is tempting to speculate that modeling familial risks using a SC framework would give different risk estimates.

5.3 PAPER III

Paper III can be argued to use both SC and the LTSD as models. If viewed from the SC angle, this means female sex would be a piece of one of the pies. That pie would also include an environmental factor and it seems like this yet unknown factor has increased throughout the 21st century and the number of patients getting the disease because of this pie has thus increased. Alternatively, several pies including female sex and the environmental factors could be causing disease.

When viewed from the LTSD framework, the results could be interpreted as some environmental factor, present in the way described above, lowers the threshold for women, making them more susceptible to disease. Another interpretation could be that the liability distribution amongst women has gradually shifted towards more women with a higher load of liability.

This is a study of prevalence in different birth cohorts. The data contain no information as to when the oldest cohort got the MS diagnosis, and the registries does not permit us to estimate an age at onset. If onset date had been available, it could be utilized to speculate as to when in time the suggested environmental factor started increasing. Paper III also assumes there were no differences in age at onset distribution between males and females. Upon stratification of the age at onset distribution on gender, no such differences were detectable (data not shown).

An alternative way of approaching the problem would be to see the data as a truncated dataset. An individual is born and will some time die. During this period, an event could possibly happen. There will also be a possibility that the individual is lost to follow up, called censoring. The censoring or death could happen before or after the disease event. The possibility that censoring or death could happen before the disease event introduces

some uncertainty, and thus, modeling with a time dependent model, such as the Cox regression used in Project II, would be interesting.

5.3.1 Findings and implications

In this paper, an increasing sex-ratio was found in Sweden, confirming a global trend with a rapid increase for MS prevalence in women. This implies something in the environment is involved, as a genetic factor would take longer time to show such a large effect.

A sampling bias in a register due to men dying at a higher rate than women was also found. The healthcare data analyzed in the paper implied no systematic structural differences in age from onset to diagnosis between the sexes. An improvement with shorter time to health care between the birth cohorts was found. This decreasing time to health care part is partly confounded by the relatively late introduction to PAR for the oldest birth cohorts, but even excluding the oldest cohorts, the time to diagnosis is decreasing.

The increase in age at giving birth to the first child did not differ between the cases and the general population, implicating this is probably not a cause for the increasing proportion of women with MS.

5.3.2 Conclusions and future perspective

The women-to-men ratio for MS does increase in Sweden. The difference in results between the studies from Sweden could be attributed to a sampling bias in the previous study, due to men dying faster than women.

Future investigation of what might be causing the increasing women-to-men ratio needs to be made. A factor that seems common to western lifestyle is suggested by many researchers, like smoking, which could unfortunately not be looked into in this project.

The higher rate of mortality in men could be further investigated and compared to the general population.

In this project it is assumed that the increase is driven by relapsing remitting MS because the proportion of individuals with primary progressive disease is so small. Again access to the data needed was unfortunately not available to investigate this, but had the data been accessible it would be very interesting to look into the sex proportions.

5.4 PAPER IV

In Paper IV the SC model was used as the underlying model, or rather a probabilistic version estimating the probability that a certain mechanism caused disease in a certain individual. Moving away from the deterministic approach suggested by Rothman allows estimation of the underlying mechanisms even with more uncertainty in the data. As not all risk factors are yet identified for MS, a model incorporating uncertainty is beneficial.

The model proposed here could quite easily be extended into a leaky noisy-OR. The leaky parameter could be interpreted as accounting for "subgroups not yet identified" or as a measure of uncertainty in the model.

An advantage with the CEM algorithm is that is gives the opportunity to investigate the risks and interaction without assuming a linear dependence, as is done in for example, logistic regression. Although the formulas for models might look similar, the CEM is not constrained to a linearly dependent solution.

This model is an attempt to classify individuals according to already hypothesized mechanisms, and as such, prior knowledge of the mechanisms is needed. Data completely at random gave completely at random answers in the prediction accuracy (data not shown). The AUC of 0.89 shows high classification accuracy, however this must be compared against the prediction accuracy of other classification methods such as support vector machines.

A way of estimating confidence intervals, to achieve some sort of accuracy estimate for the resulting parameters, would be to perform a bootstrap analysis and, using the results from the bootstrap, estimate the standard deviation to use in the calculation of CI:s. This could be easily done in R.

