Testing for isotonic inference in genetics

H. El Barmi, D. Mazzaro, F. Pesarin, L. Salmaso

2000.12

Dipartimento di Scienze Statistiche
Università degli Studi
Via S. Francesco, 33
35121 Padova

Settembre 2000

Testing for isotonic inference in genetics

Working Paper no. 2000.12

Department of Statistics, University of Padova

El Barmi H.[†]**, Mazzaro D.**[‡]**, Pesarin F.**[‡]**, Salmaso L.**[‡]

† Department of Statistics, Kansas State University, United States.

‡ Department of Statistics, University of Padova, Italy.

Table of Contents

INTRODUCTION

The major task of genetic epidemiology and molecular genetics is to map the diseases loci, that is the identification of genes causing the pathologies. However, the more common inherited disorders are very difficult to study, because a combination of various genes and different environmental factors is often involved. Discovering the major susceptibility locus can be the starting point to advance in understanding the causes of a disease. Furthermore, recently, the primary topic of interest has shifted from simple Mendelian diseases, where genotypes of some gene cause them, to more complex diseases, where genotypes of some set of genes together with environmental factors merely alter the probability that an individual has the disease, although individual factors are typically insufficient to cause the disease outright. We may use, to study these candidate genes and their relations, either linkage analysis or allelic association analysis (or linkage disequilibrium analysis).

The goal of linkage analysis is to determine whether two loci segregate independently in meiosis. Alleles of loci on different chromosomes segregate independently of each other during meiosis, as do alleles of loci on opposite ends of the same chromosome. However, when two loci are close together on the same chromosome, their alleles no longer segregate independently but are co-inherited more than 50% of the time. We say these loci are linked. The closer the two loci are to each other on the chromosome, the lower the probability of recombination of their alleles. This probability is referred to as the recombination fraction θ. The genetic distance is defined to be infinity between loci on different chromosomes, and for such unlinked loci $\theta = 0.5$. For linked loci on the same chromosome, $\theta < 0.5$, and the genetic distance is an increasing function of θ. The essence of linkage analysis is to estimate the recombination fraction θ and to test whether $\theta = 0.5$.

The terms "linkage disequilibrium" and "allelic association" are sometimes used interchangeably, and, sometimes, different meanings are assigned to them. The most general definition of either is the condition in which alleles of two loci on a random chromosome from the population do not occur independently of one another. Linkage disequilibrium is sometimes used only when the two loci are tightly linked and not when such correlations exist between unlinked loci, as may occur, for example, as a result of population stratification.

We use the term "linkage disequilibrium" irrespective of whether or not the loci are linked. The term association (without allelic) is also used to refer to the correlation between the alleles of a locus and some phenotype. Here, therefore, we use the terms "allelic association" and "linkage disequilibrium" to refer to the correlation between alleles of two loci on haplotypes. As in linkage analysis, the goal of linkage disequilibrium analysis is to map loci relative to each other and thereby to estimate the genomic position of new loci of unknown position using loci of known location. There are many measurements of linkage disequilibrium between two loci (for example, locus 1 and locus 2), but the most commonly used is the disequilibrium coefficient $D = P_{11} - p_1 q_1$, where P_{11} is the observed frequency of the 1/1 haplotype (generally, the "1" is the most common allele present in that locus), p_1 is the frequency of the "1" allele at locus 1 in the general population and q_1 is the population frequency of the "1" allele at locus 2. The coefficient D ranges from –0.25 (linkage equilibrium) to 0.25 (linkage disequilibrium). It was shown that the rate of decay of linkage disequilibrium is dependent on the distance between loci: $D_t = D_0(1- \theta)^t$, where t is the current generation number, D_t is the current amount of disequilibrium and D_0 is the disequilibrium at generation 0.

Linkage analysis is generally conducted on pedigrees of known structure, whereas linkage disequilibrium analysis is most often conducted on populations, which can be viewed as extremely large pedigrees with many generations of indeterminate structure. Allelic association analysis is used to locate regions of the genome shared by affected individuals more often than by a random sample of individuals from the population, because it is hypothesized that affected individuals share their phenotype because they also share some disease-predisposing allele identical by descent from a common ancestor. Thus allelic association analysis is a form of linkage analysis on the largest possible hypothetical pedigree (Terwilliger and Göring, 2000).

Risch and Merikangas (1996) argue that the method which has been used successfully (linkage analysis) to find major genes has limited power to detect genes of modest effect, but that a different approach by association studies that utilizes candidate genes has far greater power, even if one needs to test every gene in the genome. Thus, they say that the future of the genetics of complex diseases is likely to require large-scale testing by association analysis.

Allelic association studies may be "population-based" or "family-based": the former is essentially performed over the comparison between one sample (cases) of patients and one sample (controls) of unrelated unaffected individuals (it exists also the situation where they are not unrelated); the latter is performed over a set of family unities composed, at least, of one affected individual (there are many types of family-based association analysis).

Introduction

The purpose of this work is to illustrate a new statistical approach to test allelic association and genotype-specific effects in the genetic study of a disease. There are some parametric and non-parametric methods available for this end. We deal with population-based association studies, but comparisons with other methods will be performed too, analysing advantages and disadvantages of each one, specially regarding to power properties with small sample sizes. In this framework, we will work out some nonparametric statistical permutation tests and likelihood-based tests to perform case-controls analyses to study allelic association between marker, disease-gene and environmental factors. Permutation tests, in particular, will be extended to multivariate and more complex studies, where we deal with several genes and several alleles together. Furthermore, we show some simulations under different assumptions on the genetic model and we analyse real data sets with the simple study of one locus with the permutation test.

We think these arguments could help the researchers to decide the best test to use, particularly with regard to complex genetic problems and unusual systems of hypotheses, especially in presence of small sample sizes.

1. ASSOCIATION STUDIES IN GENETICS

1.1 Case-Control Studies in Genetics

The case-control method is usually applied in genetic epidemiology to elucidate the role of genetic factors and their interaction with environmental factors in the aetiology of human diseases. Now, the map of human genome will make it increasingly feasible to search for disease susceptibility genes using case-control methods in both population and family settings. Actually, the interest of epidemiology concerns with the relation between the environment (in its more general meaning) and the occurrence of human diseases, while that one of genetics concerns with evaluating the effects of population structure and selection forces on the frequency of genetic traits. So, of course, the primary purpose of genetic epidemiology is in studying genetic variation in human populations and its relation to normal and pathologic phenotypic variation. Genetic epidemiologists need to evaluate the distribution and determinants of genetic traits in human populations and address the role of genetic factors and their interaction with environmental factors in the aetiology of human diseases. Both population and family methods are used to achieve this goal.

Case-control methodology may not be applicable in all settings and should always be integrated with family studies using genetic analytic techniques such as segregation and linkage methods.

In population studies, case-control approaches are used 1) to study determinants of human mutations, 2) to evaluate the role of non-specific genetic indicators (such as inbreeding and racial admixture) in the etiology of diseases, and 3) to assess the role of specific genetic traits in the etiology of diseases. They are particularly useful in the study of mutations, because most mutations are individually rare and their ascertainment involves a combination of clinical and laboratory testing, which makes cohort studies prohibitively expensive.

An important aspect to consider in these studies is the presence of confounding factors. Confounders could be other unmeasured genetic determinants or environmental factors that could produce spurious differences in allele frequencies between cases and controls. Several confounding factors have been recognised in this type of study, which also make comparisons between investigations difficult. First, there are phenotypic differences in cases due to several definitions for cases and controls in different studies, to subliminal differences in enrolled cases because of variations in investigators' clinical skills and to phenotypic heterogeneity of the disease. Second, genetic background of cases and controls are not identical (specifically in multiracial areas). Third, there are practical difficulties due to low number of cases, lack of specificity and methodological artefacts.

Different components can make unclear the correct methodological and statistical procedure to follow in clinical investigators. *A)* a complex disease is a multistage process where several genetics or environmental events mark each stage and can interact between them. *B)* Interindividual variations in response to environmental factors, due to genetic heterogeneity of populations, are present in the data, so a specific polymorphism determines the type of the response to exposure to a specific environmental factor. *C)* There is a biological relation between the environmental factors, the particular polymorphism and one (or more) of disease phenotypes. *D)* Under special conditions, an allele can be no longer neutral.

If the two first assumptions can be checked by direct data observation, the latter two are not supported by any clear-cut evidence. Sometimes, a phenotype is considered indication of the disease when it is not the case. If linkage with the disease or intermediate phenotype of the disease emerges from linkage studies, this finding would strongly support the idea that the candidate gene is in some way involved with the disease. Further, with respect to the latter point, before to use linkage association studies, it should be demonstrated that an allelic variant of the gene analysed is non-neutral. Clinical investigators must allow for several points in genetic studies. A gene has not only one allelic polymorphism, but there could be more of them, so it is dangerous to discard a candidate gene because one of its allelic variants is not found in association with the disease. If a polymorphism is related with the disease, we are not sure that it is the disease allele, because it could be in strong linkage disequilibrium with that allele. Because of the, so-called, founder effect, allelic association studies can be misleading if they are done on different populations. A positive association allele-disease can be found in a specific population and several negative studies in different populations could not reject the first result. It is very important to select rigorously case-control subjects to guard against possible confounding effects (Gambero *et Al.*, 2000). To minimize confounding in case-control studies, investigators need to carefully select controls from the same racial/ethnic genetic background as that from which cases are derived. Relative controls have been used in an attempt to match for genetic background. Analyses should always be stratified with regard to potentially confounding variables. For example, in Down's syndrome, advanced maternal age is the most strictly factor, therefore, in examining the potential association between a risk factor and the syndrome, possible confounding by that factor should always be considered. In case-control methods, therefore, to plan the correct study design can address the confounding: for instance, in the previous situation, by matching cases and controls by maternal age, or by stratification in the analysis.

When we want to evaluate the role of specific genes in the etiology and pathogenesis of common diseases such as cancer, coronary heart disease, birth defects, etc., we are searching for correlations between specific alleles and diseases, so we need the so called "association studies" in human genetics. They are different from "linkage studies", in which evidence is sought for cosegregation between a marker locus and a disease in families. In effect, Greenberg showed, by computer simulations, that if the disease frequency among persons with the susceptibility allele is less than 10 times greater than the disease frequency among persons without this allele, it may be quite difficult to detect linkage even in data sets consisting of 30 nuclear families with two or more affected individuals. Under these conditions, the usual linkage approaches may lack sufficient statistical power to detect linkage or may get false rejection of linkage hypothesis and suffer from the multiple testing problem. This is the primary reason for increasing usefulness of case-control association methods to look for genetic risk factors.

Case-control evaluation of genetic traits in disease aetiology is generally guided by a "candidate" gene approach, which refers to examining allelic variation (measured either at the protein level or at the DNA level) in loci known or suspected to have some role in the pathogenesis of the disease.

In designing, analysing and interpreting case-control studies of genetic trait-disease associations, it is important to consider several methodological issues. The primary problem is that the causes of many simply and complex diseases are related to confused interactions between genetic susceptibility and environmental factors. Case-control studies provide, in this context, an efficient tool wherewith to search for genetic susceptibility factors along with environmental exposures. Many types of patterns of gene-environmental interaction had been discussed: additive, multiplicative, etc.; and let be stress the importance of power and sample size considerations for case-controls studies of these gene-environmental interactions (Foppa and Spiegelman, 1997).

