Mathematical Statistics
Stockholm University

A New Method Applied to Gene Mapping

Bei Yang

Examensarbete 2009:1

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Examensarbete **2009:1**,
http://www.math.su.se/matstat

A New Method Applied to Gene Mapping

Bei Yang*

February 2009

Abstract

In gene mapping of complex traits, classical association approaches, including the standard chi-squared statistics or logistic regression methods, have been used to find susceptibility genes with modest effects. A novel statistical method, recursive partitioning, has recently been introduced in association studies. We use this new method to assess association between the human leukocyte antigen system (HLA) and an autoimmune disease, multiple sclerosis (MS). In particular, we model the association between HLA class II loci and MS using recursive partitioning and then model the association between HLA class I loci and MS, controlling for class II loci, using logistic regression. We have access to genotype data on 3174 MS patients and healthy controls from the Swedish and Norwegian populations. Our results differ slightly from previous studies that use logistic regression exclusively on the same data.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: beiyang 2000@hotmail.com Supervisor: Juni Palmgren

Acknowledgments

I am grateful to my supervisor, Professor Juni Palmgren, for her designing the project and guiding me to apply the new statistic method to gene mapping in my master thesis. Her encouragement helped me reaching my goal.

I thank also Professor Jan Hillert and his colleagues Jenny Link and Boel Brynedal in Huddinge hospital for providing data sets, fruitful discussion and enthusiasm of the collaboration.

Contents

1 Background

1.1 HLA-region

1.1.1 The HLA system and autoimmune diseases

The human leukocyte antigen system (HLA) is the name of the major histocompatibility complex (MHC) in humans. The major histocompatibility complex (MHC) is a large genomic region or gene family found in most vertebrates. It is the most gene-dense region of the mammalian genome and plays an important role in the immune system, autoimmunity, and reproductive success. The human MHC is called the HLA (Human Leukocyte Antigen) system because antigens were first identified and characterized using alloantibodies against leukocytes.

The HLA system has been well known as transplantation antigens, encode cell-surface antigen-presenting proteins and many other genes. But the primary biological role of HLA molecules is the regulation of immune response.

Autoimmune diseases arise from an overactive immune response of the body against substances and tissues normally present in the body. In other words, the body attacks its own cells. Autoimmune diseases are a major cause of immune-mediated diseases, and are commonly referred to as Autoimmune and Inflammatory Diseases (AIID). Women tend to be affected more often by autoimmune disorders; nearly 79% of autoimmune disease patients in the USA are women. Also they tend to appear during or shortly after puberty. It is not known why this is the case, although hormone levels have been shown to affect the severity of some autoimmune diseases such as multiple sclerosis. Other causes may include the presence of fetal cells in the maternal bloodstream (wikipedia).

1.1.2 Genomic organization of the HLA system and encoding function

The human MHC map to the short arm of chromosome 6 (6p21) and spans approximately 3,600 kilobases of DNA. The human MHC is divided into three regions (Figure 1)

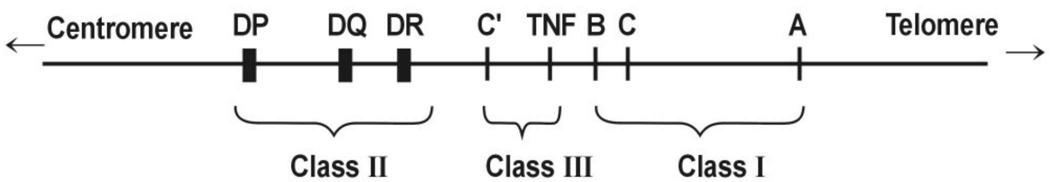


Figure1.1 The human MHC on the short arm of chromosome 6. The class I region contains A,B,C genes; The HLA-DR, DP,and DQ regions constitute class II, TNF (tumor necrossis factors), C' (complement genes) are class III genes (Sung et al. 2007).

The class I region contains the HLA-A, HLA-B and HLA-C genes which present peptides from inside the cell (including viral peptides if present). These peptides are produced from digested proteins that are broken down in the lysozomes. The peptides are generally small polymers, about 9 amino acids in length. Foreign antigens attract killer T-cells (also called CD8 positive cells) that destroy cells.

The class II region consists of a series of subregions. The DR gene family consists of a single DRA gene and up to nine DRB genes (DRB1 to DRB9). The DQA1 and DQB1 gene products associate to form DQ molecules, the DPA1 and DPB1 products from DP molecules. HLA class II antigens present antigens from outside the cell to T-lymphocytes. These particular antigens stimulate T-helper cells to reproduce and these T-helper cells then stimulate antibody producing B-cells, self-antigens are suppressed by suppressor T-cells. The class III region does not encode HLA molecules, but contains gene for complement components (C2, C4, factor B), tumor necrosis factors (TNFs), and some others.

In the HLA locus, there are also many of multiple-allele markers identified in the HLA-A and HLA-C genes in class I and the HLA-DRB1 genes in class II.

1.1.3 HLA haplotype and molecular typing of HLA alleles

Recombination

During cell meiosis, the process that leads to the formation of new gene combinations on chromosomes is called recombination.

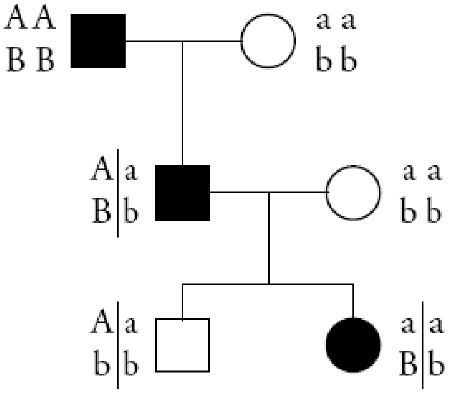


Figure 1.1 Recombination in meiosis process. The gamete Ab, aB are new gene combinations on chromosomes.

Haplotype

A sequence of alleles from different loci received from the same parents is called a haplotype. Figure 1.2 is a pedigree of HLA haplotypes, showing no recombination.

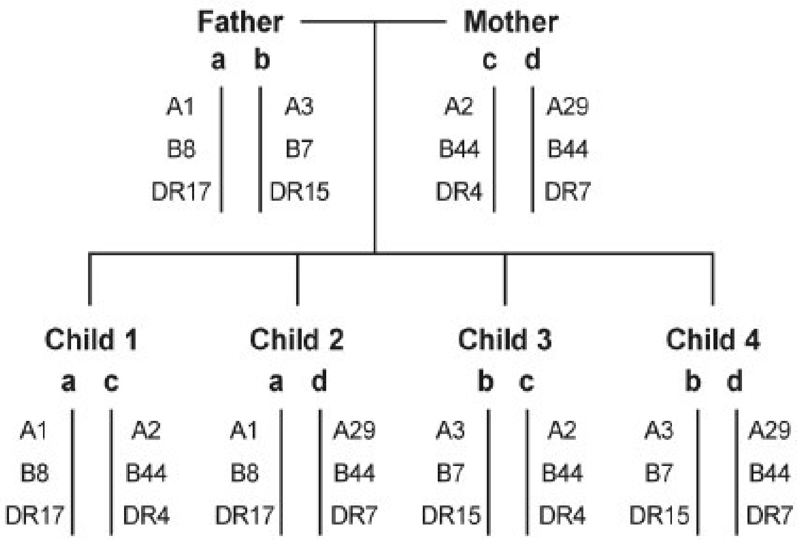


Figure 1.2 Haplotypes in the pedigree. The paternal HLA haplotypes are A1, B8, DR17 (a) and A3, B7 and DR15 (b); and the maternal HLA haplotypes are A2, B44, DR4(c) and A29, B44, DR7(d). (Sung 2007)

Linkage disequilibrium (LD)

Linkage disequilibrium is a phenomenon where alleles at two loci in a population appear at the same time more often than what would be expected by chance. The rationale for this is that nearby loci must have correlated inheritance patterns, because crossover occurs between the two loci with low probability. Genes that lie on the same chromosome tend to be inherited as a group, a tendency that declines with increasing distance between the loci. As a result, haplotype frequency may deviate from expectation based on allele frequencies, the phenomenon called linkage disequilibrium.

Linkage disequilibrium may lead to some marker alleles being over represented among affected individuals. For example, HLA-A1, B8, DR17 is the most common HLA haplotype among Caucasians, with a frequency of 5%. Haplotype of an ancient disease founder is left intact through many generations in a chromosomal region surrounding the disease locus. This theory underlies association analysis, one uses the fact that markers in close vicinity of a disease locus might be in linkage disequilibrium with the disease locus. (Almgren et al. 2003).

Molecular typing of HLA alleles

Genotyping refers to the process of determining the genotype of an individual by the use of biological assays. Current methods of doing this include PCR, DNA sequencing, ASO probes, and hybridization to DNA microarrays or beads. The technology is important in clinical research for the investigation of disease-associated genes.

SNPs

A single nucleotide polymorphism (SNP, pronounced *snip*) is a DNA sequence variation occurring when a single nucleotide - A, T, C, or G - in the genome differs between members of a species. For example, two sequenced DNA fragments from different individuals, AAGCCTA to AAGCTTA, contain a difference in a single nucleotide. In this case we say that there are two *alleles* : C and T. Almost all common SNPs have only two alleles (Wikepedia). SNPs are the most common type of genetic variation. A SNP is a single base pair mutation at a specific locus. Because SNPs are evolutionarily conserved, they have been proposed as markers for use in association studies.

1.2 MS

1.2.1 Clinical aspects of MS

Multiple sclerosis (MS) is a chronic inflammatory disease of the central nervous system (CNS), affecting the brain, optic nerves and spinal cord, resulting in relapsing or progressive demyelination in CNS. It may cause various neurological symptoms and signs. The course and severity of MS differ greatly between patients. It makes its appearance most frequently between the age of 20-40, while the onset of MS is rare before 15 and after 60 years. MS is more common in woman than in men (about 2:1). There is no diagnostic laboratory test for MS, and clinical criteria, therefore, have to be used (Poser et al.1983). However, the phenotype (clinical manifestation) is well defined in MS, particularly since the development of diagnostic techniques such as cerebrospinal fluid analysis and magnetic resonance scanning (MRI).