The framework proposed here does not necessarily rule out use of the LT model in the project. If "high genetic load" is considered a piece of one pie, it would be possible to incorporate the LT model.

The OR expression is used as the logic gate, but extending the model to allow other types of gates could be another way to incorporate the LT model. Allowing too many combination of factors will, however, very quickly result in a computationally too complex problem.

5.4.1 Findings and implications

Project IV presents a promising algorithm for classification of patients into subgroups. This algorithm needs to be investigated further. The method provides measures of interaction corresponding to the ones used traditionally in epidemiology. It furthermore shows high classification accuracy and should be developed further.

5.4.2 Conclusions and future perspectives

Paper IV presents a potentially interesting model with a high classification accuracy suggested to be investigated further. This model is however not yet fully developed. As such not all of its properties are yet known and much work is still to be done. It offers however a theoretical framework enabling us to classify patients into subgroups based on their genotypic data. These subgroups must be hypothesized *a priori* and such an identification of subgroups is outside the scope of Paper IV.

In the future, it should be further investigated how the model behaves so that results from more complex data sets can be interpreted.

The algorithm proposed here could fairly easily be extended to a leaky noisy-OR gate, to account for yet unidentified factors. Furthermore, Bayesian inference could be used to model the underlying genetic architecture of the disease [92], and the result from the analysis incorporated in the model. The model can also be developed further to allow other possible logic gates such as AND/AND-OR/etcetera, to test hypothesis about etiological subgroups.

As not only genetic factors are associated to MS, environmental factors, such as smoking, can be used. Allowing more complex models such as the "genetic risk score" as pieces of the pie could potentially also help explain more of the disease etiology.

5.5 GENERAL CONCLUSIONS AND FUTURE PERSPECTIVES

The work in this thesis failed to provide evidence for rare variants contributing to disease. Assuming this is true, and taking the lack of differences in transmission to offspring, and equal risks between the sexes into account, this implies the same genetic load, or a very small difference, is required for both sexes.

However, Paper III confirms the global trend of an increasing women-to-men ratio.

Viewing the above stated from an LTSD angle this would imply a different distribution of liability amongst women, with a larger proportion having the required amount of liability to cross the threshold. If liability does not have to be constrained to genes, this could mean the liability has increased due to some environmental factor that has increased in presence amongst women during the 20th century (Figure 5).

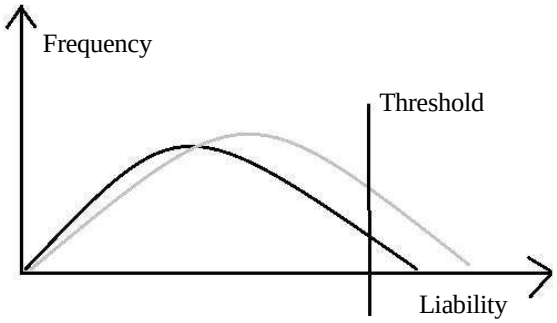

Figure 5: If the same genetic load is required for both sexes, and liability is not constrained to genetics, an environmental factor shifting the liability distribution for women (grey line), pushing more women across the threshold, could exist.

An interpretation from the SC perspective could mean that some environmental factor interacting with the female sex is becoming more and more common, increasing the frequency of that pie (or pies). A gene-environment interaction that is not necessarily stronger in effect, but more frequent and thus easier to find in women when stratifying on sex could account for this (Figure 6).

Other possible explanations of the increasing sex-ratio could of course be more awareness of the disease and better diagnostics for particularly women, but Paper III saw no differences between age an onset and age at hospitalization for the sexes. The age at onset distribution from SMSreg did not differ either between men and women.