As in case-control studies of epidemiology, it is very important how we choose the samples of patients and of unrelated healthy individuals. If we are studying a most uncommon disease, we probably choose, for cases, all the patients in the same region we have, and who present the same feature with regard to environmental factors and, if necessary, confounding effects. For control subjects, we can go by many types of selection according to the specific situations we are: hospital control series, random selection, friends of cases (we assure the same environmental factors), etc. At the beginning, we may choose more than one control group to see the different results we obtain, but after we must decide only one group to have a suitable study, that is more appropriate for characteristics of the disease, of cases subjects and of the other factors (Wacholder *et* Al., 1992).

One of the major issues in the design of case-control studies concerns the size of the study. If we are bound by the frequency of the disease with regard to cases, controls size is joined to the number of specific environmental and confounding factors that play a role in the influence of the disease, but, most of all, to the statistical powerful of the test we choose (Smith and Day, 1984).

Gene-disease association studies that fail to examine the role of environmental exposures along with the genetic traits of interest may lead to considerable dilution in measures of association if the genetic factor confers disease susceptibility only in the presence of other genes or environmental determinants. Therefore, in designing case-control studies in genetic epidemiology, environmental risk factors should be examined along with genetic markers of interest as interacting factors with the genetic factor of interest.

In studying associations between genetic traits and disease, indirect methods are often used to assign individuals' genotypes. Such indirect measurement of the underlying genotype can lead to non-differential genotypic misclassifications, and therefore would dilute the magnitude of the relative risks found. Nevertheless, genotypic misclassification can arise when the disease itself interferes with genotypic classification. If genotypes are measured at the DNA level, misclassification due to linkage disequilibrium can occur. Under ideal conditions, if the gene of interest has been completely sequenced, the presence and location of one or more mutations within the gene could be correlated with an altered gene product and then with case-control status in epidemiologic studies. However, in many of these studies, the researchers only have markers in the general region of the candidate gene or in a non-expressed portion of that one.

Unless the actual site of a deleterious mutation involved in the disease is targeted, any DNA variation between cases and controls in the region of a candidate gene could reflect DNA variation in linkage disequilibrium with the actual mutation(s) associated with the disease. Linkage disequilibrium can arise when the mutation has occurred relatively recently or if there is selective advantage for specific haplotypes, so that they are preferentially maintained in the population. Under complete equilibrium, there would be no association between any marker allele and the disease susceptibility allele (i.e., the odds ratio should be 1 in a case-control study). Under linkage disequilibrium, a marker allele may well occur more often with the disease susceptibility allele. However, the association between the marker allele and the disease susceptibility allele may not be perfect, and thus, if the marker allele is used as a proxy for the susceptibility allele to study disease risk, some non-differential misclassification could easily occur. This would dilute the magnitude of the odds ratio between the marker allele and the disease toward the null, and would underestimate the effect of the genetic locus in the aetiology of the disease. One analytical approach wherewith to address the issue of linkage disequilibrium in case-control studies is to construct specific haplotypes composed of alleles at tightly linked loci within the area of the candidate gene.

A big statistical importance in these studies concerns with the type I and type II errors. In case-control studies involving many genetic traits and other risk factors, statistically significant associations can due to chance. These type I errors will be increasingly important in case-control studies involving multiple DNA markers at many candidate loci. As researchers sequence more genes and as DNA polymorphisms are delineated throughout the genome, a major challenge in genetic epidemilogy will be to discriminate the biologically meaningful associations from the multitude of spurious ones. The establishment of a cause-effect relation depends on many issues, including consistency of the association across studies and the presence of a biologically meaningful model underlying such associations. Finally, to address issues related to statistical power (type II errors), investigators must ensure adequate sample sizes in designing case-control studies to search for causal genetic factors, especially to look for evidence of gene-environment interactions (Khoury and Beaty, 1994).

1.2 Other Methods

The main shortages of genetic case-control association studies are the lack of large numbers of patients with a condition of interest, the lack of an adequate control group and the ethnic heterogeneity. Erroneous results may occur because of several confounding factors, which can present each one at a time or together, as the population stratification (founder effect), multiple hypothesis testing and sub-group analysis. It is important to point out that in these studies any positive association should be reproduced in large cohorts and be tested for linkage in family based studies.

An alternative but related study design is to collect "trios", consisting of two parents (irrespective of phenotype) with one affected offspring. The case sample consists of the alleles or haplotypes that were transmitted from the parents to an affected child, whereas the control sample consists of those alleles that were not transmitted to the affected child. The key advantage of this so-called haplotype relative risk (HRR) design is that it ensures that case and control samples come from the same genetic population. The statistical methods for analysis of both study designs are similar for 2-point methods but may differ somewhat in multipoint analysis, because the HRR design sometimes provides a means of reconstructing multilocus haplotypes, whereas case-control analysis provides only genotype information.

The most used method which uses "trios" and that demonstrated very powerful of showing both association and linkage is the transmission disequilibrium test (TDT), which requires DNA from an affected patient and its parents, and which examines the transmission of alleles from an heterozygous parent to the affected offspring. A significant difference from the expected Mendelian ratio of 50 : 50 would suggest that the allele has a role in the susceptibility to the disease in question. The TDT was been proposed by Spielman et al. (1993) in response to the problem of spurious associations. This approach takes advantages of population-level associations but it is not susceptible to spurious associations that result from stratification. When applied exclusively to trios, the TDT is equivalent to a valid McNemar test of linkage disequilibrium. Risch and Merikangas (1996) recommended allelic association studies as the study design of choice. Allelic association analysis can be powerful when the affected individuals in a sample share the same allele identical by descent at the same disease locus from some common ancestor (Terwilliger and Göring, 2000).

Although very elegant, the TDT design is usually more labour intensive than a simple case-control design that uses affected individuals and unrelated controls. It may take considerable effort, or may even be impossible, to collect DNA samples from the parents of probands, particularly for late-onset diseases. It may also be difficult to collect DNA from other relatives for which TDT-like statistics have been proposed (Boehnke and Langefeld 1998, Lazzeroni and Lange 1998, Spielman and Ewens 1998). For this reason the simple case-control approach would often be an attractive study design, were it not for the problem of spurious associations due to population stratification. Still, Pritchard and Rosenberg (1999) showed that the case-control design can be a valid test for association if we include an explicit test for stratification. If we use only a few marker loci, the possibility that the association are due to population stratification cannot be eliminated, however, by typing additional unlinked markers, it is possible. Their basic idea is that if stratification is present, the unlinked markers must also show association with the phenotype.

Another method, very useful with respect to the TDT in several diseases (Schaid), is to use the affected sib-pair approach, which has been utilised successfully in many research works (for example into type-1 diabetes). The last methods need multi-centre collection of families and large cooperative groups, but they represent the way forward in unravelling the complex genetics of polygenic diseases. Of course, this type of studies can result too much expensive and very slow, specially, if we consider the fact that, sometimes, their results could be negative or, at least, far from that we expected. There could be the objection that a study, wherefrom we did not obtain any significant positive result, from a strictly statistical point of view, it is not "unsuccessful" or "superfluous", on the contrary, it has as much information and importance than another study that leads to expected conclusion, in particular, in point of possible future studies in the same field. Nevertheless, from a more practical point of view, it is undeniable that such an employment of resources and such a strain could be considered "excessive" with respect to a possible "negative" conclusion. Therefore, case-control studies still have a role in hypothesis testing, but they must involve large numbers to provide meaningful results (Chowdhury, 2000). However, association methods, in many cases, had had modest results in the study of genetic polymorphisms and complex diseases, and some authors attribute this fact almost entirely to the incompetence of clinical researchers and their lack of understanding of basic genetic principles (Cheung and Kumana, 2000).

In spite of the problems of case-control association studies, as the population heterogeneity, they are appealing because they do not require additional family members for cases, which can be very expensive. Then, Devlin and Roeder (1999) developed a method that has the advantages of both case-control and family-based designs. They proposed such a method for either SNP (single nucleotide polymorphism) association scans or tests of candidate genes. They use for case-control data the genome itself to induce controls similar to family-based studies and to determine what constitutes a significant departure from the null model of linkage disequilibrium. An advantage of dense association genomic scans is that they can detect loci having a small impact on risk to human disorders, while a disadvantage is the large number of false positives occurrences when many significance tests are conducted. Instead of a traditional Bonferroni correction they proposed a Bayesian outlier test as a means of determining which markers exhibit significant linkage disequilibrium with the disorder, that is, the outlier test bypasses the usual rigid assumptions required to obtain chi-square distributed random variables in favour of more flexible statistics and weaker assumptions. Another feature of their methodology is that it allows for violations in the usual model assumption, the independence of observations, which, when violated, leads to extra variance in the (parametric) test statistic, indeed, for case-control studies, affected individuals are more likely to be related than are control individuals because they share a genetic disorder and, ideally, a common genetic basis for the disorder (founder effect). For this reason, simple marker-by-marker hypothesis tests will almost surely produce false positives, even after a Bonferroni correction. However, their simulations are not so decisive, moreover, the situations they considered are quite particular.

1.3 Case-Control Design Study

Allelic association may be due to pleiotropy, linkage disequilibrium, meiotic drive, selection or population stratification. Somebody, talking about association analysis, distinguishes between model-free methods from model-based methods. The use of case-control studies to detect an association relapses in the latter studies. Classical case-control studies are important in genetic epidemilogy, even though they can only establish an association and other designs are necessary to determine whether such associations are casual. In tabulating such data, a question arises as to whether one has one or two observations per person. One approach classifies individuals according to their genotypes, that is, according to the pair of alleles that each individual has. Another classifies each allele. Intuitively, one might feel that, provided the alleles are independent, either approach should give a valid analysis, but Sasieni (1997) showed that this is not true.

The data appear as a standard Fisherian 3 * 2 or 2 * 2 table for which chi-squared statistics and odds ratios were developed. Table 1 presents the number of cases (subjects with the disease) and controls with 0 (negative), 1 (heterozygous) and 2 (homozygous) copies of the rare allele *A*.

Table 1

Genotypes	Cases	Controls
aa	x_1	y_1
aA	x_2	y_2
AA	x_3	y_3

Since each heterozygous individual has one copy of *A* and each homozygous has two copies of that one, we can produce an allele table with twice sample size (Table 2).

Table 2

Subjects	Cases	Controls
A	$x_2 + x_3$	$y_2 + y_3$
Other	x_1	y_1

Finally, Table 3 presents the data in terms of the number of subjects with and without the rare allele *A*, ignoring the difference between homozygous and heterozygous genotypes. Such a tabulation is common when it is not possible to distinguish heterozygous from homozygous individuals (a situation of perfect dominance or recessiveness).

Table 3

Allele	Cases	Controls
A	$2x_3 + x_2$	$2y_3 + y_2$
Other	$X_2 + 2x_1$	$y_2 + 2y_1$

Traditionally, one estimates odds ratios from case-control data because it is not possible to estimate the risk of diseased in each exposure group directly. Instead, one relies on the identity between the ratio of odds of exposure in the diseased to that in the controls and the ratio of the odds of disease in the exposed to that in the unexposed. Provided the disease is rare, the odds ratio will be a close approximation to the relative risk. With genotype data, one can estimate the relative risk of a rare disease associated with the heterozygous genotype and with the homozygous genotype, or one could combine these two groups (as is done in Table 3) and estimate the relative risk associated with the gene. Formulas for these estimators are given in Table 4.