1.2.2 Epidemiology

Multiple sclerosis, as a modern clinical entity, was first described in the second half of the 19th century. The etiology of MS, however, is still poorly understood. Epidemiology surveys may ideally offer important information to find clues, but are methodologically difficult to carry out. In particular, it has long been debated whether genetic or environmental factors are most critical in causing MS.

MS is a disease with an uneven world-wide distribution. High prevalence (>30/100,000) is seen in northern Europe, southern Canada, northern US and some part of New Zealand and Australia, while areas of Asia and Africa consistently show low prevalence (<5/100,000). The highest prevalence rate of MS is seen in northern Europe, about 100/100,000 (Kurzke 1983). The very high rate is found in the Orkney and Shetland Islands situated to the north of the Scottish mainland with 309 and 184/100,000 respectively (Poskanzer et al. 1980). All high-risk areas are among predominantly white populations, whereas Blacks and Orientals and possibly Indians in US have much lower rates. Thus, the high-risk rate (for instance a prevalence over 50/100,000) in individuals of Northern European decent clearly makes MS a common disease in this particular population group.

1.2.3 Autoimmune mechanism in MS

The pathological characteristic of MS is inflammatory cell infiltration and focal loss of myelin sheath scattered in white matter of the CNS, while axons and nerve cell bodies are relatively preserved. Autoimmunity to the self myelin protein is the commonly assumed model for MS pathogenesis, classifying MS as an autoimmune disease. However, the alternative possibility, that these immune responses are secondary events, cannot be excluded easily. Indeed, increasing evidence supports MS as a T cell mediated autoimmune disease by the following lines:

1) Histopathological features with restricted organ-specific inflammation and selective myelin destruction in which macrophages, T cells and antibodies may be engaged (Lassman et al.1991);

2) An increase of myelin antigen-reactive T cells in peripheral blood as well as in cerebrospinal fluid (CSF), which are specific for various myelin proteins including myelin basic protein(MBP) (Olsson 1990), proteolipid protein (PLP) (Sun et al.1991a), myelin associated glycoprotein (MAG) (Link et al. 1992) and myelin oligodendrocyte glycoprotein (MOG)(Sun et al.1991b).

3) Pronounced B cell responses within the CNS, reflected in the CSF by presence of oligoclonal immunoglobulin (Ig) bands in the CNS, elevated levels of intrathecal antibody synthesis (Link et al. 1971), autoantibodies to MBP, PLP, MAG, MOG, (Möller et al.1989; Warren et al.1994; Xiao et al.1991), and further higher numbers of myelin protein-specific antibody producing B cells (Olsson et al. 1990; Sun et al.1991b).

4) Identification of the human leukocyte antigen (HLA) class II genes, immune response genes, as associated with an increased risk of MS (Jersild et al. 1973; Olerup and Hillert 1991).

1.2.4 Genetic aspect in MS

The study of genetic factors influencing the development of MS is not a new topic, it was proposed almost a century ago. Nobody disputes the involvement of genes in monogenic disorders that consistently show simple Mendelian inheritance patterns. However, with non-Mendelian disease (complex traits), it is necessary to prove claims of genetic determination. Several approaches can be used to evaluate familial aggregation including familial recurrence rate with population comparison, twin studies, adoption studies and genealogy. Genetic mapping by linkage or association analysis can finally determine whether genetic factors are involved in a trait.

1.3 Relationship between HLA and MS

The candidate gene approach has successfully established the importance of HLA class II genes in MS by simple case-control association studies (Jersild et al. 1973; Olerup and Hillert1991), which was later confirmed by linkage studies (Tienari et al. 1993; Fogdell et al. 1997). Furthermore, these finding were also confirmed by four genome search studies (Ebers et al. 1996; Sawcer et al. 1996; The Multiple Sclerosis Genetic group 1996; Kuokkanen et al.1997).

Recently HLA class I alleles that increase and decrease the genetic susceptibility to MS were identified in 200 Swedish MS patients and 210 Swedish healthy controls (Fogdell-Hahn et al. 2000). In this report, the HLA-A*031(class I) allele increases the risk of MS (odds ratio =2.1) independently of DRB1*15 (classII) and DQB1*06(classII); HLA-A*021(classI) decrease the overall risk (odds ratio =0.52).

Later research found that the combination of HLA-A (class I) and HLA-DRB1 (class II) alleles, represented by HLA-A*02 and HLA-DRB1*15 increase the risk of MS 23-fold and the influence is independent of other variation (B. Brynedal et al. 2007). However HLA-C*05 (classI) association with MS was reported by conditioning on DRB1*01 (class II) absence (Yeo et al. 2007).

These later findings suggest that genes in HLA class I family influence MS. Is this due to linkage disequilibrium between HLA class I and HLA class II gene? Is the MS association with class I genes independent of other effect?

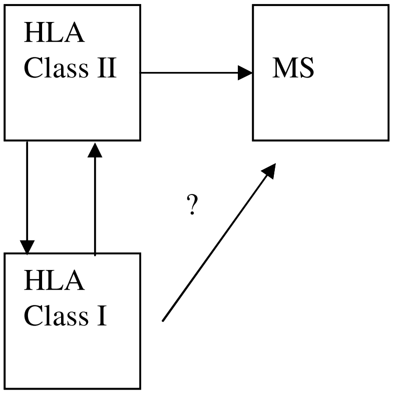


Figure 1.3 Relationship between HLA genes and multiple sclerosis (MS). The influence of HLA class II gene on MS is known as well as the presence of LD between class I and class II regions. The influence of HLA class I gene on MS susceptibility is our research aim.

2 Recursive partitioning method

2.1 What is recursive partitioning?

Recursive partitioning (RP) can be viewed as a tree based conditional gene finding approach. The basic statistical motive is to divide a data set into parts where the objects in the parts are homogenous. Once a split is made based upon one gene, then the subsequent splits are conditional on the presence or absence of a specific variant of that gene. Recursive partitioning is most easily described using an example. Consider the figure below, a hypothetical example. The RP diagram is read in the following way. In the parent node, N, there are 2000 individuals and 212 of them have the disease.

A gene scan is done over a number of candidate genes (in this example bi-allelic SNPs) and it is determined that a person who is homozygous 1_1 for Gene i belongs to node N1 in Figure 2.1. There are 400 such individuals and 100 of them have the disease. The remaining 1600 individuals belong to Node N0, 112 of which have the disease. At this point in the analysis the original data set has been divided into two groups based on discrimination with disease status. At node N1, the proportion of patients is much higher than at N0. Next a gene scan is done over the 1600 individuals in Node N0 and again a split is determined based on discrimination with disease status. It is determined that 400 individual with 1_1 or 1_0 for gene j, 100 of which have the disease form Node N01. Of the remaining 1200 individuals from Node N00, 12 individuals have the disease. Nodes N1 and N01 can be viewed as two distinct forms of the disease based upon Genes i and j. The process is repeated many times with the aim to reach the maximal proportion of patients in the other nodes. For example the first splitting could be based upon another gene k which gives N1 50/400 and the second splitting gene h gives N01 30/400. This alternative leads to a less proportion of patients in the N1 and N01group. The partition that gives the maximal proportion of patients is selected as in Figure 2.1.

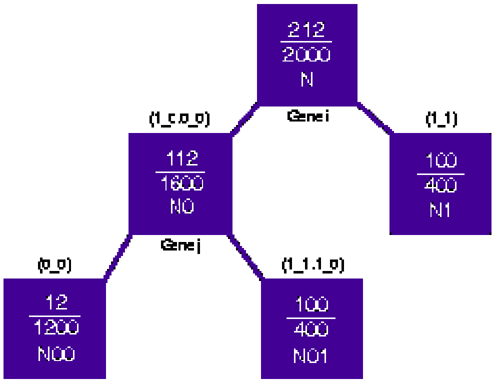


Figure 2.1 Helix tree after recursive partitioning

The analysis is conditional. When attempting to split node N0, the search is in the 1600 with genotype 1_0, 0_0 individuals for gene i, so the search for a new gene is dependent on its combined effect with genotype 1_0, 0_0 of gene i.

The gene i is associated to the disease since in node N1 the proportion of patients is much higher than in N0 and N. By conditioning on the gene i variant, we eliminate disease association to gene i when searching for another gene which is associated with the disease.

2.2 The RP method

There are two steps involved when the RP approach is used. The first is the partitioning of the gene space. The second is the pruning of the tree using validation data.

2.2.1 Recursive Partitioning

If a candidate multi-allele gene i is analysed, we assume that there are k variants i_1, i_2 ,i_3, i_4, i_5….i_k defined. We view the observed gene variants as points in k dimensional space and simply illustrate the RP method in a square shown in Figure 2.2 and 2.3. For example, at first, we split the gene i space with respect to i_4. The individuals are divided into two groups according to genotypes, one set of genotypes in one part and the complementary sets in the other part, genotype (1_1) group and complement genotypes (1_0,0_0) group, where 1 represents the i_4 allele, 0 represents the no-i_4 allele. See Figure 2.2.

Consequently, a split with respect to another variant i_5 would be performed and this splitting repeats with the rest allele variants. The purpose of splitting is to make the sample space as homogenous or 'pure' as possible with respect to the disease status. Eventually high purity sample spaces will be reached. Ideally, in those sample spaces which is represented by some small rectangles, all the individuals are either patients or healthy controls. See Figure 2.3.

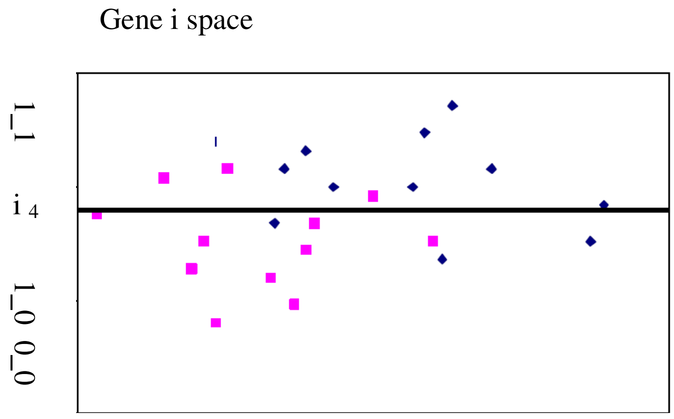


Figure 2.2 The observations are divided in to two parts according to allele variants of i_4 in gene i. In the upper part of the rectangles, the individual genotype is (1_1) of i_4, in the lower part, the individual genotype is (1_0) or (0_0) of i_4.