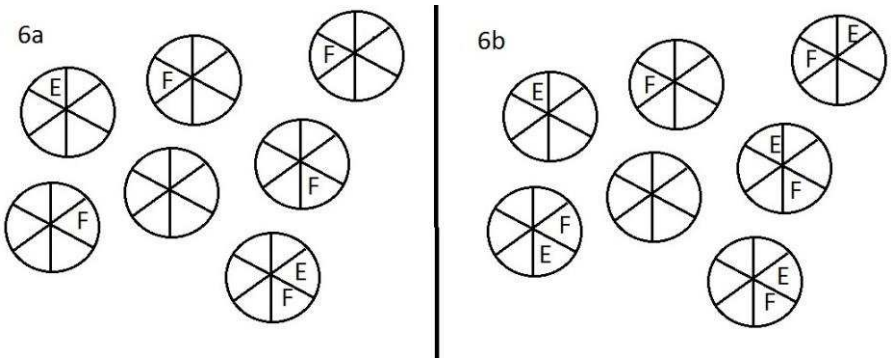

Figure 6: A possible SC-interpretation of the results. An environmental factor (E), interacting with female sex (F). As the factor is becoming more frequent, the frequency of the pie (or pies) including the factor E, has increased from figure 6a to figure 6b.

The Carter effect part of project II could be extended by taking year of birth into account and making the model more complex. By doing so, it would be possible to see if the transmission to offspring has changed over time. If it is increasing over time, this could imply some trans-generational effect, possibly epigenetic. If it is stable, this could indicate something in the environment is responsible for the increase.

5.5.1 Personal remark

This last part of the thesis is written from my personal perspective, as an engineer trained in computer science, meeting the field of medical genetics and epidemiology.

Programming is an art. It's a craft that takes years and years to develop and master. Like science, it is a creative process that requires reflection upon the work done, and it is a way of thinking that is acquired through experience of the learned theoretical framework.

Programming is not just about writing code that performs what you expect of it, it's also about designing your algorithm, choosing a proper language, appropriate data types and designing structs, objects and classes suitable for the problem and the language. Much like choosing study design, analysis methods and statistics appropriate for the scientific study.

Programming builds on knowledge about computer architecture, where the program will be used and by whom, and knowing how to adapt to those conditions. It is not about

writing many lines of code, but code that efficiently does the job without being over optimized. Programming is about knowing how to properly validate all the small parts that together make up the program, and for validation purposes, it is essential that the code is released as open source so other researchers can validate, understand what is going on and help develop further.

Much like science is about knowing what has been done in the field and strength and weaknesses of different studies and methods, setting up an hypothesis that can be tested in a reliable way and publishing both positive and negative results in a way that makes it reproducible and gives the full picture, will avoid publication bias.

With the increasing amount of data generated in the medical sciences that are analyzed in more and more complex ways, complexity of the analysis is an issue that must be given more attention. Spreadsheets are since some years back no longer sufficient to even open the data files, less so calculate statistics. There is a general interest and willingness among many researchers in the field to go towards more advanced statistical tools, such as R, but for more advanced problems R is not enough.

I believe the field of medical science as a whole, and genetics in particular, would benefit from basic knowledge about how to handle data in *nix systems and master basics concepts about programming. If nothing else but to give scientists tools to write simple scripts to test a new idea or concept. But only knowledge of basic programming is not enough. The infrastructure must be supportive and small computational clusters that are easy to access and use would be a step in this direction. Having the tools to easily test a new idea would further encourage creativity amongst researchers.

To be prepared for the future of more complex analysis, investments must be made in hardware capable of fast and efficient processing of the data as well as education of researchers. But also inter-disciplinary collaborations involving more programming expertise in combination with statistical knowledge is essential to develop novel, reliable, efficient methods for analyzing large sets of data in complex ways.

I have been fortunate to work in an encouraging environment willing to invest in hardware and send me to courses to further develop my programming skills. I have also been fortunate to get to know people passionate about programming and computational problems and together we have created a network to help and support each other. It is my hope that this collaboration will continue and expand to a level where it can be a resource to Karolinska Insitutet and contribute to solving the above raised concerns.

6 ACKNOWLEDGEMENTS

First and foremost I would like to thank my main supervisor professor **Jan Hillert,** for having the courage to take on a computer scientist for PhD studies in medical sciences, trusting me to know my things, and for teaching me so much about science and always helping out when needed.

I would also like to thank my co-supervisor professor **Timo Koski** for taking me under his wings, for his immense effort in paper IV and for all our talks about life, academia and life in academia. Thanks to my co-supervisor **Ryan Ramanujam**, for always taking the time to help and always having a solution to the problem up your sleeve. Also big thank you to my two co-supervisors **Ingrid Kockum** and **Iza Lima** for helping me out, and a thank you to my mentor **Peter Arner** for all valuable advice on life as a researcher and on my thesis.