Table 4

Table	Odds ratio	Formula
1: hetero	θ_{hetero}	$(x_1 \cdot y_2)/(x_2 \cdot y_1)$
2: homo	θ_{homo}	$(x_2 \cdot y_3)/(x_3 \cdot y_2)$
3: allele	θ_{allele}	$[(2x_3 + x_2) \cdot (y_2 + 2y_1)] / [(x_2 + 2x_1) \cdot (2y_3 + y_2)]$
4: serological	θ_{sero}	$[(x_2 + x_3) \cdot y_1] / [x_1 \cdot (y_2 + y_3)]$

The odds ratio from allele data is the relative odds of the allele in cases and in controls. For a rare allele, this is approximately the relative gene frequency in cases and controls. It is not however immediately obvious how to translate this odds ratio into a statement about the risk of disease. Whereas one can discuss the risk of disease in an individual with a given genotype, it does not make sense to talk about the risk of an allele getting the disease. The best we can do is to say that the known allele is chosen at random. By contrast, the odds ratio from the serological table does have a reasonable interpretation. For a rare disease, it will give the relative risk of disease for an individual (chosen at random from among all individuals) with at least one copy of the allele. Thus, we need not assume that homozygotes and heterozygotes have the same risk. The serological odds ratio is appropriate whenever we do not have information to distinguish homozygotes from heterozygotes.

There is, however, a special case under which the allelic odds ratio will coincide with the genotypic odds ratio. Suppose that the Hardy-Weinberg equilibrium holds in both cases and controls; that is, the relative proportions of the different genotypes is $p_i^2 : 2p_i(1-p_i) : (1-p_i)^2$, $i = 1, 2$, where p_1 and p_2 are the allelic frequencies of the more common allele in cases and controls, respectively. Recall that the equilibrium holds under the pair of assumptions of random mating and no selection. The assumption of no selection in cases implies that the gene is not associated with the disease, but the equilibrium could hold under weaker assumptions, too. Statistically, the Hardy-Weinberg equilibrium simply states that the alleles are independent.

Sasieni (1997) showed that it is not recommendable the use of the allelic odds ratio and chi-squared statistic, even when it is possible to assume that the effect of different alleles at a given locus are codominant. Indeed, these statistics are not robust against departures from the assumptions of Hardy-Weinberg equilibrium in controls and codominance between the alleles.

1.4 The problem of allelic association analysis

Now, suppose to have two random samples, one of M cases (individuals with disease), and one of N controls (without the disease), where each person is classified as having a particular marker allele (a, the more common or A the rarer). Indicate by x_1, x_2, x_3 the numbers of affected individuals who carry (respectively) zero, one or two copies of rare allele, while by y_1, y_2, y_3 the corresponding control subjects. So we obtain the following 3×2 contingence table:

Marker	**Cases**	**Controls**	**Total**
aa	x_1	y_1	$S_1 = x_1 + y_1$
aA	x_2	y_2	$S_2 = x_2 + y_2$
AA	x_3	y_3	$S_3 = x_3 + y_3$
Total	$M = x_1 + x_2 + x_3$	$N = y_1 + y_2 + y_3$	$S = M + N = S_1 + S_2 + S_3$

so that the odds ratio

$$P(\text{disease}|\ aa)/P(\text{no disease}|\ aa)\ /\ P(\text{disease}|\ aA)/P(\text{no disease}|\ aA)$$

and

$$P(\text{disease}|\ aA)/P(\text{no disease}|\ aA)\ /\ P(\text{disease}|\ AA)/P(\text{no disease}|\ AA)$$

or, equivalently,

$$P(aa| \text{ disease})/P(aA| \text{ disease}) \,/\, P(aa| \text{ no disease})/P(aA| \text{ no disease})$$

and

$$P(aA| \text{ disease})/P(AA| \text{ disease}) \,/\, P(aA| \text{ no disease})/P(AA| \text{ no disease})$$

are consistently estimated, respectively, by $\theta_{aA} = (x_1 \cdot y_2)/(x_2 \cdot y_1)$ and $\theta_{AA} = (x_2 \cdot y_3)/(x_3 \cdot y_2)$. Significance of the deviation of these ratios from 1 can be tested by the usual chi-square statistic with 1 degree of freedom or, for small samples, by the exact test of Fisher. If the controls are obtained by a random sample from the population, rather than a sample of persons without the disease, then θ_{aA} and θ_{AA} are consistent estimations of the more meaningful *relative risks* (Elston, 1998).

In genetic epidemiology of diseases of complex etiology, association studies are useful to investigate candidate disease genes. Association studies are case-control population-based studies on a comparison of unrelated affected and unaffected individuals. An allele *A* at a gene of interest is said to be associated with the disease if it occurs at a significantly higher frequency among affected compared with control individuals. For a bi-allelic locus with common allele *a* and rare allele *A*, individuals may carry none (subjects with genotype *aa*), one (subjects with genotype *aA*) or double (subjects with genotype *AA*) copies of the *A* allele. Conventionally, therefore, a test for allelic association is to test for the distribution of case/control genotypes using the likelihood ratio chi-square statistic (asymptotically distributed as χ^2 with 2 df) or the Fisher exact test.

But testing only for overall effects of a gene rather than genotype-specific effects may be less powerful. For example, in a case-control study on the role of R353Q genetic variants of factor VII (a plasma protein involved in the blood coagulation) on myocardial infarction, Iacoviello et al. (1998) showed a great protection against myocardial infarction due to the rare genotype QQ (found in a 5% of controls but only in 0.6% of cases) but only a small difference in the distribution of RQ genotype (see Ch. 5).

It is therefore necessary to test for genotype-specific risks. However, this approach requires some attention as all models are not necessarily biologically plausible: the effect of an allele can be expressed only in one of the following way:

1) recessive – there is an effect only in the presence of two copies of *A* allele (genotype *AA*), while the heterozygous condition (genotype *Aa*) is the same like the reference and commonest condition (genotype *aa*).

2) codominant – there is an additive effect of the *A* allele: genotype *Aa* is of risk (or of protection) in comparison with the genotype *aa*, and *AA* is of risk (or of protection) in comparison with the genotype *Aa*. Obviously, *AA* is of great risk (or of great protection) in comparison with the genotype *aa*.
3) dominant – the effect of the *A* allele is the same in *AA* and *Aa* genotype. In this situation, there is no relative risk (or protection) between *AA* and *Aa*, but only between *AA* (or *Aa*) and *aa*.

For these reasons, differences in the risk should be tested for while maximizing over the restricted parameter space that corresponds to plausible biological models: ($R_{AA} \geq R_{Aa} \geq \mathrm{R_{aa}}$) or ($R_{AA} \leq R_{Aa} \leq R_{aa}$), where R_g are the genotype-specific risks.

In case-control studies it is easy to obtain genotype-specific relative risk from odds ratios: $\theta_{AA} = R_{AA}/R_{Aa}$ and $\theta_{Aa} = R_{Aa}/R_{aa}$ (of course $\theta_{aa} = R_{aa}/R_{aa} = 1$), and the null and alternative hypotheses became:

$$H_0\colon \theta_{AA} = \theta_{Aa} = 1$$

$$H_1\colon \{(\theta_{AA} \geq 1) \cap (\theta_{Aa} \geq 1)\}\ XOR\ \{(\theta_{AA} \leq 1) \cap (\theta_{Aa} \leq 1)\},$$

where at least one inequality is strong.

This particular system of hypotheses was proposed for the first time by Chiano and Clayton in 1998. From a statistical point of view, the alternative hypothesis is of isotonic type, that is the variables are ordered in one sense, however, there is the further complication due to the "*XOR*", that is an exclusive "or". This approach allows to study genetic diseases for which we do not known the relative effect of the putative allele (dominant, recessive or codominant) or, if we are studying a related genetic polymorphism that may be protective or deleterious with respect to the disease.

In the following, we deal with this particular statistical problem, by using different approaches.

2. The Nonparametric Permutation Methodology

2.1 Basic Concepts on the Theory of Permutation Tests

We introduce terminology, definition and the general theory of permutation tests. The permutation tests are essentially conditional procedures, where conditioning is made with respect to the permutation sample space associated with the whole data set, which is a set of sufficient statistics under the null hypothesis. It was shown (Pesarin, 2001) that this conditioning makes permutation tests invariant, under the null hypothesis, with respect to the underlying population distribution, which may be partially or even completely unknown. Consequently, permutation tests are distribution-free and nonparametric.

We denote by X a response random variable whose values are points of the sample space χ. The probability distribution P on χ, associated with a symbolic random experiment characterizing X, is defined on an additive class B of subsets of χ. Sometimes, associated with P and with respect to a dominating measure ξ, we may refer to the density f of χ. Here, χ is a one-dimensional Euclidean space and B is a family of Borel sets. A random sample from X is a random experiment whose result is a sample point $\mathrm{X}^n = \{X_1,\ldots,X_n\}$. Given a sample X of n i.i.d. observations from X, we wish to test the null hypothesis H_0 that the unknown probability distribution P on (χ,B) generating X belongs to a certain class P_0, against the alternative class P_1. To be precise, we denote the null hypothesis by H_0: $\{P\in \mathrm{P}_0\}$ and the alternative by H_1: $\{P\in \mathrm{P}_1\}$, where of course $\mathrm{P}_1 = \mathrm{P} - \mathrm{P}_0$. The sample point X takes values on the sample space χ^n. The most common situation is that P_0 contains only one element, the null hypothesis in this case is said to be simple, otherwise it is said to be composite.

We indicate by P^n the probability distribution induced on χ^n by the sampling experiment. Associated with any sample point $\mathbf{X}$ there is the orbit $(\chi^n|\mathbf{X})$, also called conditional sample space, containing all points of χ^n which are equivalent to the given sample point $\mathbf{X}$ with respect to a group of transformations characterized by suitable invariance properties. The invariance properties in question are that conditional distribution $P^n{}_{|\mathbf{X}}$ on points of the conditional sample space $(\chi^n|\mathbf{X})$ is not dependent on population distribution P $\forall P\in \mathbf{P_0}$ (see Pesarin, 2001).

> **Definition:** any test statistic T: $\chi^n \rightarrow \Re^1$, whose conditional c.d.f. $F_T(t|\mathbf{X})$ is induced by $P^n_{|\mathbf{X}}$ and is invariant under H_0 on $(\chi^n|\mathbf{X})$, is said to be an invariant test for testing H_0 against $H_{1.}$

However, as for any given testing problem we may condition with respect to different sets of sufficient statistics, we may also take into consideration various groups of invariant transformations. From this point of view, on one side we should condition with respect to a minimal set of sufficient statistics, on the other side we should take into consideration a group of maximal invariant transformations (see Pesarin, 2001).

It is important to consider that:

1) the conditional sample space $(\chi|\mathbf{X})$ has always a finite number of points, provided that the sample size n is finite;
2) denote by K the cardinality of $(\chi|\mathbf{X})$: $K = \#\{\mathbf{X}^* \in (\chi|\mathbf{X})\}$, where # means the number of point satisfying condition (.);
3) on $(\chi|\mathbf{X})$ we may define an algebra of events $(B|\mathbf{X})$ containing all sub-sets of interest, so that $\{(\chi|\mathbf{X}), (B|\mathbf{X})\}$ is a conditional measurable space;
4) for every event $A \in (B|\mathbf{X})$ we have that $\Pr\{A|\mathbf{X}\} = \int_A dP_{|\mathbf{X}}$.