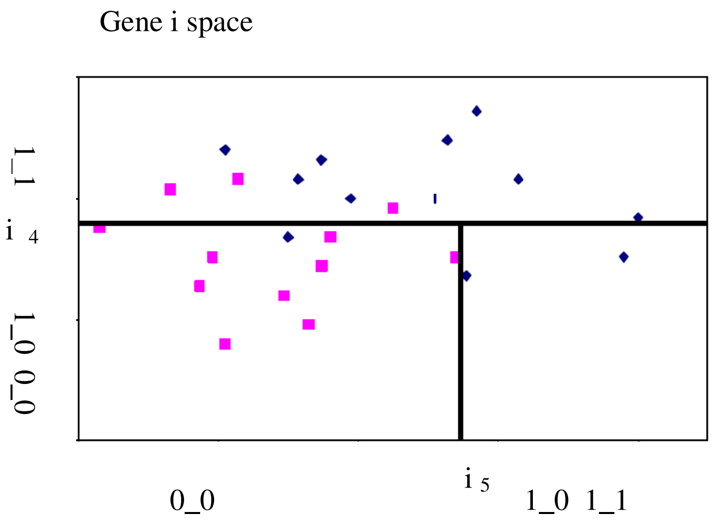


Figure 2.3 The observations are divided into two parts according to allele variants of i_5 in gene i. In the left part of the lower panel, the individual genotype of i_5 is (0_0); in the right part, the individual genotype of i_5 is (1_0) or (1_1). By i_5 splitting, a homogenous space (the right-lower rectangle) is obtained.

The impurity at each partition can be evaluated by the Gini impurity index

$$I(A) = 1 - \sum_{k=1}^{C} P_k^2, \qquad (2.1)$$

where P_k is the fraction of the observations in rectangle A that belongs to class k. C is the total number of classes for the disease variable, (in our application C=2). The partition method with minimal Gini impurity index will be selected. (On line Lecture notes 3 "classification trees" *www.myoops.org*)

For example

When p_1=1, p_2=0,

I_1(A)=0.

When p_1=0.5, p_2=0.5,

$I_2(A)=0.5$.

where p_1 is the proportion of individuals with a specified genotype set, for example (1_1). p_2 is proportion of individuals with the complement genotype set, for example (1_0), (0_0). Hence, $p_1+ p_2 = 1$. Since $I_1(A)< I_2(A)$, we chose the first alternative to split the observation space.

2.2.2 Pruning of the tree

The second step of recursive partitioning is to use validation data for pruning a tree that is growing excessively. The pruning step is similar to the backward deletion in the ordinary linear regression. In our examples, the last few splits resulted in rectangles with a very few observations (indeed four rectangles in the full tree have just one observation). Over-fitting may cause inaccurate interpretation.

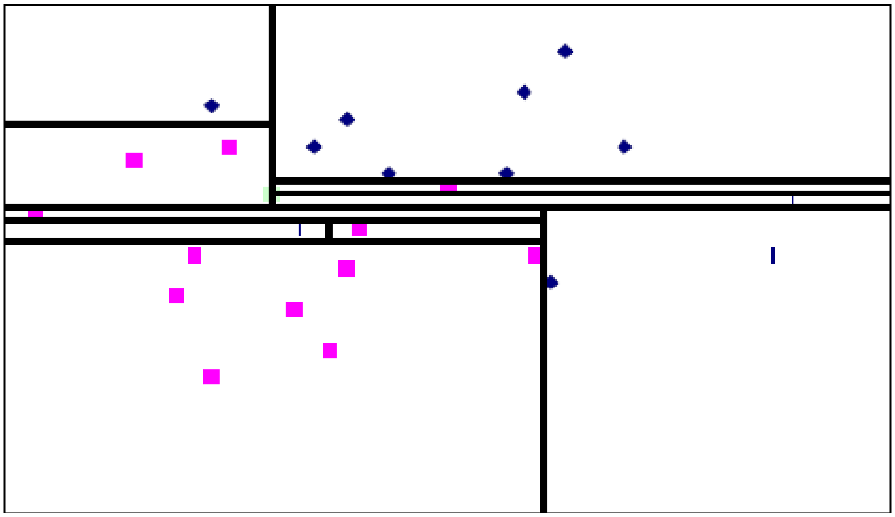


Figure 2.4 The observations are divided into small rectangles based on a series of alleles. In some rectangles, there is only one observation. After pruning, those branches will be cut off.

We can intuitively see that these final splits capture features specified to the training set. We call this situation over-fitted. Pruning is involved in successively selecting a decision node and re-designating it as a terminal node. The best tree is defined as the one in the sequence that gives the smallest misclassification error in cross-validation.

Cross-validation, sometimes called rotation estimation, is the statistical practice of partitioning a sample of data into subsets such that the analysis is initially performed on a single subset, while the other subset(s) are retained for subsequent used in confirming and validating the initial analysis.

Figure 2.5 gives an example of how to choose the pruning stopping point. The decreasing number of decision nodes increases the error in cross validation with a slow trend in the beginning of pruning, from 30 decision nodes down to 10 nodes. The error goes up sharply when the tree is quite small. So we should avoid choosing a tree with few number nodes during the pruning.

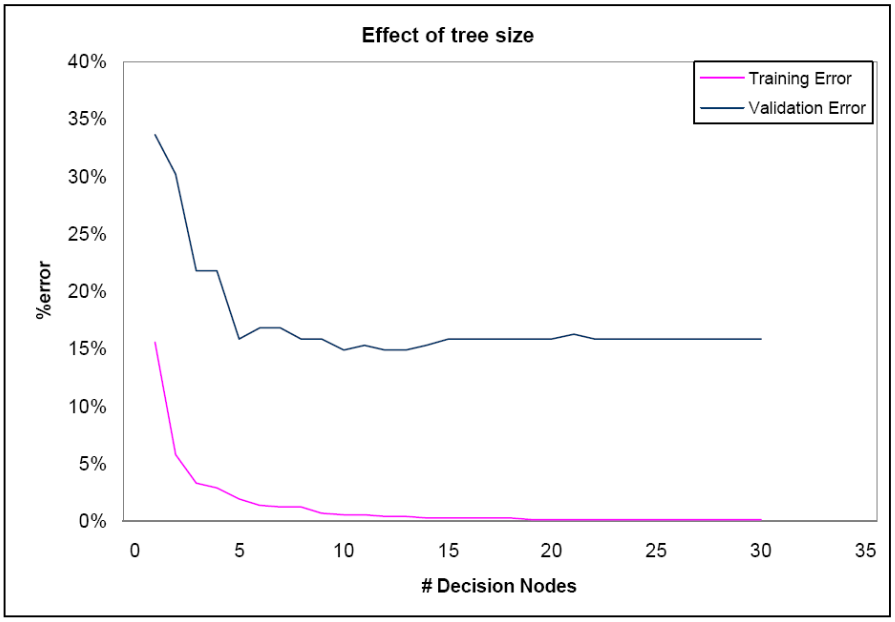


Figure 2.5 How to select minimum error tree? Here a ten nodes tree will be chosen because it reaches minimal validation errors. The upper curve represents the error with regard to node numbers using validation data.

However, cross validation is not a unique pruning method; an alternative approach using χ^2 tests is reported by Zhang and Bonney (2000). To test if a split is significant, a 2×2 table is created with the number of patients and controls in presence/absence of the gene. A split is regarded as unnecessary if the χ^2 tests from this split as well as its further splits are not significant at a pre-specified level. All nodes resulting from unnecessary splits are removed. To illustrate this process, let us begin with the tree in Figure 2.6 and explain how the nodes are pruned.

Under each internal node (represented by a circle), we list a raw χ^2 statistic. For example, we have “raw: 59.3” under node 1, indicating that the χ^2 statistic from a 2×2 tables with cell values of 54, 600 (from node 2), 193 and 637 (from node 3) equals 59.3.
We also report a maximum χ^2 statistic obtained as follows. Let us take two representative nodes (3 and 5) from Figure 2.6 and show how their maximum χ^2 statistics are derived. For node 3, we have a raw χ^2 statistic of 13.7. Node 3 has two offspring internal nodes (6 and 12), and their raw χ^2 statistics are 10.9 and 4.8. Then, the maximum χ^2 statistic for node 3 is the maximum of 13.7, 10.9, and 4.8, which is 13.7 and turns out to be the same as the raw χ^2 statistic of node 3. For node 5, however, its raw χ^2 statistic is 4.6 and its offspring node (10) has a larger raw χ^2 statistic of 8.4. Thus, the maximum χ^2 statistic for node 5 becomes 8.4, which is the maximum of 4.6 and 8.4. Likewise, a maximum χ^2 statistic can be assigned for any internal node as displayed in Figure 2.6. After the maximum χ^2 statistics are computed for all internals, we then set a critical χ^2 level, e.g., 10.83 at the significance level of 0.001. An internal node becomes a terminal node (in other words, its offspring nodes are pruned) if its maximum χ^2 statistic is less than the critical level. Consequently, nodes 4, 5, and 12 become terminal nodes because their maximum χ^2 statistics are less than 10.83, and nodes 8 through 11 and 13 through 19 are pruned. It is useful to note that the pruned internal nodes (e.g., node 10) cannot have maximum χ^2 statistics greater or equal to the critical level because of the way by which the maximum χ^2 statistics are defined. This explains how we obtained the tree in

Figure 2.7 at the significance level of 0.001. For the significance level of 0.005 or any other level, the pruning is done in the same way.