Without **KIRC**, this wouldn't have been doable. Thank you **Daniel Uvehag** for sorting everything technical out, **Henrik Källberg** for always helping out and all our discussions and your enthusiasm, I'm really looking forward to working with you soon! Thank you **Robert Karlsson** for your always brilliant input in everything we've done, **Jonas Forsslund** for all our studies together and for introducing me to the brilliant entrepreneur **Johan Acevedo**, without whom there would be no KIRC, **Andreas Gillberg** for designing the front page for my thesis, and to **Boel Brynedal** an extra special thank you for choosing my application to the student position in the Hillert-group.

I also owe all my co-authors a big thank you. **Kerstin Imrell** for starting the segmental sharing project, **Jim Stankovich** for all your help literally all the way from the other side of the earth, and all other **co-authors** for commenting on the manuscript. The familial risk project wouldn't have been if professor **Paul Lichtenstein** hadn't proposed the project and so generously allowed us to use both his method and material. Thank you, and thank you **Ralf Kuja-Halkola** for teaching me about twin studies, and **Marcus Boman** and **Christina Norrby** for your help with the Crime database. I also owe professor **Matteo Bottai** a big thank you for making me part of his Biostat community and always taking time to share his immense knowledge of statistics.

Thank you **Inger Boström**, for all our talks on MS epidemiology, **Leszek Stawiarz**, for always taking time to chat about science and sharing your knowledge, and **Anne-Marie Landtblom** and **Catarina Almqvist** for all your comments and help with the sex-ratio paper.

Thank you all past and present colleagues in the Hillert research group: **Christina Hermanrud** for all our talks, **Anna Fogdell Hahn** good luck with your very own group, **Anna Glaser**, **Jenny Link**, **Sahl Bedri**, **Andrius Kavaliunas**, **Wangko Lundström**, **Rasmus Gustafsson**, **Malin Ryner**, **Elin Engdahl**, **Ingegerd Löfving Arvholm**, **Roger Ljungedal** and **Anna Mattsson**. An extra thank you to **Ali Manouchehrinia** for reading and commenting on my thesis. Thank you all our brilliant clinicians working both as researchers and seeing patients: **Virginija Karrenbauer**, **Katharina Fink**, **Tomas Masterman**.

Thank you all past and present **colleagues** and **co-workers** all over CMM and W5 for all the fikas and all the chats during lunch.

Thank you everyone involved in the administrative work, **Olle Gartell** at CMM IT for turning the ship around, and **Elin Johansson** and **Gullan Rydén**, "my" administrators. **Linda**, **Nina**, **Carolina**, **Karin** and **Karin** and everyone else involved in the collection of all MS patients. **Marjan**, **Merja**, **Cecilia** and all the **nurses**, all the **clinicians** and of course all the **patients** at the clinic.

Thank you to everyone in the **Biostat group** for all your talks and discussions on statistics, I've learned so much from you guys!

Thanks to all my friends outside work: **Anette**, **Jenny**, **Jessica**, **Anneli**, **AnnaKarin**, **Malin**, **Gunilla**, **Joakim** and **Andrea** who keep my feet on the ground and head away from working all the time. Thanks to all the dog breeders that have let me practice my knowledge of genetics in your practical work and to all my friends at past and present dog training clubs for all the fun we've had at shows and at trainings. Thank you **Carina** for providing the best of care to my dogs when I'm at work.

Thanks to my mum **Lilian Hamilton** for all the help and support, taking care of the dogs when needed and proofreading my thesis, and thank you to my **grandparents** and the rest of my **family** for all your support and encouragement.

Last of all: thank you **Bubba**, **Dennis** and **Sheldon** for always being the best of friends and making sure I get exercise and have fun outside work.

7 REFERENCES

1. **Compston A, Coles A**. Multiple sclerosis. *Lancet*. 2008; **372**(9648):1502–17.

2. **Confavreux C, Vukusic S**. The clinical course of multiple sclerosis. *Handb. Clin. Neurol.* 2014; **122**:343–69.