Another important concept is the permutation equivalence of two statistics:

> **Definition:** Two statistics T_1 and T_2, both mapping χ into $\Re^1$, are said to be permutationally equivalent when, for all points $\mathbf{X} \in \chi$ and $\mathbf{X}^* \in (\chi|\mathbf{X})$, the relationship $\{T_1(\mathbf{X}^*) \leq T_1(\mathbf{X})\}$ is true if and only if $\{T_2(\mathbf{X}^*) \leq T_2(\mathbf{X})\}$ is true, where $\mathbf{X}^*$ indicates any permutation of $\mathbf{X}$ and $(\chi|\mathbf{X})$ is the conditional sample space.

Formally, the randomized version of the permutation test Φ_R associated with $(T|\mathbf{X})$ is defined as:

$$\Phi_R = \begin{cases} 1 \; if \; T_{ob} > T_\alpha \\ \gamma \; if \; T_{ob} = T_\alpha \\ 0 \; if \; T_{ob} < T_\alpha \end{cases}$$

where α is the significance level of the test T, T_{ob} is the observed value of statistic test, T_α is the critical value of the statistic and $\gamma=[\alpha - \Pr\{T_{ob} > T_\alpha|\mathbf{X}\}]/\Pr\{T^* = T_{ob}|\mathbf{X}\}$.

2.2 Sampling Inspection of Permutation Space

We first observe that, under H_0 and due to the assumed exchangeability of data with respect to symbolic treatment levels, all points of the conditional sample space $(\chi|\mathbf{X})$ are equally likely. Therefore, one way of inspecting into $(\chi|\mathbf{X})$ is by means of a Monte Carlo simulation. Among the different Monte Carlo techniques, the simplest is by means of simple random sampling.

Without loss of generality, hereafter we assume that univariate permutation test statistics T of interest are significant for large values. The permutation distribution of any test statistic T is denoted by the notation $F_T(z|\mathbf{X}), \forall\, z\in \Re^1$. A general simulation procedure for estimating the c.d.f. $F(z|\mathbf{X})$ and the associated p-value λ induced by a statistic T applied on data set $\mathbf{X}$ is described in the following steps:

1) calculate the observed value of T: $T_{ob} = T(\mathbf{X})$;
2) consider a data permutation $\mathbf{X}^*$ randomly selected from $(\chi|\mathbf{X})$, where all points of $(\chi|\mathbf{X})$ are equally likely, and consider the value of test statistic T on $\mathbf{X}^*$: $T^* = T(\mathbf{X}^*)$;
3) independently, repeat step 2) B times; the set of CMC-Iterations results $\{T^*, i = 1,\ldots,B\}$ is thus a random sample from the permutation distribution of T;
4) the E.D.F. $\hat{F}_B^*(z) = \sum_{i=1}^{B} I(T_r^* \le z)/B, \forall z \in \Re^1$, where I (.) = 1 if relation (.) is true and 0 otherwise, is a consistent estimate of the permutation distribution $F(z|\mathbf{X})$ of T; moreover: $\hat{\lambda} = \sum_{r=1}^{B} I(T_r^* \ge T_{ob})/B$ is an unbiased and consistent estimate of the permutation p-value λ = $\Pr\{T^* \ge T_{ob}|\mathbf{X}\}$;
5) if, for any fixed significance level α, the result is $\hat{\lambda} < \alpha$, then reject H_0.

The following figure summarizes the conditional Monte Carlo procedure (CMC-Procedure):

Fig. 1: The Conditional Monte Carlo Procedure.

$\mathbf{X}$	$\mathbf{X}_1^*$	$\cdots$	$\mathbf{X}_r^*$	$\cdots$	$\mathbf{X}_B^*$
T	T_1^*	$\cdots$	T_r^*	$\cdots$	T_B^*

In statistics, usually the researcher works with complex problems that involve hypothesis systems which can be decomposed into more sub-problems with simplest systems of hypotheses: the global null hypothesis becomes the intersection of all partial null hypotheses, while the alternative global hypothesis corresponds to the union of all partial alternative hypotheses. In these situations the use of nonparametric combination methodology for dependent tests shows be very useful and efficient to obtain a good solution.

2.3 The Nonparametric Combination of Dependent Tests

We can consider a set of generic partial tests $\{T_i,\ i = 1,\ldots,k\}$. Let us assume that the following assumptions are satisfied.

1) All permutation partial tests T_i must be marginally unbiased and significant for large values, so that they are stochastically larger under H_1 than under H_0.

These assumptions, formally, mean that $Pr\{T_i \geq T_{i\alpha}|\mathbf{X},H_{1i}\} \geq \alpha, \forall\, \alpha > 0$, $i = 1,\ldots,k$, and $Pr\{T_i \leq z|\mathbf{X},H_{0i}\} = Pr\{T_i \leq z|\mathbf{X},H_{0i} \cap H_i^{+}\} \geq Pr\{T_i \leq z|\mathbf{X},H_{1i}\} = Pr\{T_i \leq z|\mathbf{X},H_{1i} \cap H_i^{+}\}$, $i = 1,\ldots,k$, $\forall\, z \in \Re^1$, where irrelevance with respect to the complementary set of hypotheses $H_i^{+}\!: \{\bigcup_{j\neq i}(H_{0j} \cup H_{1j})\}$ means that it does not matter which among H_{0j} and H_{1j}, $j \neq i$, is true when testing for the i-th sub-hypotheses.

2) Partial tests T_i must be consistent, that is: $Pr\{T_i \geq T_{i\alpha}|H_{1i}\} \rightarrow 1, \forall\, \alpha > 0$, $i = 1,\ldots,k$, as n tends to infinity, where $T_{i\alpha}$, which is assumed to be finite, is the marginal critical value T_i.

These assumptions, especially the former, imply that the set of p-values $\lambda_1,\ldots, \lambda_k$, associated with the partial test statistics in $\boldsymbol{T}$, are positively dependent under the alternative, and this is irrespective of dependence relations among component variables in $\boldsymbol{X}$. They also imply that partial tests T_i, $i = 1,\ldots,k$, must be considered in such a way that their permutation distributions are monotonically related to underlying entities not implied by sub-hypotheses H_{0i} or H_{1i}, but possibly implied by H_{0j} or H_{1j}, for some $j \neq i$. In practice, when each partial test is related to a different component variable, as for instance is usual in many multidimensional testing on locations, this property is easily satisfied, provided that each partial test T_i is unbiased for the proper sub-hypothesis H_{0i} against H_{1i}, $i = 1,\ldots,k$.

Sometimes positive dependence or marginal unbiasedness are only approximately satisfied. One important example is when H_{01} and H_{02} are respectively related to locations and scale coefficients in a testing problem where symbolic treatment may influence both. Therefore, for the positive dependence and marginal unbiasedness properties to be satisfied, on one hand, T_1 must be unbiased for H_{01} against H_{11}, irrespective of whether H_{02} is true or not; on the other hand, T_2 must be unbiased for H_{02} against H_{12}, irrespective of whether H_{01} is true or not.

For the sake of simplicity and uniformity of analysis, but without loss of generality, we only refer to combining functions applied to *p*-values associated with partial tests. Because of assumption 1), partial tests are permutationally equivalent to their *p*-values: $T_i \approx Pr\{T_i \geq T_o | \mathbf{X}\} = \lambda_i$, $i = 1,...,k$. Of course, this is a direct consequence of the monotonic nonincreasing behavior with respect to t of significance level functions $L_i(t) = Pr\{T_i^* \geq t | \mathbf{X}\}$. Thus, the nonparametric combination in a single second-order test $T'' = \psi(\lambda_1,\ldots,\lambda_k)$ is achieved by a continuous, non-increasing, univariate and non-degenerate real function $\psi : (0,1) \rightarrow \Re^1$. Of course, ψ satisfies the measurability property as every other function does in the permutation context. In order to be suitable for test combination, all combining functions ψ must satisfy at least the following reasonable properties:

a) the function ψ must be non-increasing in each argument: $\psi(\ldots,\lambda_i,\ldots) \geq \psi(\ldots,\lambda_i',\ldots)$ if $\lambda_i < \lambda_i'$, $i \in \{1,\ldots,k\}$;

b) every combining function ψ must attain its supremum value $\overline{\psi}$, possibly not finite, when at least one argument attains the zero: $\psi(\ldots,\lambda_i,\ldots) \rightarrow \overline{\psi}$ if $\lambda_i \rightarrow 0$;

c) $\forall\, \alpha > 0$, the critical value of every ψ is assumed to be finite and strictly smaller than the supremum value: $T_\alpha'' < \overline{\psi}$.

These properties of combining functions are quite reasonable and intuitive, and are generally easy to justify. Property *a)* is related to the unbiasedness of combined tests; *b)* and *c)* are related to the consistency. Further, these properties define a class C of combining functions, which contains the well-known combining functions of Fisher, Lancaster, Liptak, Tippett. Class C also contains the Mahalanobis quadratic form for invariance testing against alternatives lying at the same quadratic distance from H_0.

Furthermore C contains a class of admissible combining functions of independent tests characterized by convex acceptance regions, when these are expressed in terms of *p*-values λ's. In particular, C includes all combining functions which take account in a nonparametric way of the underlying dependence structure among *p*-values λ_i, $i = 1,...,k$.

Thus, a problem arises naturally: how to choose, for any given testing problem, the best combining function in class C. This seems to be very difficult and we believe it is unsolvable in the case of finite sample sizes and without any further restriction. At the moment, only ”asymptotic optimal combinations” may sometimes be obtained. Moreover, if $D_i = \gamma_i(T_i)$, $i = 1,\dots,k$, where γ_i are continuous monotonically increasing transformations of partial tests, then $\forall\, \psi \in C$, $T_D'' = \psi(\lambda_{D1},\dots, \lambda_{Dk})$ is permutationally equivalent to $\psi(\lambda_1,\dots, \lambda_k) = T''$, because the p-values are invariant under continuous monotonic increasing transformations of test statistics. Note that, if partial tests are all exact permutation tests, then for every combining function $\psi \in C$, the combined test T_ψ'' is an exact permutation test.

Consider a two-phase algorithm for the nonparametric combination. This algorithm is used to obtain a Monte Carlo estimate of the permutation distribution of a combined test. The first phase concerns the estimate of k-variate distribution of $\boldsymbol{T}$, the second derives the estimate of permutation distribution of combined test T_ψ'' by using the same simulation results as in the first phase. Note that, when it is clear from the context which combining function ψ has been adopted, in place of T_ψ'' we simply use T''.