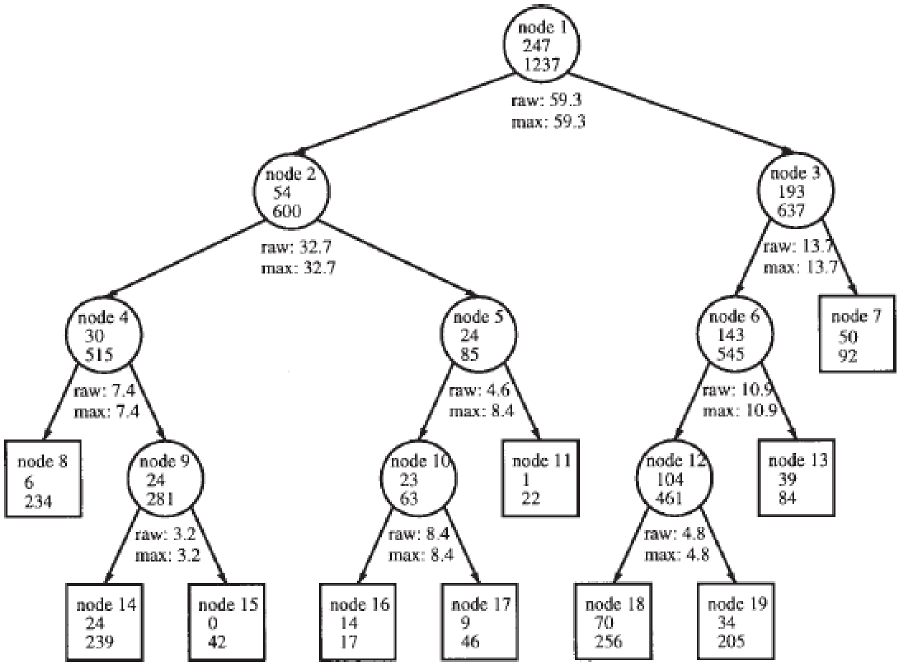


Figure 2.6 Illustration of tree pruning. Inside each node are the node number **(top)**, the numbers of affected **(middle)**, and unaffected **(bottom)** individuals. Under each internal node is the raw and maximum $\chi 2$ statistics, as described in the text in detail.

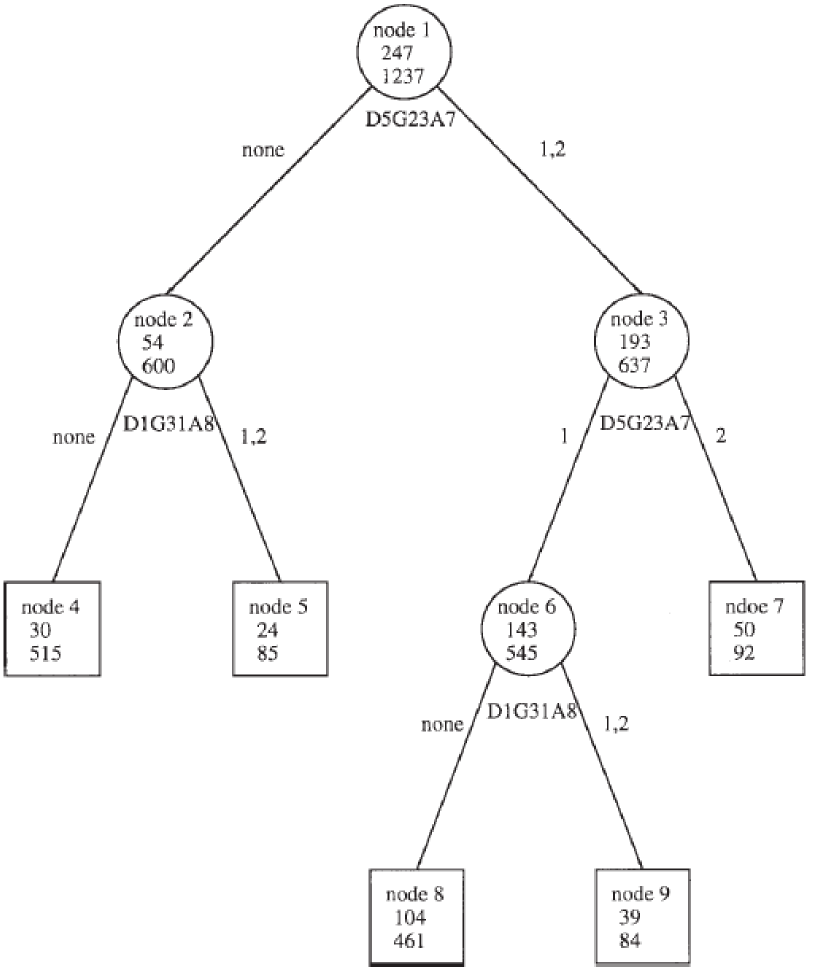


Figure 2.7 The pruned tree at significance level 0.001. Inside each node are the node number **(top)**, the numbers of affected **(middle)**, and unaffected **(bottom)** individuals. Under each internal node is the split based on the genotype. For example, node 1 is split based on the number of D5G23A7 alleles. The 'none' means there are no D5G23A7 allele; '1,2' means the individuals have 1 or 2 D5G23A7 alleles.

3 The logistic regression model

We define the risk of disease in term of odds k = p/(1-p), with p the probability of a randomly sampled individual from the population to be a patient and 1-p probability to be a healthy control. A logistic regression model is defined as

$$Log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 ...,$$

where X_i =0 or 1, i=1,2 ….. and $e^{\beta i}$ is the odds ratio for comparing the odds of X_i=1 with X_i=0. Thus the variable 1 parameter exponential ($e^{\beta 1}$) compares the odds of being a case for a subject who is variable X_1 positive to the odds of being a case for a subject who is variable X_1 negative. β_1 is the log odds ratio if variable X_1 changes 1 unit.

When data are stratified with regard to a variable A, the model should include variable A as a confounding variable.

$$Log(\frac{p}{1-p}) = \beta_0 + \beta_1 A + \beta_2 X_1 + \beta_3 X_2 + \beta_4 X_3$$

A class II gene involved in the recursive partitioning is defined as a confounding variable in the logistic regression model for class I genes. This implies that the effect of X_i, i=1,2….n is calculated conditionally on a specific value of A, the effect of X_i is the same for all values of A. In our study, we have two values for each of A, which are 0 and 1. Moreover, all the HLA class II genes may be confounding variables in the logistic regression model to class I genes.

4 Aims of the present study

I To apply the recursive partitioning method to describe association between HLA class II loci and MS.

II To model the association between HLA class I loci and MS, conditional on a partition of HLA class II loci.

5 Materials and Method

5.1 Materials

Totally 3174 individuals were included in the present study. Among them, 983 Swedish MS patients were recruited by the Department of Neurology, Karolinska University Hospital, Stockholm and 546 MS patients were collected from Norway. All patients fulfilled the McDonald criteria of MS (McDonald, 2001) with mean age of 54.1years and the female to male ratio of 2.5. The ethical board of Karolinska Institutet approved the study and informed consent was obtained from all participants. The controls (mean age of 47.3 and female to male ratio of 1.8) consisted of 1075 Swedish blood donors from Sweden and 570 Norwegian bone marrow donors from Norway.

All patients and controls in this study were typed for the three HLA loci, HLA-DRB1, HLA-A and HLA-C using Olerup SSPTM HLA Low resolution Kits (Olerup, 1992).

5.2 Methods

5.2.1 Helix tree generation

Two softwares, Xlminer (Cytel Statistical software ver 3.0) and the recursive partitioning library in R (Rpart, http://cran.r-project.org;. refs 20-22), were used to generate a helix tree. All possible binary splits of the data corresponding to presence or absence of various different genotypes at the HLA class II loci were considered. For each of the class II loci, data were divided into two groups that maximally reduced the impurity. The minimal Gini impurity index was applied in the Xlminer software.

In the R partitioning program, another index called information with the form f(p)=-2plog(p) was used in stead of the Gini impurity index in XLminer, where p is the proportion of observations in a node that has the different genotypes. For example p_1=1, that means all the observations have different genotypes. Then the information index f(p_1)=0. Assume p_2=0.5, then f(p_2)=0.15. Since f(p_1)< f(p_2), we chose the second way to split the observations. A HLA class II Helix tree was generated by choosing a maximal information index in R partitioning program.

5.2.2 Pruning and evaluation of the Helix tree

In order to prune the Helix tree by use of the Rpart program, a cp plot was performed. Complex parameter (cp) indicates how observations vary in one group. It is usually computed using a variance estimate from the largest model under consideration. This will be done automatically when the `cpplot` function is used. The outcome of cp plot is a cp table to assess how the models fit. It gives a visual representation of the cross-validation result in an Rpart object.

The χ^2 test method was also used in the tree pruning. By comparing the tree model from different pruning methods, a suitable model was selected.

5.2.3 Testing class I locus in the case-control collections

A backward stepwise logistic regression procedure was used to test whether any of the 17 loci typed in the HLA class I has an effect on MS in addition to the HLA class II. The class II loci according to the partitioning model were placed in the regression model as confounders and other class I loci were added. The significance of class I genes conditional on the class II genes in the model were evaluated by a likelihood ratio test. The significant loci remained in the model. Interactions between HLA class II loci were tested as well.

We used a family-wise error rate (FEW) to calculate a critical value (a threshold). Null hypothesis: none of the markers are significant.

FEW = P {Reject at least one true null hypothesis}

In the present study,

FEW = p-value for a single marker × (17 + the number of variables in the confounding model)

where 17 represents the number of tested HLA class I loci.

We chose a threshold of FEW 0.05.

6 Results and discussion

6.1. Confounding variable identification

6.1.1 Helix tree growing and pruning

All the class II variables were analysed by the XLminer software. After classification, a full helix tree was obtained and showed in Figure 6.1. The circle represents split nodes and the rectangle represents terminal nodes. The data were split into 5 groups showing minimal impurity according to Geni impurity index. Totally, 1407 individuals carried at least one of HLA DR15 allele in the genotype (DR15+), 1767 individual carried complement alleles (DR15-). The obtained groups were 1) DR15+; 2) DR15- DR4+; 3) DR15-, DR4-, DR1+; 4) DR15-,DR4-,DR1-, DR3+ and 5)DR15-,DR4-,DR1-,DR3-. According to the $\chi 2$ method (Zhang et al 2000), we determined the model based on the maximal of χ^2 statistics for the nodes. The maximal of χ^2 statistics is a maximal value of χ^2 statistics from the split node to the terminal node. For instance, the χ^2 statistics of the node DR4 was 1.152, but the maximal of χ^2 statistics of DR4 was 1.961. It is because χ^2 statistics of the node DR1 located downstream of the DR4 was 1.961. We chose the critical value of 1.96 for the χ^2 statistics. The maximal χ^2 statistics of node DR1 was 1.961>1.96, therefore we retained node DR1. The node DR3 was cut off since the maximal χ^2 statistics of the node DR3 was 0.3032, which was < 1.96. After pruning, the model consisted of split nodes DR15, DR4 and DR1. See Figure 6.2. The pruning result was confirmed by likelihood ration tests as well.