3. **McDonald WI, Compston A, Edan G, *et al.*** Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. *Ann. Neurol.* 2001; **50**(1):121–7.

4. **Confavreux C, Vukusic S**. Age at disability milestones in multiple sclerosis. *Brain*. 2006; **129**(Pt 3):595–605.

5. **Cross AH, Naismith RT**. Established and novel disease-modifying treatments in multiple sclerosis. *J. Intern. Med.* 2014.

6. **Berg-Hansen P, Moen S, Harbo H, Celius E**. High prevalence and no latitude gradient of multiple sclerosis in Norway. *Mult. Scler.* 2014.

7. **Koch-Henriksen N, Sørensen PS**. The changing demographic pattern of multiple sclerosis epidemiology. *Lancet Neurol.* 2010; **9**(5):520–32.

8. **Ahlgren C, Odén A, Lycke J**. High nationwide prevalence of multiple sclerosis in Sweden. *Mult Scler*. 2011; **17**(8):901–908.

9. **Orton S-M, Herrera BM, Yee IM, *et al.*** Sex ratio of multiple sclerosis in Canada: a longitudinal study. *Lancet Neurol.* 2006; **5**(11):932–6.

10. **Willer CJ, Dyment DA, NJ R, Sadovnick AD, Ebers GC an CCSG**. Twin concordance and silbling recurrence rates in multiple sclerosis. 2003; **100**(22):12877 – 82.

11. **O'Gorman C, Lin R, Stankovich J, Broadley SA**. Modelling genetic susceptibility to multiple sclerosis with family data. *Neuroepidemiology*. 2013; **40**(1):1–12.

12. **Bush WS, Sawcer SJ, de Jager PL, *et al.*** Evidence for polygenic susceptibility to multiple sclerosis--the shape of things to come. *Am. J. Hum. Genet.* 2010; **86**(4):621–5.

13. **Jersild C, Hansen G, Svejgaard A, *et al.*** HISTOCOMPATIBILITY DETERMINANTS IN MULTIPLE SCLEROSIS, WITH SPECIAL REFERENCE TO CLINICAL COURSE. *Lancet*. 1973; **302**(7840):1221–1225.

14. **Stys PK, Zamponi GW, van Minnen J, Geurts JJG**. Will the real multiple sclerosis please stand up? *Nat. Rev. Neurosci.* 2012; **13**(7):507–14.

15. **Fogdell-Hahn A, Ligers A, Grønning M, Hillert J, Olerup O**. Multiple sclerosis: a modifying influence of HLA class I genes in an HLA class II associated autoimmune disease. *Tissue Antigens*. 2000; **55**(2):140–8.

16. **Lundmark F, Duvefelt K, Iacobaeus E, *et al.*** Variation in interleukin 7 receptor alpha chain (IL7R) influences risk of multiple sclerosis. *Nat. Genet.* 2007; **39**(9):1108–13.

17. **International Multiple Sclerosis Genetics Consortium(IMSGC), Beecham AH, Patsopoulos NA, *et al.*** Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat. Genet.* 2013; **Advance on**(http://dx.doi.org/10.1038/ng.2770 L3).

18. **Sawcer S, Hellenthal G, Pirinen M, *et al.*** Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*. 2011; **476**(7359):214–9.

19. **Hedström AK, Bäärnhielm M, Olsson T, Alfredsson L**. Tobacco smoking, but not Swedish snuff use, increases the risk of multiple sclerosis. *Neurology*. 2009; **73**(9):696–701.

20. **Bäärnhielm M, Hedström AK, Kockum I, *et al.*** Sunlight is associated with decreased multiple sclerosis risk: no interaction with human leukocyte antigen-DRB1*15. *Eur. J. Neurol.* 2012; **19**(7):955–62.

21. **Ascherio A, Munger KL, Simon KC**. Vitamin D and multiple sclerosis. *Lancet Neurol.* 2010; **9**(6):599–612.

22. **Hedström AK, Olsson T, Alfredsson L**. High body mass index before age 20 is associated with increased risk for multiple sclerosis in both men and women. *Mult. Scler.* 2012; **18**(9):1334–6.

23. **Hedström AK, Sundqvist E, Bäärnhielm M, *et al.*** Smoking and two human leukocyte antigen genes interact to increase the risk for multiple sclerosis. *Brain*. 2011; **134**(Pt 3):653–64.