Phase I. An algorithm which simulates the first phase of a procedure estimating the k-variate distribution of $\boldsymbol{T}$ should include the following steps:

1) Calculate the vector of the observed values of tests $\boldsymbol{T}$: $\boldsymbol{T}_o = \boldsymbol{T}(\boldsymbol{X})$.
2) Consider a member g^*, randomly drawn from the proper group of transformations $\boldsymbol{G}$, and the values of vector statistics $\boldsymbol{T}^* = \boldsymbol{T}(\boldsymbol{X}^*)$, where $\boldsymbol{X}^* = g^*(\boldsymbol{X})$. In most situations, data permutation $\boldsymbol{X}^*$ may be obtained by first considering a random permutation $(u_1^*,\dots,u_n^*)$ of basic label integers $(1,\dots,n)$ and then by assignment of related individual data vectors to the proper group; thus, according to the data representation given in $\boldsymbol{X}^* = \{\boldsymbol{X}(u_i^*),\ i = 1,\dots,n;\ n_1,\dots,n_C\}$ (see figure 2 below).
3) Repeat step ***I.2)*** B times independently. The set of conditional Monte Carlo iterations results $\{T_r^*,\ r = 1,\dots,B\}$ is thus a random sampling from the permutation k-variate distribution of vector test statistics $\boldsymbol{T}$.
4) The k-variate E.D.F. $\hat{F}_B(z \mid X) = \left[0.5 + \sum_r I(T_r^* \leq z)\right]/(B+1), \forall z \in \Re^k$, gives an estimate of the corresponding k-dimensional permutation distribution $F(z|\boldsymbol{X})$ of $\boldsymbol{T}$. Moreover, $\hat{L}_i(z|X) =$ $=[0.5+\sum I(T_{ir}^* \geq z)]/(B+1)$, $i = 1,\dots,k$, gives an estimate $\forall z \in \Re^1$ of the marginal permutation significance level functions $L_i(z|\boldsymbol{X}) = Pr\{T_i^* \geq z|X\}$, thus $\hat{L}_i(T_o|X) = \hat{\lambda}_i$ gives an estimate of

the marginal p-value $\lambda_i = Pr\{T_i^* \geq T_o | X\}$, relative to test T_i. All these are unbiased and consistent estimates of corresponding true values.

Figure 2 below summarizes the observed data set and one multidimensional permutation in a two-sample problem. Figure 3 summarizes the CMC-Procedure. In multidimensional problems, the CMC-Procedure only considers permutations of individual data vectors, so that: $\boldsymbol{X}^* = \{\boldsymbol{X}(u_i^*),\ i = 1,\ldots,n;\ n_1,\ldots,n_C\}$, as is explicitly displayed in the second part of figure 2, and thus all dependence relations which are present in the component variables are preserved. From this point of view, CMC-Procedure is essentially a multivariate procedure.

Fig. 2: Representation of a multivariate data permutation.

$X_1(1)$	…	$X_1(n_1)$	$X_1(1+n_1)$	…	$X_1(n)$		T_{o1}
…	…	…	…	…	…	→	…
$X_q(1)$	…	$X_q(n_1)$	$X_q(1+n_1)$	…	$X_q(n)$		T_{ok}

$X_1(u^*_1)$	…	$X_1(u^*_{n1})$	$X_1(u^*_{1+n1})$	…	$X_1(u^*_n)$		T^*_1
…	…	…	…	…	…	→	…
$X_q(u^*_1)$	…	$X_q(u^*_{n1})$	$X_q(u^*_{1+n1})$	…	$X_q(u^*_n)$		T^*_k

Fig. 3: Representation of the CMC-Procedure.

$\boldsymbol{X}$	$\boldsymbol{X}^*_1$	…	$\boldsymbol{X}^*_r$	…	$\boldsymbol{X}^*_B$
T_{o1}	T^*_{11}	…	T^*_{r1}	…	T^*_{B1}
…	…	…	…	…	…
T_{ok}	T^*_{1k}	…	T^*_{rk}	…	T^*_{Bk}

With respect to standard E.D.F. estimators, 1/2 and 1 have been added respectively to the numerators and denominators of relationships in step ***I.4)***. This is done in order to obtain estimated values of c.d.f. $F(z|\boldsymbol{X})$ and of p-values in the open interval (0,1), so that transformations by inverse c.d.f. of continuous distributions, such as $-\log(\lambda)$ or $\Phi^{-1}(\lambda)$, etc. (where Φ is the standard normal c.d.f.) are continuous. However, as B is generally large, this minor alteration is substantially irrelevant, because it does not modify test behaviour or consequent inferences, neither for finite sample sizes nor asymptotically. In particular, this proposition is valid:

Proposition. As B tens to infinity, $\hat{F}_B(z|\boldsymbol{X})$ almost surely converges to permutation c.d.f. $F(z|\boldsymbol{X})$, $\forall z \in \Re^k$.

For the proof of this statement, see Pesarin (2001).

Phase II. The second phase of the algorithm for simulating a procedure for nonparametric combination should include the following steps:

1) The k observed p-values are estimated on data $\boldsymbol{X}$ by $\hat{\lambda}_i = \hat{L}_i(T_{oi}|\boldsymbol{X})$, where $T_{oi} = T_i(\boldsymbol{X})$, $i = 1,...,k$, represent the observed values of partial tests and $\hat{L}_i$ are the i-th marginal significance level functions estimated by the CMC-Procedure on data set $\boldsymbol{X}$.
2) The combined observed value of the second-order test is again evaluated through the same conditional simulation results of the first phase, and is given by: $T^*{}_o = \psi(\hat{\lambda}_1,...,\hat{\lambda}_k)$.
3) The r-th combined value of vector statistics are then calculated by $T''^*{}_r =.\psi(\lambda^*_{1r},...,\lambda^*_{kr})$, where $\lambda^*_{1r} = \hat{L}_i(T^*_{ir}|\boldsymbol{X})$, $i = 1,\ldots,k$, $r = 1,\ldots,B$.
4) Hence, the p-value of combined test T'' is estimated as: $\hat{\lambda}''_\psi = \sum_r I(T''^*_r \geq T''_{io})/B$.
5) If $\hat{\lambda}''_\psi < \alpha$, global null hypothesis H_0 is rejected at significance level α.

Figure 4 below displays the nonparametric combination.

Fig. 4: Nonparametric combination.

T_{ol}	T^*_{11}	...	T^*_{rl}	...	T^*_{Bl}
...	...	...	...	...	...
T_{ol}	T^*_{1k}	...	T^*_{rk}	...	T^*_{Bk}

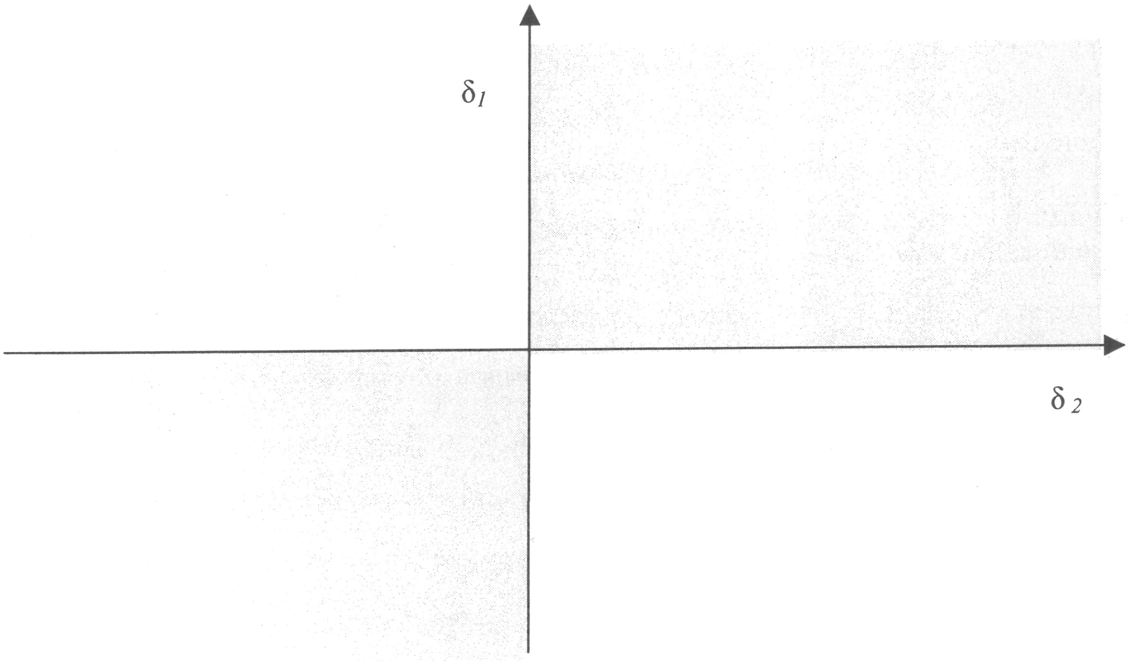
$\hat{\lambda}_1$	λ^*_{11}	...	λ^*_{1r}	...	λ^*_{1B}
...	...	...	...	...	...
$\hat{\lambda}_k$	λ^*_{k1}	...	λ^*_{kr}	...	λ^*_{kB}

T''_o	T''^*_1	...	T''^*_r	...	T''^*_B

The CMC-Procedure gives unbiased and consistent estimates of both true permutation distribution $F_\psi(t|\mathrm{X})$, $\Pr\{T''^* \leq t|\mathrm{X}\}$, $\forall t \in \Re^1$ and true p-value $\lambda''_\psi = \Pr\{T''^* \geq T''_o|\mathrm{X}\}$. In fact, $\hat{\lambda}_i \to \lambda_i$ with probability one (i = 1,…,k), as B tends to infinity. Hence, $\hat{\lambda}''_\psi$ converges to λ_ψ with probability one, k being a fixed finite integer and combining function ψ being continuous by assumption. This combination is a proper nonparametric method for multidimensional testing problems, because it takes into consideration only the whole joint k-variate permutation one inequality in each "sub-alternative" is strong. This situation is displayed in Fig.1, where the two gray sections ($\delta_1 > 0$, $\delta_2 > 0$) and ($\delta_1 < 0$, $\delta_2 < 0$) represent the alternative hypothesis, as the half-lines ($\delta_1 > 0$, $\delta_2 = 0$), ($\delta_1 < 0$, $\delta_2 = 0$), ($\delta_1 = 0$, $\delta_2 > 0$) and ($\delta_1 = 0$, $\delta_2 < 0$), while the null hypothesis is only the single point ($\delta_1 = 0$, $\delta_2 = 0$). The points in ($\delta_1 < 0$, $\delta_2 > 0$) and ($\delta_1 > 0$, $\delta_2 < 0$) are not relevant for the analysis. This kind of hypotheses arise when two variables are such that under the alternative at least one of them stochastically increases *XOR* decreases, whereas the other variable may remain either affected or not affected.

In our genetic context, this happens when a gene is associated with a given disease so that, on affected individuals (cases), at least one of genotype frequencies with putative allele increases *XOR* decreases with respect to non-affected individuals (controls).

Fig. 1: bivariate isotonic hypotheses.



Of course, as under the null hypothesis, the pooled data set ***X*** is a set of sufficient statistics for the problem, so that, the partial tests to take into consideration are:

$$T_h^* = \sum_i X_{h2i}^* - \sum_i X_{h1i}^*, \, h = 1,2.$$

In the present problem, under H_1, p-values of partial tests are either stochastically smaller than α or stochastically larger than 1 – α. So that, we need to modify assumptions 1) and 2) in the Sec. 2.3, into:

1) all partial tests T_i, $i = 1,2$, are marginally unbiased and significant either for large or small values, so that their permutation distribution under H_1 are either stochastically larger or smaller than under H_0.
2) all partial tests T_i, $i = 1,2$, are consistent.