We analysed ratio of health control (HC) and MS in the nodes. From Figure 6.2, in split node DR15, HC: MS was 1.073. After the first split with respect to DR15, in the terminal node DR15 (DR15+), the ratio reduced to 0.503; in the split node DR4 (DR15-), the ratio raised to1.98. Hence, after the first splitting, the individuals were divided into two groups with either more proportion of HC or MS than that at the initial situation. The splitting continued based on the nodes DR4 and DR1. Consequently, the ratio of HC and MS reached to 2.4 in the terminal node DR1.

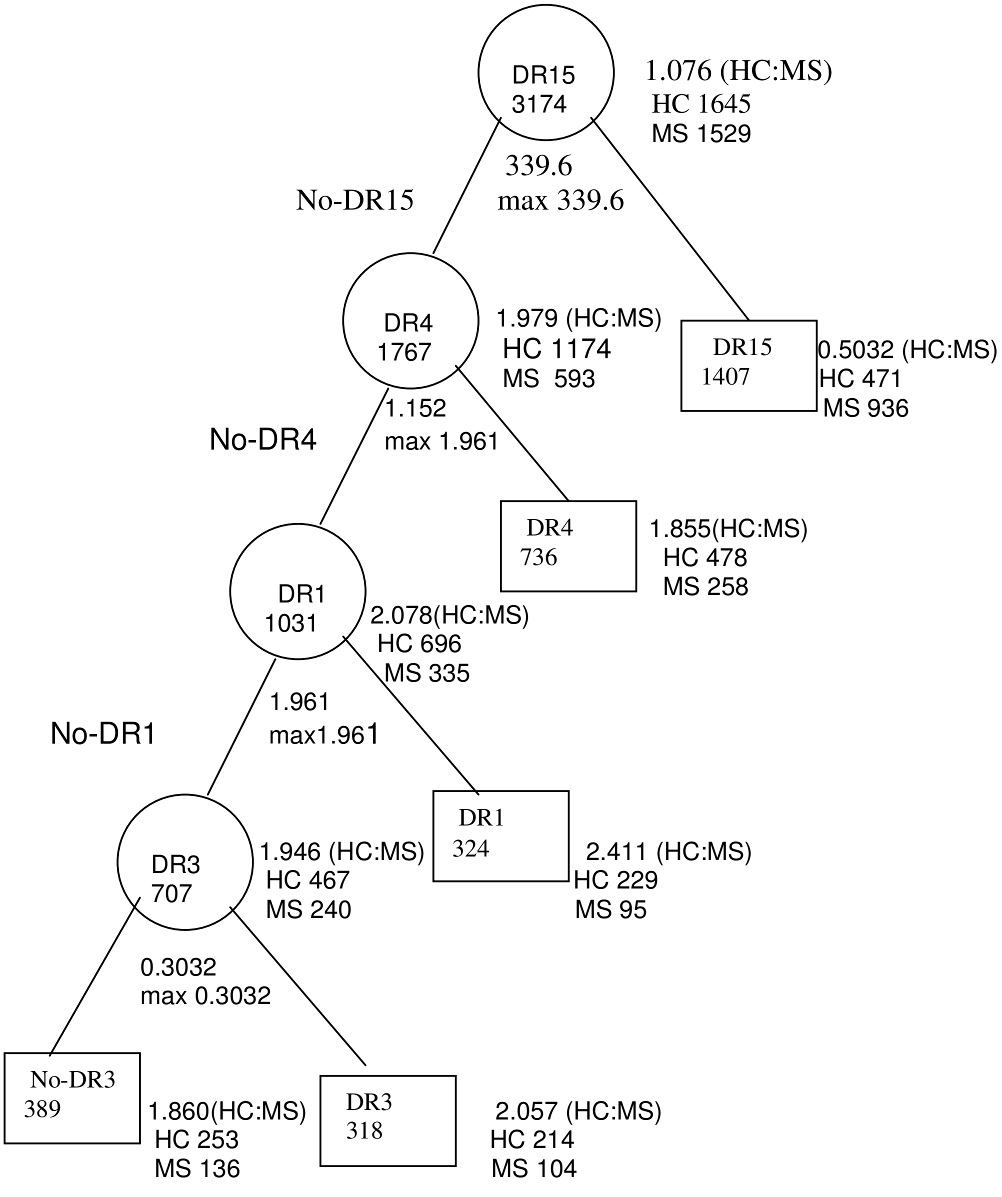


Figure 6.1 The classification of HLA-class II gene by the Xlminer software. Numbers at right of the nodes are ratios of healthy controls with MS patients within the groups. Numbers under split nodes are the χ^2 statistics and the maximal of χ^2 statistics. The circles represent the split nodes and the rectangles represent the terminal nodes.

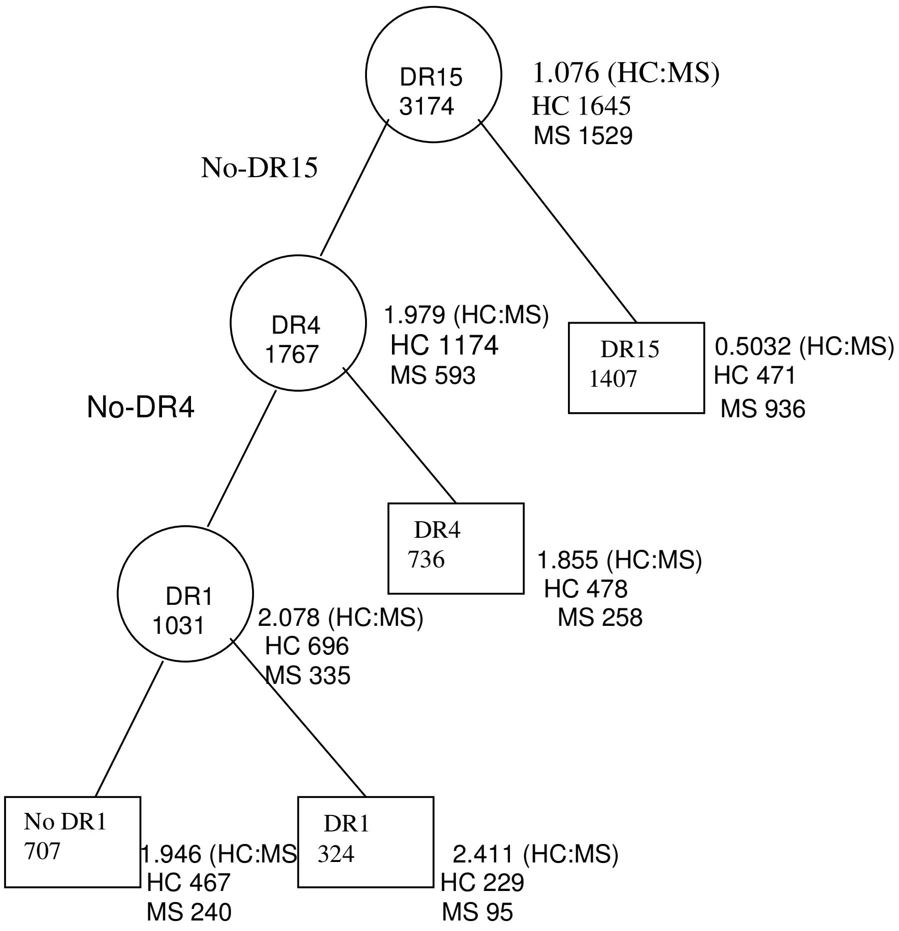


Figure 6.2 After pruning with the χ^2 method, DR15 and DR4 and DR1 retained in the confounding model.

The pruning was also performed automatically by XLminer program with validation data. The program chose a subset of data randomly and tested all the alternative models with cross validation methods. The result displayed with only one node is the ‘main effect’ variable. See Figure 6.3.

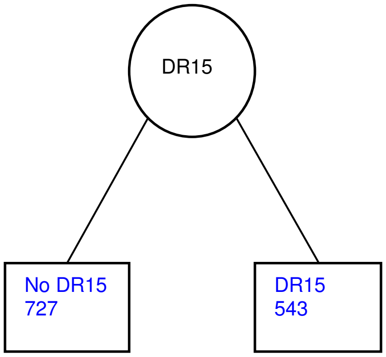


Figure 6.3 The best pruning tree from XLminer.

The data were also analysed by R in the Rpart library to carry out tree partitioning. The saturate partitioning tree was 11 nodes. By choosing the minimal number in a node (minsplit) 100, and cp=0.001, we obtained a model with 9 nodes.

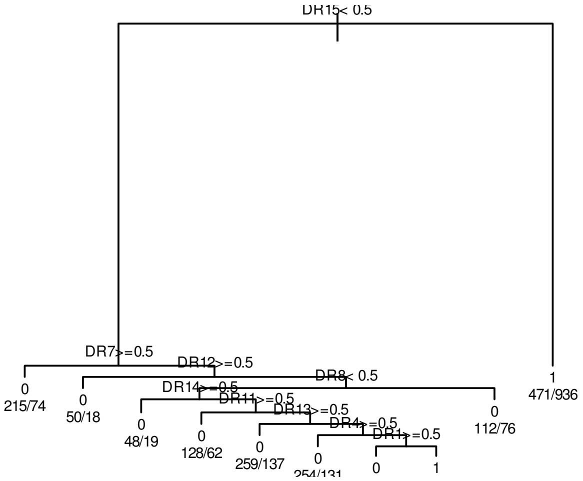


Figure 6.4 The helix tree from Rpart library, cp=0.001 minsplit=100

The Helix tree grown by recursive partitioning using the Rpart program was pruned. The CP plot gives a table to find tree size with respect to cross-validation errors of the model. A good choice of cp for pruning is often the leftmost value for which the mean lies below the horizontal line. The result indicated that the best tree size is 2.