24. **Serafini B, Rosicarelli B, Franciotta D, *et al.*** Dysregulated Epstein-Barr virus infection in the multiple sclerosis brain. *J. Exp. Med.* 2007; **204**(12):2899–912.

25. **Willis SN, Stadelmann C, Rodig SJ, *et al.*** Epstein-Barr virus infection is not a characteristic feature of multiple sclerosis brain. *Brain.* 2009; **132**(Pt 12):3318–28.

26. **Salvetti M, Giovannoni G, Aloisi F**. Epstein-Barr virus and multiple sclerosis. *Curr. Opin. Neurol.* 2009; **22**(3):201–6.

27. **Do R, Kathiresan S, Abecasis GR**. Exome sequencing and complex disease: practical aspects of rare variant association studies. *Hum. Mol. Genet.* 2012; **21**(R1):R1–9.

28. **Auton A, McVean G**. Estimating recombination rates from genetic variation in humans. *Methods Mol. Biol.* 2012; **856**:217–37.

29. **Reich DE, Cargill M, Bolk S, *et al.*** Linkage disequilibrium in the human genome. *Nature*. 2001; **411**(6834):199–204.

30. **Price AL, Patterson NJ, Plenge RM, *et al.*** Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 2006; **38**(8):904–9.

31. **Bradford Hill A**. The Environment and Disease: Association or Causation? *Proc. R. Soc. Med.* 1965:295–300. Available at: http://www.edwardtufte.com/tufte/hill? [Accessed April 17, 2014].

32. **Pearl J**. *Causality: Models, Reasoning and Inference*. second edi. New York: Cambridge University Press; 2009.

33. **Rothman KJ, Greenland S, Lash TL**. *Modern epidemiology*. Wolters Kluwer Health; 2008.

34. **Lykken DT, McGue M, Tellegen A**. Recruitment bias in twin research: the rule of two-thirds reconsidered. *Behav. Genet.* 1987; **17**(4):343–62.

35. **Carter CO**. The inheritance of congenital Pyloric Stenosis. *Br Med Bull*. 1961; **17**:251–254.

36. **Farrall M, Green FR, Peden JF, *et al.*** Genome-wide mapping of susceptibility to coronary artery disease identifies a novel replicated locus on chromosome 17. *PLoS Genet.* 2006; **2**(5):e72.

37. **Broadbent HM, Peden JF, Lorkowski S, *et al.*** Susceptibility to coronary artery disease and diabetes is encoded by distinct, tightly linked SNPs in the ANRIL locus on chromosome 9p. *Hum. Mol. Genet.* 2008; **17**(6):806–14.

38. **Wedrén S, Lovmar L, Humphreys K, *et al.*** Oestrogen receptor alpha gene haplotype and postmenopausal breast cancer risk: a case control study. *Breast Cancer Res.* 2004; **6**(4):R437–49.

39. **Purcell S, Neale B, Todd-Brown K, *et al.*** PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 2007; **81**(3):559–75.

40. **R Core Team**. R: A Language and Environment for Statistical Computing. 2013.

41. **MATLAB**. *version 7.9.1*. Natick, Massachusetts: The MathWorks Inc.; 2010.

42. **Ludvigsson JF, Andersson E, Ekbom A, *et al.*** External review and validation of the Swedish national inpatient register. *BMC Public Health*. 2011; **11**:450.

43. **Magnusson PKE, Almqvist C, Rahman I, *et al.*** The Swedish twin registry: establishment of a biobank and other recent developments. *Twin Res. Hum. Genet.* 2013; **16**(1):317–29.

44. **Blom G, Enger J, Englund G, Grandell J, Holst L**. *Sannolikhetsteori och statistikteori med tillämpningar*. 5th ed. Lund: Studentliteratur; 2005.

45. **Browning SR, Thompson EA**. Detecting rare variant associations by identity-by-descent mapping in case-control studies. *Genetics*. 2012; **190**(4):1521–31.

46. **Greenland S, Thomas DC**. On the need for the rare disease assumption in case-control studies. *Am. J. Epidemiol.* 1982; **116**(3):547–53.