Furthermore, we also need to modify the properties of combining functions ψ ("a", "b" and "c" of Sec. 2.3), into:

a) a continuous combining function ψ must be monotonically decreasing in each argument: $\psi(\ldots, \lambda_i,\ldots) > \psi(\ldots, \lambda'_i,\ldots)$, if $\lambda_i < \lambda'_i$, $i = 1,\ldots,k$;
b) it must attain its supremum positive value $\overline{\psi}$, possibly not finite, when at least one argument attains 0 (zero): $\psi(...,\lambda_i,...) \rightarrow \overline{\psi}\ if\ \lambda_i \rightarrow 0$; moreover it must attain its infimum negative value $\underline{\psi}$, possibly not finite, when at least one argument attains 1: $\psi(...,\lambda_i,...) \rightarrow \underline{\psi}\ if\ \lambda_i \rightarrow 1$;
c) $\forall\alpha$ >0, its acceptance region is bounded: $\underline{\psi} < T''_{\alpha/2} < T'' < T''_{1-\alpha/2} < \overline{\psi}$.

Further, we need to modify also step II.5 in Sec. 2.3, into:

(II.5) if $1-|2\hat{\lambda}''_{\psi}-1| \leq \alpha$, then reject H_0 at significance level α.

If the exchangeability property is satisfied under H_0, the nonparametric combination methods lead to exact, unbiased and consistent permutation tests (Pesarin).

An allele A at a gene of interest is said to be associated with the disease if it occurs at a significantly higher or smaller frequency among affected compared with control individuals. For a bi-allelic locus with common allele a and rare allele A, individuals may carry zero (subjects with genotype aa), one (subjects with genotype Aa) or two (subjects with genotype AA) copies of the A allele. Therefore, conventionally testing for allelic association implies to test for the joint equality in distribution of genotype frequencies against an alternative of XOR dominance of cases with respect to controls by using a proper test statistic. In doing this, it should be taken into consideration that, by referring to genotype-specific risks $R_h = f_{h1}/f_{h2}$, $h = AA$, Aa, aa, (where f_{hj}, $j = 1,2$, are the observed frequencies in cases and controls, respectively) the effect of an allele can be expressed according to only one of the following ways:

1.Recessive: there is an effect only in the presence of two copies of A allele (genotype AA), whereas the behaviour in heterozygous condition (genotype Aa) is the same as the reference and commonest condition (genotype aa), so that: ($R_{AA} > R_{Aa} = R_{aa}$, in presence of a protective effect) *XOR* ($R_{AA} < R_{Aa} = R_{aa}$ for a deleterious effect).

2.Codominant: there is an ordering on effects associated with the A allele: genotype Aa is of risk (or of protection) in comparison with the genotype aa, and AA is of risk (or of protection) in comparison with the genotype Aa. Obviously, AA is of great risk (or of great protection) in comparison with the genotype aa, so that ($R_{AA} > R_{Aa} > R_{aa}$, for a protective effect) *XOR* ($R_{AA} < R_{Aa} < R_{aa}$ for a deleterious effect).

3.Dominant: the effect of the A allele is the same in AA and Aa genotype. In this situation, there is no relative risk (or protection) between AA and Aa, but only between AA (or Aa) and aa, so that: ($R_{AA} = R_{Aa} > R_{aa}$, protection) *XOR* ($R_{AA} = R_{Aa} < R_{aa}$, risk).

For these reasons, differences in risk should be tested for over the restricted parameter space, which properly fits the plausible biological models, defined as: ($R_{AA} \geq R_{Aa} \geq R_{aa}$) *XOR* ($R_{AA} \leq R_{Aa} \leq R_{aa}$).

Following Chiano and Clayton (1998), in order to reduce the analysis from three to two dimensions, because in a 2×3 contingency table there are only 2 degrees of freedom, we may consider odds ratios of genotype-specific relative risks, which contain all relevant information and are defined as $\theta_{AA} = R_{AA}/R_{Aa}$ and $\theta_{Aa} = R_{Aa}/R_{aa}$, respectively. Thus, the hypotheses under testing may be equivalently expressed as: H_0: $\{\theta_{AA} = \theta_{Aa} = 1\}$, against H_1: $\{[(\theta_{AA} \geq 1) \cap (\theta_{Aa} \geq 1)]$ *XOR* $[(\theta_{AA} \leq 1) \cap (\theta_{Aa} \leq 1)]\}$, where at least one inequality in both directions is strong. This system of hypotheses is equivalent to the previous.

In order to solve the problem within the permutation approach, it should be noted that relation defining the null hypothesis

$$H_0: \{(\theta_{AA} = 1) \cap (\theta_{Aa} = 1)\}$$

is equivalent to

$$H_0: \{(f_{AA,cases} \cdot f_{Aa,controls} \overset{d}{=} f_{Aa,cases} \cdot f_{AA,controls}) \cap (f_{Aa,cases} \cdot f_{aa,controls} \overset{d}{=} f_{aa,cases} \cdot f_{Aa,controls})\},$$

which is easier for computations because expressed in terms of products of frequencies.
The permutation solution is based on two partial statistics:

$$T_{AA} = f_{AA,cases} \cdot f_{Aa,controls} \,/\, (f_{Aa,cases} \cdot f_{AA,controls})$$
$$T_{Aa} = f_{Aa,cases} \cdot f_{aa,controls} \,/\, (f_{aa,cases} \cdot f_{Aa,controls})$$

which test the respective partial hypotheses:

$$H_{0AA}\text{: } \{\theta_{AA} = 1\} \text{ against } H_{1AA}\text{: } \{\theta_{AA} > 1 \text{ or } \theta_{AA} < 1\}$$
$$H_{0Aa}\text{: } \{\theta_{Aa} = 1\} \text{ against } H_{1Aa}\text{: } \{\theta_{Aa} > 1 \text{ or } \theta_{Aa} < 1\}.$$

Note, in fact, that:

$$\{\theta_{AA} = 1\} \Leftrightarrow \{f_{AA,cases} \cdot f_{Aa,controls} \overset{d}{=} f_{Aa,cases} \cdot f_{AA,controls}\},$$

so that, the two relations are equivalent.
To explain how the test is done, we start from the CMC method. We construct a vector of dimension n ($n = n_{cases} + n_{controls}$), and we assign three different values to observations of different genotypes, for instance: 1 to all the n_{AA} subjects who stay in the cells (AA, cases) and (AA, controls), 2 to all the n_{Aa} subjects who stay in the cells (Aa, cases) and (Aa, controls), and 3 to all remaining n_{aa} subjects who stay in the cells (aa, cases) and (aa, controls).

Now, we insert randomly the $f_{AA,cases}$ values 1, the $f_{Aa,cases}$ values 2 and the $f_{aa,cases}$ values 3 in the first n_{cases} positions of the vector, and, in the same way, all the others values in second $n_{controls}$ positions of the vector. We obtain a vector as that one in Fig. 2.

Fig. 2: Vector of the data for the permutation test of the allelic association problem.

Values:	2	1	…	…	2	3	…	…	1
Position:	1	2	…	…	n_{cases}	$n_{cases} + 1$	…	…	n

In this way, we preserve all the marginal values of an association table (n_{cases}, $n_{controls}$, n_{AA}, n_{Aa}, n_{aa}).
The permutation statistics T^*_{AA} and T^*_{Aa} are calculated on the same vector, after executing a random permutation of its n elements. For example, the estimation of partial p-value λ_{AA} is obtained using B CMC-Iterations, as:

$$\hat{\lambda}_{AA} = \frac{\#(T^{*}_{AA} \geq T^{oss}_{AA})}{B},$$

this partial p-value is distributed as $U(0,1)$ and it makes to reject of H_{0AA} if it will result $\hat{\lambda}_{AA} \leq \alpha/2$, or $\hat{\lambda}_{AA} \geq 1-\alpha/2$, at a fixed significance level α. By using the same B vectors, previously obtained, we estimate the p-values $\lambda^{'}_{AAs} = \Pr(T^{*}_{AA} \geq T^{*}_{AAs} | \mathbf{X})$ too, where $s \in (1,\ldots,B)$:

$$\hat{\lambda}^{'}_{AAs} = \frac{\#(T^{*}_{AA} \geq T^{*}_{AAs})}{B}.$$

Now, with the two partial p-values and the other B p-values of first type for each of them, we use the combining function of Liptak to construct the combined test which verifies the initial hypothesis system. The final p-value λ_L is estimated by:

$$\hat{\lambda}_{L} = \frac{\#^{B}_{s}\{[\Phi^{-1}(1-\hat{\lambda}^{'}_{AAs})+\Phi^{-1}(1-\hat{\lambda}^{'}_{Aas})] \geq [\Phi^{-1}(1-\hat{\lambda}_{AA})+\Phi^{-1}(1-\hat{\lambda}_{Aa})]\}}{B}.$$

Also the final p-value follows a distribution U(0,1). Further, if $\hat{\lambda}_{L} \leq \alpha/2$, we consider the rare allele of risk, whereas, if $\hat{\lambda}_{L} \geq 1-\alpha/2$, we reputed it of protection.

3.2 Exact Nonparametric Solution for the Genetic Problem

We can represent the previous problem by a simple case-control contingency table (see Fig. 2).

Fig. 2: Case-control table for allelic association study.

Genotype/haplotype:	*Cases*	*Controls*	Size
AA	X_1	Y_1	$S_1 = X_1 + Y_1$
Aa	X_2	Y_2	$S_2 = X_2 + Y_2$
Aa	X_3	Y_3	$S_3 = X_3 + Y_3$
Size	$M = X_1 + X_2 + X_3$	$N = Y_1 + Y_2 + Y_3$	$\mathrm{S} = M + N = S_1 + S_2 + S_3$

It should be noted that in all this types of studies, the data may be represented in a fixed (in this case 3×2) contingency table with fixed marginal values. The total of the cases, M, and the total of the

controls, N, are fixed numbers, obtained from experimental observations. At the same time, also the number of genotypes AA, in cases and controls together, S_1, is fixed, and so on for S_2 and S_3.
With the usual representation by data file, we have the following structure:

Fig. 2: Data representation by a file.

Observation	1	2	3	4	…	M	$M+1$	…	$S = M + N$
Genotype	Aa	AA	AA	aa	…	Aa	aa	…	AA
Permutation order	u_1	u_2	u_3	u_4	…	u_M	u_{M+1}	…	u_S

where, in the first M observations (or subjects), we have X_1 genotypes AA, X_2 genotypes Aa and X_3 genotypes aa. It does not matter what order between the first M subjects we have (and the same for the second N subjects), because the result in the contingency table does not change if we take two random permutations into these sub-vectors, and the frequencies X_1, X_2, X_3, Y_1, Y_2 and Y_3 remain the same. So that, if we consider the overall permutation space associated to the data in the previous paragraph, (S!), it may be very large to explore exhaustively, also for the more modern computer (and if it would be possible in some situations, its time of execution will be very high).

Now, instead, we think to look exclusively at those specific combinations and recombinations of the permutated genotypes/haplotypes in the table, which give us a particular structure of the cells. Observe the following example to explain this concept. We have a particular permutation in the data which allows to obtain the dataset represented in Tab. 1.

Tab. 1: A particular result of a permutation in the dataset.

	$Ca.$	$Co.$	
AA	x_1^*	y_1^*	S_1
Aa	x_2^*	y_2^*	S_2
aa	x_3^*	y_3^*	S_3
	M	N	S

The marginal sums are identical for any permutation, only the frequencies in the cells may change. The relative data file is illustrated in the Fig. 4.