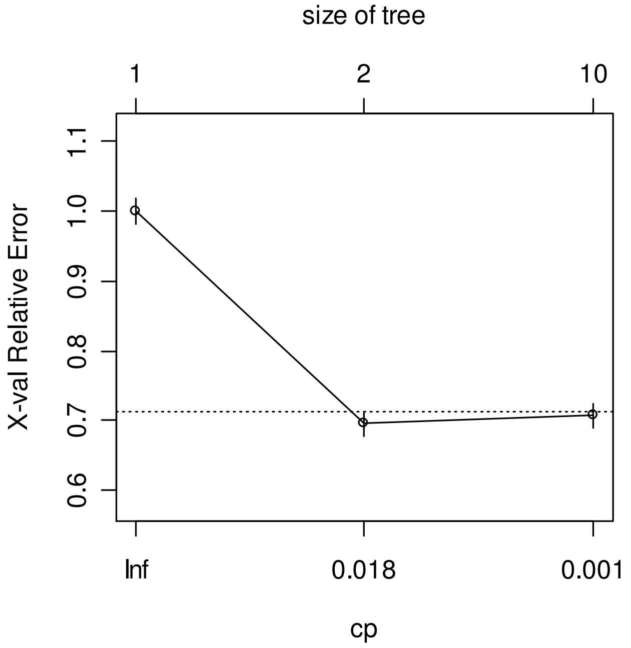


Figure 6.5 Cp plot from Rpart library, the value of complex parameter (cp) with respect to the cross validation error. The size of the tree corresponding to the cp is shown above the graph.

The helix tree (9 nodes) from the Rpart program with cp=0.001 minsplit=100 was confirmed by the χ^2 pruning method, we used likelihood ratio test to approximate the χ^2 test. Since the last split node DR1 exceeded to the threshold, no node located upstream of the DR1 could be deleted.

6.1.2 Evaluation of the helix tree models

How to choose the confounding variables depends on the pruning result. But it is difficult to judge which pruning result, for example, by cross validation or the χ^2 test procedure, is the best. The rationale for minimizing the number of variable in a model is that the resulting model is more likely to be numerically stable (Hosmer and Lemeshow 2000). A overfitted model would produce numerically unstable estimates. However the number of variables must be sufficient to prevent residual confounding (Nejentsev 2007).

From the partitioning and pruning results, two logistic regression models were selected. Model 1 was obtained from the XLminer program, pruned by χ^2 methods. Model 2 was obtained by the recursive partitioning using the Rpart program confirmed by χ^2 pruning methods.

$Model\ 1:$

$$Log(\frac{p}{1-p}) \quad = \quad \alpha + \beta_1 DR1 + \beta_2 DR4 + \beta_3 DR15$$

$Model\ 2:$

$$Log(\frac{p}{1-p}) = \alpha + \beta_1 DR15 + \beta_2 DR7 + \beta_3 DR12 + \beta_4 DR8 +$$

$$\beta_5 DR14 + \beta_6 DR11 + \beta_7 DR13 + \beta_8 DR4 + \beta_9 DR1$$

$$where\ \frac{p}{1-p}\ is\ odds\ of\ MS.$$

Using the real data test, we found that the two models had similar deviances and Akaike's information criterions (AIC). See Table 6.1.

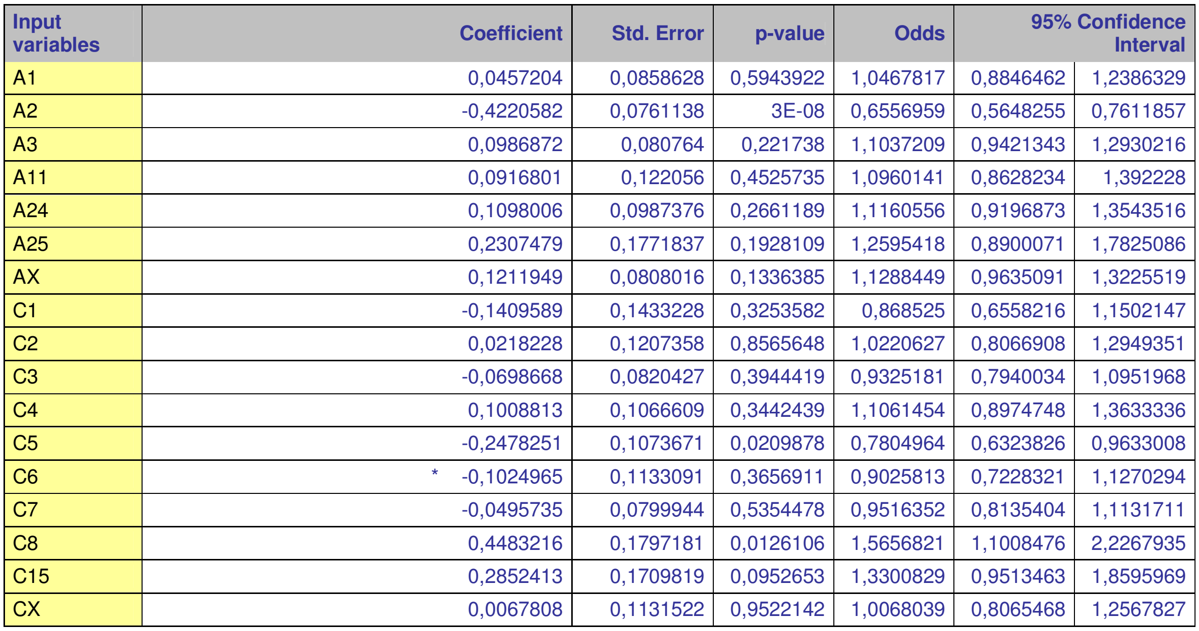
Table 6.1 Deviance and AIC of the two models

	Model 1	Model 2
deviance	4030.8	4016.8
AIC	4042.8	4034.8

We also tested the confounding model stability by adding class I variables to the class II confounding models. One class I allele was added at two different class II models. If we found the same coefficient and p-values of the class I allele between two different confounding models, it indicated that the confounding models were effective and accurate. The test results in Table 6.2 show that the two models give similar results.

Table 6.2 Confounding model stability testing with class I variables

Model 1



Input variables	Coefficient	Std. Error	p-value	Odds	95% Confidence Interval	
A1	0,0457204	0,0858628	0,5943922	1,0467817	0,8846462	1,2386329
A2	-0,4220582	0,0761138	3E-08	0,6556959	0,5648255	0,7611857
A3	0,0986872	0,080764	0,221738	1,1037209	0,9421343	1,2930216
A11	0,0916801	0,122056	0,4525735	1,0960141	0,8628234	1,392228
A24	0,1098006	0,0987376	0,2661189	1,1160556	0,9196873	1,3543516
A25	0,2307479	0,1771837	0,1928109	1,2595418	0,8900071	1,7825086
AX	0,1211949	0,0808016	0,1336385	1,1288449	0,9635091	1,3225519
C1	-0,1409589	0,1433228	0,3253582	0,868525	0,6558216	1,1502147
C2	0,0218228	0,1207358	0,8565648	1,0220627	0,8066908	1,2949351
C3	-0,0698668	0,0820427	0,3944419	0,9325181	0,7940034	1,0951968
C4	0,1008813	0,1066609	0,3442439	1,1061454	0,8974748	1,3633336
C5	-0,2478251	0,1073671	0,0209878	0,7804964	0,6323826	0,9633008
C6	* -0,1024965	0,1133091	0,3656911	0,9025813	0,7228321	1,1270294
C7	-0,0495735	0,0799944	0,5354478	0,9516352	0,8135404	1,1131711
C8	0,4483216	0,1797181	0,0126106	1,5656821	1,1008476	2,2267935
C15	0,2852413	0,1709819	0,0952653	1,3300829	0,9513463	1,8595969
CX	0,0067808	0,1131522	0,9522142	1,0068039	0,8065468	1,2567827

Model 2

Input variables	Coefficient	Std. Error	p-value	Odds	95% Confidence Interval	
A1	-0,040862	0,0899451	0,6496134	0,9599616	0,8048083	1,1450259
A2	-0,4149122	0,0769327	7E-08	0,6603982	0,5679638	0,7678761
A3	0,1020157	0,0812513	0,2092763	1,1074008	0,9443732	1,2985721
A11	0,1269389	0,1234286	0,303743	1,1353476	0,8913869	1,4460772
A24	0,0936535	0,0991339	0,3448032	1,0981792	0,9042536	1,333694
A25	0,2416516	0,1787808	0,1764828	1,2733504	0,8969525	1,8077003
AX	0,1701454	0,081938	0,0378464	1,1854773	1,0095956	1,3919991
C1	-0,1516555	0,1447556	0,2947924	0,8592842	0,6470245	1,1411771
C2	0,02382	0,1227341	0,8461149	1,0241059	0,8051438	1,3026156
C3	-0,0835495	0,083877	0,3192039	0,9198456	0,7804025	1,0842044
C4	0,1398572	0,1077637	0,1943512	1,1501095	0,9311307	1,4205868
C5	-0,2260517	0,1092966	0,0386173	0,7976769	0,6438631	0,9882354
C6	* 0,0309449	0,1213559	0,7987286	1,0314287	0,8130941	1,3083911
C7	-0,1498677	0,0841031	*0,0747571	0,8608218	0,7300028	1,015084
C8	0,5267255	0,1822166	*0,0038444	1,6933783	1,184816	2,4202323
C15	0,3001238	0,1725285	0,0819362	1,3500259	0,962688	1,8932092
CX	0,0809605	0,1153596	0,4827978	1,0843281	0,8649011	1,3594239

* represents the different value between two models.

6.2 Logistic regression models

6.2.1 Tests for interaction of confounding

After the 'main effect' variables of the class II genes were determined, the interactions between the main effect variables were tested. The forward stepwise method was used to test for the interaction of confounding variables in model 1. The result in Table 6.3 demonstrated the interactions between DR1 and DR4, and between DR1 and DR15.