47. **Tenesa A, Haley CS**. The heritability of human disease: estimation, uses and abuses. *Nat. Rev. Genet.* 2013; **14**(2):139–49.

48. **Browning BL, Browning SR**. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*. 2013; **194**(2):459–71.

49. **Browning BL, Browning SR**. A fast, powerful method for detecting identity by descent. *Am. J. Hum. Genet.* 2011; **88**(2):173–82.

50. **Risch N**. Estimating morbidity risks with variable age of onset: review of methods and a maximum likelihood approach. *Biometrics*. 1983; **39**(4):929–39.

51. **Mills M**. *Intoducing survival and event history analysis*. London: Sage; 2011.

52. **Aalen O, Borgan Ö, H G**. *Survival and Event History Analysis*. Softcover . New York: Springer-Verlag; 2008.

53. **Boker S, Neale M, Maes H, *et al.*** OpenMx: An Open Source Extended Structural Equation Modeling Framework. *Psychometrika*. 2011; **76**(2):306–317.

54. **Pearl J**. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann; 1988.

55. **Boström I, Callander M, Kurtzke JF, Landtblom A-M**. High prevalence of multiple sclerosis in the Swedish county of Värmland. *Mult. Scler.* 2009; **15**(11):1253–62.

56. **Binzer M, Forsgren L, Holmgren G, Drugge U, Fredrikson S**. Familial clustering of multiple sclerosis in a northern Swedish rural district. *J. Neurol. Neurosurg. Psychiatry*. 1994; **57**(4):497–9.

57. **Pihlaja H, Rantamäki T, Wikström J, *et al.*** Linkage disequilibrium between the MBP tetranucleotide repeat and multiple sclerosis is restricted to a geographically defined subpopulation in Finland. *Genes Immun*. 2003; **4**(2):138–46.

58. **Aulchenko YS, Hoppenbrouwers IA, Ramagopalan S V, *et al.*** Genetic variation in the KIF1B locus influences susceptibility to multiple sclerosis. *Nat. Genet.* 2008; **40**(12):1402–3.

59. **Koski T**. *Hidden Markov Models for Bioinformatics*. Dordrecht, The Netherlands: Kluwer Academic Publishers; 2001.

60. **Gusev A, Lowe JK, Stoffel M, *et al.*** Whole population, genome-wide mapping of hidden relatedness. *Genome Res.* 2009; **19**(2):318–26.

61. **Ebers GC, Sadovnick AD, Risch NJ**. A genetic basis for familial aggregation in multiple sclerosis. Canadian Collaborative Study Group. *Nature*. 1995; **377**(6545):150–1.

62. **Sadovnick AD, Baird PA**. The familial nature of multiple sclerosis: age-corrected empiric recurrence risks for children and siblings of patients. *Neurology*. 1988; **38**(6):990–1.

63. **Sadovnick AD, Risch NJ, Eberg GC**. Canadian collaborative project on genetic susceptibility to MS, phase 2: rationale and method. Canadian Collaborative Study Group. *Can J Neurol Sci*. 1998; **25**(3):216–221.

64. **Ebers GC, Sadovnick AD, Dyment DA, *et al.*** Parent-of-origin effect in multiple sclerosis: observations in half-siblings. *Lancet*. 2004; **363**(9423):1773–4.

65. **Herrera BM, Ramagopalan S V, Lincoln MR, *et al.*** Parent-of-origin effects in MS: observations from avuncular pairs. *Neurology*. 2008; **71**(11):799–803.

66. **Ramagopalan S V, Yee IM, Dyment DA, *et al.*** Parent-of-origin effect in multiple sclerosis: observations from interracial matings. *Neurology*. 2009; **73**(8):602–605.

67. **Hader WJ, Yee IM**. The prevalence of familial multiple sclerosis in saskatoon, Saskatchewan. *Mult. Scler. Int.* 2014; **2014**:545080.

68. **Robertson NP, Fraser M, Deans J, *et al.*** Age-adjusted recurrence risks for relatives of patients with multiple sclerosis. *Brain*. 1996; **119**(2):449–55.

69. **Carton H, Vlietinck R, Debruyne J, *et al.*** Risks of multiple sclerosis in relatives of patients in Flanders, Belgium. *J. Neurol. Neurosurg. Psychiatry*. 1997; **62**(4):329–33.