Fig. 4: Representation by the data file of the permutation.

Observation	1	2	3	4	…	M	$M+1$	…	$S=M+N$
Genotype	aa	Aa	Aa	AA	…	aa	Aa	…	Aa
Permutation order	u_1^*	u_2^*	u_3^*	u_4^*	…	u_M^*	u_{M+1}^*	…	u_S^*

Here, $\forall i,i^{'}(i \neq i^{'}), u_i^* = u_j^*$ $and\, u_{i^{'}}^* = u_{j^{'}}^*, where\;\; j \neq j^{'}, and\, i,i^{'},j,i^{'} \in \{1,...,S\}$, furthermore, in the first M observations (or subjects), we have x_1^* genotypes AA, x_2^* genotypes Aa and x_3^* genotypes aa. Again, the orders of the two sub-vectors (firs M elements and second N elements) are not important.

We see that there are not (S)! different results for the permutations, but many permutations with different numbers give a specific structure of the cells $x_1^*, x_2^*, x_3^*, y_1^*, y_2^*$ and y_3^*, which are the important parameters for our statistics.

So, we can construct the exact permutation distribution for the statistics, associating to the statistics their related frequencies, that is, the times these values of the statistics appear into the (S!) permutations. We do not need expensive iterations, by computer, in doing that, but we can use the combinatorial calculus. Then, we are looking for the frequencies associated, in the exploration of the total sample space, to all the different configurations of the table (Tab. 1), that is all the sets $\{ x_1^*, x_2^*, x_3^*, y_1^*, y_2^*, y_3^* \}$ where at least one cell is different from the others.

For the data in Tab. 1, we can obtain all the different table configurations by the following algorithm :

1) $x_1^* \in [\max(0, S_1 - N), \min(M, S_1)]$;

2) $y_1^* = S_1 - x_1^*$;

3) $x_2^* \in [\max(S_2 - (N - y_1^*)), \min(M - x_1^*, S_2)]$;

4) $y_2^* = S_2 - x_2^*$;

5) $x_3^* = M - x_1^* - x_2^*$;

6) $y_3^* = S_3 - x_3^*$.

Then, for a specific set *i* $\{ {}_ix_1^*, {}_ix_2^*, {}_ix_3^*, {}_iy_1^*, {}_iy_2^*, {}_iy_3^* \}$ we have the frequency:

$$f_i^* = M! \cdot N! \cdot \begin{pmatrix} S_1 \\ {}_ix_1^* \end{pmatrix} \cdot \begin{pmatrix} S_2 \\ {}_ix_2^* \end{pmatrix} \cdot \begin{pmatrix} S_3 \\ {}_ix_3^* \end{pmatrix} = (M!N!S_1!S_2!S_3!)/({}_ix_1^*!\,{}_ix_2^*!\,{}_ix_3^*!\,{}_iy_1^*!\,{}_iy_2^*!\,{}_iy_3^*!);$$

and, of course, the sum of all the frequencies is:

$$\sum_{i}^{I} f_i^* = (M+N)! = (S!);$$

where the total number of all these different configurations is:

$$I = \sum_{x_1^*}^{\min(M,S_1)+1-\max(0,S_1-N)} [\min(M-x_1^*, S_2)+1-\max(0, S_2-(N-(S_1-x_1^*)))];$$

so that, the relative frequencies (more simple in the computations by computer) are $p_i^* = f_i^* /(S!)$. Of course, the highest relative frequency is associated to configuration where x_1^* and x_2^* are close maximally (if possible, equal) to, respectively, y_1^* and y_2^*; that coincides (in general) with the case of no association between cases and controls.

Instead, we can see that the sampling distribution is in the form of bell shape (where the parameters are: the mean of the cell configurations, that is, in general, equal to the configuration which has the maximum relative frequency; the variance between the cell configurations). But, note that this distribution is not continue, because the data are discrete.

So, we can repeat the same previous test with the nonparametric combination to have an exact *p*-value associated to the hypothesis system. We can call this type of procedure CEP (Conditional Exact Procedure) to distinguish it from the CMC-Procedure (Conditional Monte Carlo Procedure) shown previously.

3.3 The Parametric Approach of Chiano and Clayton

Chiano and Clayton (1998) started from this specific system of hypotheses with the odds ratio, that considers the admissible genetic model in the definition of the model, but after, for convenience, the used the log transformations of the odds: β_{AA} = log θ_{AA} and β_{Aa} = log θ_{Aa}; so that the parametric space under the null hypothesis is:

$$\Omega_1 : \{(\beta_{AA} \geq 0) \cap (\beta_{Aa} \geq 0)\} XOR \{(\beta_{AA} \leq 0) \cap (\beta_{Aa} \leq 0)\},$$

while the null space is Ω_0, the origin of the axes $\{(\beta_{AA} = 0) \cap (\beta_{Aa} = 0)\}$, that is equivalent to that one shown in Fig. 1 (Sec. 3.1).

We indicate by β the vector $[\beta_A, \beta_A]'$, by using the standard theory of estimation, they obtain $\lim_{n\to\infty}\sqrt{n}(\hat{\beta}_1 - \hat{\beta}_0) \to N(0, \sum_{\beta_0})$, where $\sum_{\beta_0}$ is the variance-covariance matrix of β evaluated into the null hypothesis. Then, if all regularity conditions hold, inference would be made by referring the likelihood ratio chi-squared statistic of Wilks:

$$\Lambda = 2\sum nP(\hat{\beta})\log\left[\frac{P(\hat{\beta})}{P(\beta_0)}\right],$$

to a standard χ^2 on 2 df.

Unfortunately, under such order restriction, Wilk's regularity assumptions are not met and the null point (origin) is on the boundary. The likelihood is therefore maximized subject to order constraints as follows. First, we obtain the unrestricted maximum likelihood estimate $\hat{\beta}$ of β:

I) if $\hat{\beta} \in \{(\beta_{AA} \leq 0) \cap (\beta_{Aa} \geq 0)\} \cup \{(\beta_{AA} \geq 0) \cap (\beta_{Aa} \leq 0)\}$ (the unshaded region in Fig. 1 of Sec. 3.1), $\hat{\beta}$ is remaximized subject to the constraint that $\beta_{Aa} = 0$ or $\beta_{AA} = 0$ (whichever is maximum) and

II) if $\hat{\beta}$ is in the shared region (Fig. 1 of Sec. 3.1), it is left as it is.

In other words, when $\hat{\beta}$ falls in the unshaded region, it is projected onto the line $\beta_{Aa} = 0$ or $\beta_{AA} = 0$ and the contribution to the overall distribution is χ_1^2. However, when $\hat{\beta}$ falls in the shaded region, the contribution is χ_2^2, with probability $\lambda(\beta)$ proportional to the area of the shaded region (for details, see Chiano and Clayton, 1995). Then, it turns out that the distribution of the likelihood ratio chi-square statistic can be represented as a mixture of 2 chi-square distributions:

$$\Lambda = \lambda(\beta)\chi_2^2 + (1 - \lambda(\beta))\chi_1^2,$$

where $\lambda(\beta)$, the mixing probability, can be approximated to

$$\lambda(\beta) = \cos^{-1}\left(\frac{I_{12}}{\sqrt{I_{11}I_{22}}}\right)/\pi,$$

I_{ij} being the components (i,j) of the variance-covariance matrix $\sum_{\beta_0}$ evaluated at the null.

We denote by α the usual required significance level, therefore, we have to find x such that:

$$\mathrm{P}(\Lambda \geq x_\alpha) = \lambda(\beta)\mathrm{P}(\chi_2^2 \geq x) + (1 - \lambda(\beta))\mathrm{P}(\chi_1^2 \geq x).$$

3.4 The Maximum Likelihood Solution

This solution was developed by Hammou El Barmi referring to the previous work of Dykstra, Kochar and Robertson (1995), within a collaboration project where I am involved.

The case-control table can be interpreted as two independent vectors of data. The random sample of $\boldsymbol{M}$ cases is taken from a multinomial distribution with probability vector $\boldsymbol{p} = (p_1, p_2, p_3)$ where it refers to the observed values ($p_1 = X_1/\boldsymbol{M}$, $p_2 = X_2/\boldsymbol{M}$, $p_3 = X_3/\boldsymbol{M}$), while the $\boldsymbol{N}$ controls are taken from a multinomial (independent from the other) distribution $\boldsymbol{q} = (q_1, q_2, q_3)$, where $q_1 = Y_1/N$, $q_2 = Y_2/N$, $q_3 = Y_3/N$. We can derive the nonparametric MLE's of the probability vectors $\boldsymbol{p}$ and $\boldsymbol{q}$ under the two hypothesis systems:

$$H_0\text{: } \{\boldsymbol{p} = \boldsymbol{q}\} \text{ VS } H_1\text{: } \{\boldsymbol{p} \overset{LR}{>} \boldsymbol{q}\},$$

and

$$H_0\text{: } \{\boldsymbol{p} = \boldsymbol{q}\} \text{ VS } H_1\text{: } \{\boldsymbol{p} \overset{LR}{<} \boldsymbol{q}\},$$

And then use these estimates to construct a likelihood ratio test.

The symbols $\overset{LR}{>}$ or $\overset{LR}{<}$ mean that it exists a likelihood ratio ordering between the distributions of two vectors: $(X \overset{LR}{>} Y) \Rightarrow \forall a, b\ (a < b)$ the conditional distribution of X given $X \in (a, b)$ is stochastically greater than that of Y given $Y \in (a, b)$, or, equivalently, $[f_x(t)/f_y(t)]$ is nondecreasing in t, with f_x and f_y the density functions of X and Y.

Now, we need to express the likelihood function of (p, q), vectors of parameters (p_1, p_2, p_3) and (q_1, q_2, q_3), as $L \propto \prod_{i=1}^{3} p_i^{X_i} q_i^{Y_i}$. We reparameterize by letting: $\theta_i = M{\cdot}p_i/(M{\cdot}p_i + N{\cdot}q_i)$ and $\Phi_i = M{\cdot}p_i + N{\cdot}q_i$, to obtain $p_i = \theta_i\Phi_i/M$ and $q_i = \Phi_i(1 - \theta_i)/N$, $i = 1,2,3$. With some passages (see Dykstra, *et al.* 1995), we obtain the MLE's of p and q under H_0 and H_1. Under H_0, that is $p = q$, we have $p_i = q_i = (X_i + Y_i)/(M + N)$. Instead, under H_1, the MLE's are:

$$\text{if } p > q,\ p_i^* = [(X_i + Y_i)/M]{\cdot}E_{(X+Y)}[\boldsymbol{X}/(\boldsymbol{X} + \boldsymbol{Y})|\{(\theta_0,\theta_1,\theta_2)\text{: } \theta_0 \le \theta_1 \le \theta_2\}]_i$$

$$\text{and } q_i^* = [(X_i + Y_i)/N]{\cdot}E_{(X+Y)}[\boldsymbol{Y}/(\boldsymbol{X} + \boldsymbol{Y})|\{(\theta_0,\theta_1,\theta_2)\text{: } \theta_0 \ge \theta_1 \ge \theta_2\}]_i,$$

$$\text{if } p < q,\ p_i^* = [(X_i + Y_i)/M]{\cdot}E_{(X+Y)}[\boldsymbol{X}/(\boldsymbol{X} + \boldsymbol{Y})|\{(\theta_0,\theta_1,\theta_2)\text{: } \theta_0 \ge \theta_1 \ge \theta_2\}]_i$$

$$\text{and } q_i^* = [(X_i + Y_i)/N]{\cdot}E_{(X+Y)}[\boldsymbol{Y}/(\boldsymbol{X} + \boldsymbol{Y})|\{(\theta_0,\theta_1,\theta_2)\text{: } \theta_0 \le \theta_1 \le \theta_2\}]_i,$$

where ***X*** and ***Y*** are the data vectors (x_1,x_2,x_3) and (y_1,y_2,y_3), while the θ_i's must satisfy some conic restrictions (see see Dykstra, *et al.* 1995).