Models in forward stepwise method

$$Log(\frac{p}{1-p}) = \beta_0 + \beta_1 DR1 + \beta_2 DR4 + \beta_3 DR15 + \beta_4 DR1 \times DR4$$

$$Log(\frac{p}{1-p}) = \beta_0 + \beta_1 DR1 + \beta_2 DR4 + \beta_3 DR15 + \beta_4 DR1 \times DR15$$

$$where \ \frac{p}{1-p} \ is \ odds \ of \ MS.$$

Table 6.3 P-values of interaction variables

variables	DR1	DR4	DR15
DR1		0.000689	0.00796
DR4			0.309
DR15			

6.2.2 Class I variable tests by the backward elimination method

All the HLA class I exposure variables were simultaneously added in the class II confounding model and variables without significance were successively eliminated. The combined data set (Swedish and Norwegian) showed that the exposure variable A2 was significant. The model for combined data set and the result in Table 6.3 are based on criterion of P-values before Bonferroni correction.

Model for combined data:

$$Log(\frac{p}{1-p}) = \beta_0+\beta_1 DR1+\beta_2 DR4+\beta_3 DR15+\beta_4 DR1\times DR4+\beta_5 DR1\times DR15+\beta_6 A2$$

Table 6.3 Results of HLA class I variables tested using the combined data (Swedish and Norwegian)

	Estimate	Std. Error	z value	Pr(>lzl)
A2	-0.42278	0.07630	-5.541	3.01e-08

The data were also analysed separately according to different populations (Swedish and Norwegian). For the Swedish MS, A2 was significant with the odds ratio 0.64 that confirmed previous results (odds ratio 0.63, Brynedal 2007). A new allele Cn was also significant with a P-value 0.000853, FWE 0.0144, odds ratio 2.29. The p value of allele C5 was 0.0412, but the FWE was 0.7004 >0.05. Therefore, C5 was discarded. The model for Swedish data set and the result in Table 6.4 are based on criterion of P-values before Bonferroni correction.

Swedish data tests
Model:

$$Log(\frac{p}{1-p}) = \beta_0 + \beta_1 DR1 + \beta_2 DR4 + \beta_3 DR15 + \beta_4 DR1 \times DR4 + \beta_5 DR1 \times DR15 + \beta_6 A2 + \beta_7 C5 + \beta_8 Cn$$

Table 6.4 The result of HLA class I variables with Swedish data

	Estimate	Std. Error	z value	Pr(>lzl)
A2	-0.44690	0.09634	-4.639	3.51e-06
C5	-0.28805	0.14112	-2.041	0.041233
Cn	0.82980	0.24880	3.335	0.000853

In the Norwegian data, no significant allele was found because p value for A2 was 0.018 and FWE was 0.306. The model for Norwegian data set and the result in Table 6.5 are based on criterion of P-values before Bonferroni correction.

$$Log(\frac{p}{1-p}) = \beta_0+\beta_1 DR1+\beta_2 DR4+\beta_3 DR15+\beta_4 DR1\times DR4+\beta_5 DR1\times DR15+\beta_6 A2$$

Table 6.5 The result of HLA class I variables with Norwegian data

	Estimate	Std. Error	z value	Pr(>|z|)
A2	-0.3010	0.1280	-2.352	0.01866

6.3 Discussion about confounding model

We tested the real data sets with different confounding models, and compared the HLA class I gene findings with different class II confounding models.

6.3.1 Real data set test with different confounding models

Confounding strata with significant nodes

$$Model\ 1:\ Log(\frac{p}{1-p}) = \alpha + \beta_1 DR1 + \beta_2 DR15$$

Confounding strata with pruned recursive partitioning

$$Model\ 2:\ Log(\frac{p}{1-p}) = \alpha + \beta_1 DR1 + \beta_2 DR4 + \beta_3 DR15$$

Confounding strata with recursive partitioning

$$Model\ 3:\ Log(\frac{p}{1-p}) = \alpha + \beta_1 DR1 + \beta_2 DR3 + \beta_3 DR4 + \beta_4 DR15$$

Confounding strata with recursive partitioning and one additional interaction term

$$Model\ 4:\ Log(\frac{p}{1-p}) = \alpha + \beta_1 DR1 + \beta_2 DR4 + \beta_3 DR15 + \beta_4 DR1 \times DR4$$

Confounding strata with recursive partitioning and two additional interaction terms

$$Model\, 5:\ Log(\frac{p}{1-p}) = \alpha + \beta_1 DR1 + \beta_2 DR4 + \beta_3 DR15 + \beta_4 DR1 \times DR4 + \beta_5 DR1 \times DR15$$

Confounding strata according to the recursive partitioning from the R-part program

$$Model\ 6:\ Log(\frac{p}{1-p}) = \alpha + \beta_1 DR15 + \beta_2 DR7 + \beta_3 DR12 + \beta_4 DR8 + \beta_5 DR14 + \beta_6 DR11 + \beta_7 DR13 + \beta_8 DR4 + \beta_9 DR1$$

$$where\ \frac{p}{1-p}\ is\ odds\ of\ MS.$$

6.3.2 Comparison of HLA class I gene finding results with different confounding models

Tables 6.6-6.8 present p values of HLA class I variables in the 6 models above. All the class I variables were added to the 6 models. Using the backward eliminate method, alleles that exceeded the significant threshold were recorded in the tables. The threshold of 0.05 before Bonferroni corrections was used. Some variables included in model 1 were not significant in models 2, 3, 4, 5 and 6, for instance A1, C3, C5 in Table 6.6. Thus, the absence of sufficient confounding variables may result in false positive results.

When we calculated p values using FEW, weak positive associations of the alleles did not exceed the threshold and those false positive were discarded. All the models gave the same results. The allele A2 was significantly associated with MS in combined data of Swedish and Norwegian set. Interestingly, the allele Cn showed significant association with Swedish MS, though A2 has also shown association in Swedish data. However, no significant association of any alleles in Norwegian data was found. In Table 6.6-6.8, the threshold after Bonferroni correction for every model is showed. For instance, in model 1, threshold is 0.0026.

Table 6.6 Comparison of different confounding models with combined data

	Model 1 DR1, DR15 P<0.0026	Model 2 DR1, DR4,DR15 P<0.0025	Model 3 DR1,DR3, DR4,DR15 P<0.0024	Model 4 DR1, DR4,DR15, DR1*DR4 P<0.0024	Model 5 DR1,DR4, DR15, DR1*DR4, DR1*DR15 P<0.0023	Model 6 DR1,DR4,DR7, DR8, DR11,DR12, DR13, DR14,DR15, DR1*DR4, DR1*DR15 P<0.0018
A1	0.00462					
*A2	1.37e-07	2.94e-08	1.41e-08	3.19e-08	3.01e-08	2.53e-08
C3	0.00517					
C5	0.01489					
Cn						0.007002

* represents the significant HLA class I variable based on criterion after correction.

Table 6.7 Comparison of different confounding models with Swedish data

	DR1, DR15 P<0.0026	DR1, DR4,DR15 P<0.0025	DR1,DR3, DR4,DR15 P<0.0024	DR1, DR4,DR15, DR1*Dr4 P<0.0024	DR1,DR4, DR15 DR1*DR4, DR1*DR15 P<0.0023	DR1,DR4,DR7,DR8, DR11,DR12,DR13, DR14,DR15, DR1*DR4, DR1*DR15 P<0.0018
*A2	1.83e-06	2.95e-06	1.6e-06	3.72e-06	3.51e-06	1.22e-06
C5	0.02202	0.029066	0.0284	0.038039	0.041233	0.053
C7	0.04108					0.0062
*Cn	0.00199	0.000826	0.000845	0.000852	0.000853	0.00069

* represents the significant HLA class I variable based on criterion after correction.

Table 6.8 Comparison of different confounding models with Norwegian data

	DR1, DR15 P<0.0026	DR1, DR4,DR15 P<0.0025	DR1,DR3, DR4,DR15 P<0.0024	DR1, DR4,DR15, DR1*Dr4 P<0.0024	DR1,DR4, DR15, DR1*DR4, DR1*DR15 P<0.0023	DR1,DR4,DR7, DR8,DR11,DR12, DR13, DR14, DR15, DR1*DR4, DR1*DR15 P<0.0018
A2	0.0167	0.019333	0.01648	0.01875	0.01866	0.0133

7 Conclusions

The statistical analysis of the genotype data of HLA polymorphisms in MS revealed that HLA class I gene polymorphisms are associated with MS in Swedish population independent of the class II association. The HLA-A2 locus is associated with reduced risk of MS in Swedish and Norwegian combined data. In addition, association of the HLA-Cn locus is a new finding in the Swedish set. Interactions between HLA class II genes (HLA DR1*DR4 and DR1*DR15) were found in the combined data of Swedish and Norwegian set.

In methodology, recursive partitioning was applied to analysis of genotype data in HLA class II polymorphisms. Two statistical programs were used to grow the helix trees. Different pruning methods were performed to create a partitioning tree model. When we used the logistic regression model to test for association to HLA class I gene, the class II confounding model was first performed in genotype data for conditioning.

In previous statistical analysis using logistic regression, HLA-A2 at HLA class I and HLA-DR15 at HLA class II were associated with MS in Swedish population. The effects of these two loci were independent of each other. These results were also confirmed in the present study.

However, association of HLA-Cn with Swedish MS (983 MS patients), but no association with Norwegian MS was identified. It might be due to low power to detect a weak effect in the Norwegian data (only 546 MS patients). Alternatively, this finding suggests that there might be population heterogeneity that opens up the probability of further studies in the future.