70. **Nielsen NM, Westergaard T, Rostgaard K, *et al.*** Familial risk of multiple sclerosis: a nationwide cohort study. *Am. J. Epidemiol.* 2005; **162**(8):774–8.

71. **Hemminki K, Li X, Sundquist J, Hillert J, Sundquist K**. Risk for multiple sclerosis in relatives and spouses of patients diagnosed with autoimmune and related conditions. *Neurogenetics*. 2009; **10**(1):5–11.

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

73. **Kantarci OH, Barcellos LF, Atkinson EJ, *et al.*** Men transmit MS more often to their children vs women: the Carter effect. *Neurology*. 2006; **67**(2):305–10.

74. **Herrera BM, Ramagopalan S V, Orton S, *et al.*** Parental transmission of MS in a population-based Canadian cohort. *Neurology*. 2007; **69**(12):1208–12.

75. **Kampman MT, Aarseth JH, Grytten N, *et al.*** Sex ratio of multiple sclerosis in persons born from 1930 to 1979 and its relation to latitude in Norway. *J. Neurol.* 2013; **260**(6):1481–8.

76. **Boström I, Stawiarz L, Landtblom A-M**. Sex ratio of multiple sclerosis in the National Swedish MS Register (SMSreg). *Mult. Scler.* 2013; **19**(1):46–52.

77. **Palacios N, Alonso A, Brønnum-Hansen H, Ascherio A**. Smoking and increased risk of multiple sclerosis: parallel trends in the sex ratio reinforce the evidence. *Ann. Epidemiol.* 2011; **21**(7):536–42.

78. **Magyari M, Koch-Henriksen N, Pfleger CC, Sørensen PS**. Reproduction and the risk of multiple sclerosis. *Mult. Scler.* 2013; **19**(12):1604–9.

79. **Nielsen NM, Jørgensen KT, Stenager E, *et al.*** Reproductive history and risk of multiple sclerosis. *Epidemiology*. 2011; **22**(4):546–52.

80. **Khuri AI**. *Advanced calculus with applications in statistics*. John Wiley & Sons; 2003.

81. **Sing T, Sander O, Beerenwinkel N, Lengauer T**. ROCR: visualizing classifier performance in R. *Bioinformatics*. 2005; **21**(20):7881.

82. **Watson CT, Disanto G, Breden F, Giovannoni G, Ramagopalan S V.** Estimating the proportion of variation in susceptibility to multiple sclerosis captured by common SNPs. *Sci. Rep.* 2012; **2**(Table 1):8–11.

83. **Gusev A, Bhatia G, Zaitlen N, *et al.*** Quantifying missing heritability at known GWAS loci. *PLoS Genet.* 2013; **9**(12):e1003993.

84. **Gauvin H, Moreau C, Lefebvre J-F, *et al.*** Genome-wide patterns of identity-by-descent sharing in the French Canadian founder population. *Eur. J. Hum. Genet.* 2013.

85. **Browning SR, Browning BL**. Identity-by-descent-based heritability analysis in the Northern Finland Birth Cohort. *Hum. Genet.* 2013; **132**(2):129–38.

86. **Watterson GA, Guess HA**. Is the most frequent allele the oldest? *Theor. Popul. Biol.* 1977; **11**(2):141–60.

87. **Statistics Sweden**. *Fakta om den svenska familjen*.; 1994.

88. **Verweij KJH, Mosing MA, Zietsch BP, Medland SE**. Estimating heritability from twin studies. *Methods Mol. Biol.* 2012; **850**:151–70.

89. **Ristori G, Cannoni S, Stazi MA, *et al.*** Multiple sclerosis in twins from continental Italy and Sardinia: a nationwide study. *Ann. Neurol.* 2006; **59**(1):27–34.

90. **Ahlgren C, Odén A, Lycke J**. A nationwide survey of the prevalence of multiple sclerosis in immigrant populations of Sweden. *Mult. Scler.* 2012; **18**(8):1099–107.

91. **Hawkes CH**. Twin studies in medicine--what do they tell us? *QJM*. 1997; **90**(5):311–21.

92. **Stahl EA, Wegmann D, Trynka G, *et al.*** Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat. Genet.* 2012; **44**(5):483–9.