Appendix

A.1 The binary probability model and conditional probability model

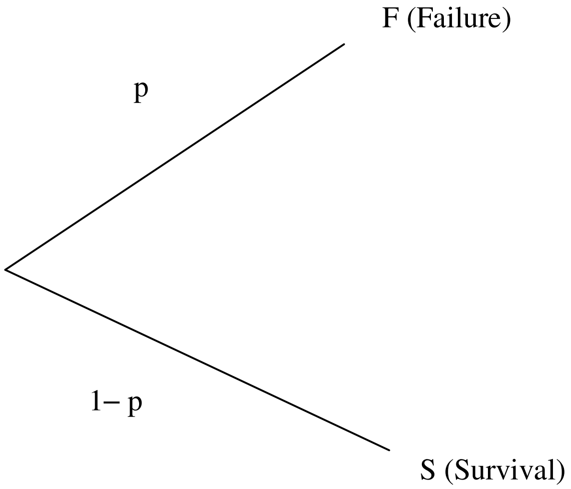


Figure A.1 The binary probability model. There are two possible outcomes. P is probability of Failure. The probability of survival is 1-p (Clayton et al. 1993 p7,)

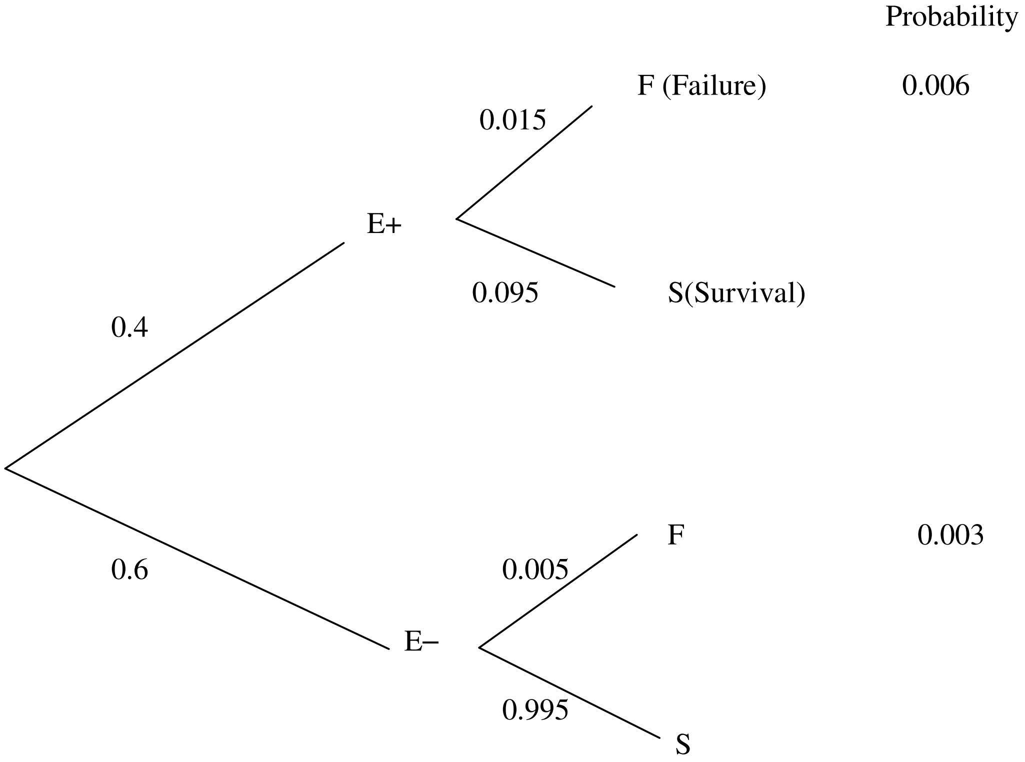


Figure A.2 A Conditional probability model (Clayton et al. 1993 p7,)

When the subjects are classified as either exposed (E+) or not exposed (E-), the conditional probability model can be presented as a tree with 6 branches. For example if 0.4 is the probability of exposure and 0.6 that of unexposure, the conditional probabilities are 0.015 (F) and 0.985(S) if a subject is exposed. The probability that a subject is exposed and fails is 0.4×0.015=0.006.

A.2 Case-control study and stratification

Case-control study

Case-control study is a type of observational study. Enrollment into the study is based on presence (``case") or absence (``control") of disease. Characteristics such as previous exposure are then compared between cases and controls.

		Disease		
		Yes	No	
Risk Factor	Yes	a	b	n_1
(Exposure)	No	c	d	n_2
		m_1	m_2	N

Confounding

Confounding is the distortion of the effect of one risk factor by the presence of another. Confounding occurs when another risk factor for a disease is also associated with the risk factor being studied but acts separately. Age, breed, gender and production levels are often confounding risk factors because animal with different values of these are often at difference risk of disease. As a result of the association between the group with the study risk factor and the control group without the study factor, the confounding is not distributed randomly between these two groups.

Correction for confounding

Confounding can be controlled by restriction, by matching on the confounding variable or by including it in the statistical analysis.

Stratification

The classical approach to experimentation is to hold constant all influence other than the experimental variable(s) of interest. For example, to avoid confounding by age, we would compare failure risk in exposed and unexposed subjects of a fixed age, or falling within a narrow range of ages. The statistical comparison would be made conditional upon age. This statistical analytical strategy is called stratification.

Reference

Almgren P, Bendahl PO, Bengtsson H, et al. (2003) Statistics in genetics. Lecture notes, Lund university, PP11.

Brynedal B, Kristina D, Jonasdottir, et al. (2007) HLA-A confers an HLA-DRB1 independent influence on the risk of multiple sclerosis. Plos one issue 7:e664

Barocellos LF, Sawcer S, Ramsay PP, Baranzini SE, Thomson G,et al.(2006) Heterogeneity at the HLA-DRB1 locus and risk for multiple sclerosis. Hum Mol Genet 15:2813-2824.

Cordell H and Clayton D (2002) A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. Am J Hum Genet 70:124-141.

Clayton D (1993) Statistical models in epidemiology. 1st Ed. Oxford science publications. pp273

Dyment DA, Herrera BM, Cader MZ, Willer CJ, Lincoln MR, et al. (2005) Complex interaction among MHC halplotypes in multiple sclerosis: susceptibility and resistance. Hum Mol Genet 14:2019-2026.

Englund Gunnar (2004) Computer intensive methods in mathematical statistics. Mathematic institution KTH, PP206-209

Ebers GC, Kukay K, Bulman DE,et al.(1996) A full genome search in multiple sclerosis. Nat Genet 13:472-476.

Fogdell A, Olerup O, Frerikson S, Vrethem M, Hillert J (1997) Linkage analysis of HLA class II genes in Swedish multiplx families with multiple sclerosis. Neurol 48:758-62.

Fogdell_hahn A, Ligers A, Gronning (2000) Multiple sclerosis: a modifying influence of HLA class I genes in an HLA class II associated autoimmune disease. Tissue Antigens 55:140-148.

Hosmer DW and Lemeshow S (2000) Applied logistic regression. 2st Ed. A wiley interscience publication. pp92

Jersild C, Fog T, Hansen GS., Thomsen M, Svejgaard A, Dupont B (1973) Histocompatibility determinants in multiple sclerosis, with special reference to clinical course. Lancet ii: 1221-1225.
Kuokkanen S, Gschwend M, Rioux JD et al.(1997) Genomewide scan of multiple sclerosis in Finnish multiplex families.Am J Hum Genet 61:1379-87.

Kurtzke JF (1983) Epidemiology of multiple sclerosis. Multiple sclerosis. Edited by Hallpike JF, Adams CWM, Tourtellotte WW.london: Chapman and Hall, pp47-95.

Lassmann H, Zimprich F, Rössler K, Vass K (1991) Inflammation in the nervous system. Basic mechanisms and immunological concepts. Rev Neurol 147:763-781.

Link H, Müller R (1971)Immunoglobulins in multiple sclerosis and infections of the nervous system. Arch Neurol 25:326-44.

Link H, Sun JB, Wang Z, Xu Z, Löve A, Fredrikson S, Olsson T(1992)Virus-reactive and autoreactive T cells are accumulated in cerebrospinal fluid in multiple sclerosis. J Neuroimmunol 38:63-73.

Möller JR, Johnson D, Brady RO, Tourtellotte WW, Quarles RH (1989)Antibodies to myelin-associated glycoprotein (MAG) in the cerebrospinal fluid of multiple sclerosis patients. J Neuroimmunol 22:55-61.

Nejentsev S, Howson JMM, Walker N, et al. (2007) Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A. Nature 450;887-892.

Olerup O, Hillert J(1991)HLA class II-associated genetic susceptibility in multiple sclerosis: a critical evaluation.Tissue Antigens.38:1-15.

Olerup O, Aldener A, Fogdell A(1993)HLA-DQB1 and -DQA1 typing by PCR amplification with sequence-specific primers (PCR-SSP) in 2 hours.Tissue Antigens 41:119-34.

Olsson T, Baig S, Höjeberg B, Link H (1990) Antimyelin basic protein and antimyelin antibody-producing cells in multiple sclerosis. Ann Neurol 27:132-6.

Poser CM, Paty DW, Scheinberg L et al. (1983) New diagnostic criteria for multiple sclerosis: guidelines for research protocols.
Ann Neurol 13:227-31.

Poskanzer DC, Prenney LB, Sheridan JL et al(1980) Multiple sclerosis in the Orkney and Shetland Islands. I: Epidemiology, clinical factors, and methodology. J Epidemiol Community Health.34:229-39.

Rao DC, (1998) Cat scans, pet scans and Genomoc Scans. Genetic Epidemiology 15:1-18.

Sawcer S, Jones HB, Feakes R et al(1996) A genome screen in multiple sclerosis reveals susceptibility loci on chromosome 6p21 and 17q22.Nat Genet 13:377-8.

Sun JB, Olsson T, Wang WZ et al.(1991) Autoreactive T and B cells responding to myelin proteolipid protein in multiple sclerosis and controls.Eur J Immunol 21:1461-8.

Sung Yoon Choo (2007) The HLA system: Genetics, immunology, clinical testing, and clinical implications. Yonsei Medical Journal 48:11-23.

Tienari PJ, Wikström J, Koskimies S (1993) Reappraisal of HLA in multiple sclerosis: close linkage in multiplex families. Eur J Hum Genet 1:257-268.

Yu K, Xu J, Rao D (2005) Using tree-based recursive partitioning methods to group haplotypes for increased power in association studies. Ann Hum Genet 69,577-589.

Yeo Tw, De Jager PL, Gregory SG et al. (2007) A second major histocompatibility complex susceptibility locus for multiple sclerosis. Ann Neurol 61:228-36.

Zaykin D and Young S (2005) Large recursive partitioning analysis of complex disease pharmacogenetic studies. II. Statistical considerations. NIH public access 6: 77-89

Zhang HP and Bonney G (2000) Use of classification trees for association studies. Genetic Epidemiology 19:323-332

Lecture notes 3 “classification trees”, *www.myoops.org*