These MLE's are consistent in the sense that the associated cdf's converge pointwise to the true cdf's when $M,N \to \infty$ and the likelihood ratio order holds.

The likelihood ratio test is the statistic:

$$\Psi = \frac{\sup_{(p,q)\in H_0} L((p.q))}{\sup_{(p,q)\in H_1} L((p.q))} = \frac{L(p^0.q^0)}{L(p^*.q^*)} = \ldots = \prod_{i=1}^{3} \left(\frac{\theta_i^0}{\theta_i^*}\right)^{X_i} \left(\frac{1-\theta_i^0}{1-\theta_i^*}\right)^{Y_i},$$

because $\Phi_i^0 = \Phi_i^*$. The test rejects H_0 for large values of $T = -2\ln\Psi$, that is for large values of

$$T = 2\sum_{i=1}^{3} \{X_i \ln\theta_i^* + Y_i \ln(1-\theta_i^*) - X_i \ln\theta_i^0 - Y_i \ln(1-\theta_i^0)\}.$$

T, under H_0, has, asymptotically, a mixed distribution of χ_1 and χ_2, but it is not simple writing it.

Now, we consider a special algorithm to find the maximum likelihood estimates. We denote by $(m_1, m_2,\ldots, m_k)$ and $(n_1, n_2,\ldots, n_k)$ the frequencies corresponding to two independent multinomial with parameters $(m,(p_1, p_2,\ldots, p_k))$ and $(n,(q_1, q_2,\ldots, q_k))$ respectively. Next we consider maximizing

$$\prod_{i=1}^{3} p_i^{m_i} \prod_{i=1}^{3} q_i^{n_i} \qquad (1)$$

subject to

$$\frac{p_1 q_2}{p_2 q_1} \geq \frac{p_2 q_3}{p_3 q_2} \geq 1 \qquad (2)$$

or

$$\frac{p_1 q_2}{p_2 q_1} \leq \frac{p_2 q_3}{p_3 q_2} \leq 1 \qquad (3)$$

and

$$\sum_{i=1}^{3} p_i = 1, \quad \sum_{i=1}^{3} q_i = 1. \qquad (4)$$

To solve this problem, we use the algorithm developed in El Barmi and Dykstra (1998). For completeness, we describe next the algorithm and then show how to apply it to (1).

Consider the problem of maximizing

$$\prod_{i=1}^{k} p_i^{\,i} \qquad (5)$$

subject $\mathbf{p} \in P$ and

$$\mathbf{p} \in K_1 \tag{6}$$

$$\ln \mathbf{p} \in K_2 \tag{7}$$

where ln **p** = (ln p_1, ln p_2,..., ln p_k), K_1 and K_2 are two cones (a cone is defined to be a subset of $\Re^k$ that satisfies $\alpha\mathbf{x}$ in the cone whenever **x** is in the cone for any $\alpha \geq 0$). Examples of cones in $\Re^k$ include any linear space and a nonnegative orthant. We assume here that K_2 contains constant vectors and note that when $K_1 = \Re^k$ and K_2 is a linear space, this optimization problem corresponds to fitting a log-linear model to the data (n_1, n_2,..., n_k).

For a given cone K, its dual (polar) cone is defined to be

$$K^* = \{\mathbf{y}, \sum_{i=1}^{K} x_i y_i \leq 0, \forall \mathbf{x} \in K\}. \tag{8}$$

It is easy to see that when K is actually a linear space (and hence a cone), then K^* is its orthogonal space.

El Barmi and Dykstra (1998) show that if $\mathbf{y}^* \in K_1^*$ and $\mathbf{z}^* \in K_2^*$ solve

$$\min \sum_{i=1}^{k} (\hat{p}_i - z_i) \ln\left[\frac{\hat{p}_i - z_i}{1 + y_i}\right] \tag{9}$$

subject to

$$\mathbf{y}^* \in K_1^* \tag{10}$$

$$\mathbf{z}^* \in K_2^* \tag{11}$$

then the vector whose i-th component is given by

$$p_i^* = \frac{\hat{p}_i - z_i^*}{1 + y_i^*},\ i = 1, 2,\ldots, k, \tag{12}$$

solves (1) subject to (2) and (3).

To find $\mathbf{y}^*$ and $\mathbf{z}^*$ they developed an iterative algorithm which is guaranteed to converge to the true solution when the constraint region is not empty. To apply the algorithm one proceeds as follows. Set $\mathbf{z}^{(0)} = \mathbf{0}$ and $\mathbf{y}^{(0)} = \mathbf{0}$ and $\nu = 1$. At the ν-th step of the algorithm, step 1, calculate $\mathbf{z}^{(\nu)}$ which solves

$$\min_{z \in K_2^*} \sum_{i=1}^{k} (\hat{p}_i - z_i) \ln\left[\frac{\hat{p}_i - z_i}{1 + y_i^{(\nu-1)}}\right]. \tag{13}$$

The second step of the algorithm amounts to finding $\mathbf{y}^{(\nu)}$ that solves

$$\max_{y\in K_1^*}\sum_{i=1}^{k}(\hat{p}_i - z_i^{(v)})\ln(1+y_i)\,. \tag{14}$$

At the end of the v-th cycle the estimate of **p** is given by the vector whose *i*-th component is

$$p_i^{(v)} = \frac{\hat{p}_i - z_i^{(v)}}{1+y_i^{(v)}}\,,\ i = 1, 2,\ldots, k. \tag{15}$$

This two step procedure is repeated until sufficient accuracy is attained.

To apply the algorithm to (1) we reparametrize the problem as follows. Consider

$$\theta_i = \begin{cases} p_i/2, & i = 1,2,3 \\ q_{i-3}/2, & i = 4,5,6 \end{cases}$$

and

$$r_i = \begin{cases} m_i, & i = 1,2,3 \\ m_{i-3}, & i = 4,5,6 \end{cases}$$

and note that maximizing (1) is equivalent to maximizing

$$\prod_{i=1}^{6}\theta_i^{r_i} \tag{16}$$

subject to

$$\frac{\theta_1\theta_5}{\theta_2\theta_4} \geq \frac{\theta_2\theta_6}{\theta_3\theta_5} \geq 1 \tag{17}$$

$$\frac{\theta_1\theta_5}{\theta_2\theta_4} \leq \frac{\theta_2\theta_6}{\theta_3\theta_5} \leq 1 \tag{18}$$

and

$$\sum_{i=1}^{3}\theta_i - \sum_{i=4}^{6}\theta_i = 0\,, \tag{19}$$

$$\sum_{i=1}^{6}\theta_i = 1\,, \tag{20}$$

in the sense that if θ_i^*, $i = 1, 2,\ldots, 6$, solve (16), then $(p_i^*, q_i^*) = (2\theta_i^*, 2\theta_{i+3}^*)$, $i = 1,2,3$, solve (1).

Let $\mathbf{z}_1^{(1)}$ = (-1,2,-1,1,-2,1), $\mathbf{z}_2^{(1)}$ = (0,-1,1,0,1,-1) and $\mathbf{z}_1^{(2)}$ = (1,-2,1,-1,2,-1), $\mathbf{z}_2^{(2)}$ = (0,1,-1,0,-1,1), then (17) and (18) can be expressed as

$$\sum_{j=1}^{6} z_{ij}^{(1)}\theta_j \leq 0\,,\ i = 1,2,$$

or

$$\sum_{j=1}^{6} z_{ij}^{(2)}\theta_j \leq 0\,,\; i = 1{,}2,$$

or equivalently, (θ_1, θ_2,…, θ_6)∈ $K_2^{(1)}$ or $K_2^{(2)}$ where $K_2^{(1)}$ and $K_2^{(2)}$ are two cones whose respective duals are given by

$$K_2^{(l^*)} = \{(\beta_1 z_{11}^{(l)} + \beta_2 z_{21}^{(l)},\ldots, \beta_1 z_{16}^{(l)} + \beta_2 z_{26}^{(l)}),\; \beta_1 \geq 0, \beta_2 \geq 0\},$$

l = 1,2. Let K_1 = {θ, $\Sigma_{i=1}^{3}\theta_i$ - $\Sigma_{i=4}^{6}\theta_i = 0$} and note that K_1^*={(αy_1, αy_2,..., αy_6), α∈ ℜ } where y_i = 1, i = 1,2,3 and y_i = -1, i = 4,5,6.

Our problem is then equivalent to

$$\max\{\max(\theta\in K_1 \cap K_2^{(1)})\prod_{i=1}^{6}\theta_i^{r_i}\,,\; \max(\theta\in K_1 \cap K_2^{(2)})\prod_{i=1}^{6}\theta_i^{r_i}\;\} \qquad (**)$$

and θ a probability vector. The algorithm described above can now be applied to solve

$$\max(\theta\in K_1 \cap K_2^{(1)})\prod_{i=1}^{6}\theta_i^{r_i} \qquad (21)$$

$$\max(\theta\in K_1 \cap K_2^{(2)})\prod_{i=1}^{6}\theta_i^{r_i} \qquad (22)$$

and θ a probability vector, individually and hence find the overall maximum.

For simplicity, we only consider how to implement the algorithm to (**). The dual problem in this case is given by

$$\min\sum_{i=1}^{6}(\hat{\theta}_i - \beta_1 z_{1i}^{(1)} - \beta_2 z_{2i}^{(1)})\ln\left[\frac{\hat{p}_i - \beta_1^{(v)} z_{1i}^{(1)} - \beta_2^{(v)} z_{2i}^{(1)}}{1+\alpha y_i}\right] \qquad (23)$$

subject to $\beta_1 \geq 0$, $\beta_2 \geq 0$ and α∈ ℜ, where $\hat{\theta}_i = r_{ij}/(m+n)$, i = 1, 2,…, 6.

To apply the algorithm, set β_i = 0, i = 1,2, and α = 0. At the step of the algorithm, step 1, we calculate $\beta_1^{(v)}$, i = 1,2, which solve

$$\min\sum_{i=1}^{6}(\hat{\theta}_i - \beta_1 z_{1i}^{(1)} - \beta_2 z_{2i}^{(1)})\ln\left[\frac{\hat{\theta}_i - \beta_1^{(v)} z_{1i}^{(1)} - \beta_2^{(v)} z_{2i}^{(1)}}{1+\alpha^{(v-1)} y_i}\right] \qquad (24)$$

subject to

$$\beta_i \geq 0,\; i = 1{,}2.$$

The second step of the algorithm amount to finding α$^{(v)}$ that solves

$$\max\sum_{i=1}^{6}(\hat{\theta}_i - \beta_1 z_{1i}^{(1)} - \beta_2 z_{2i}^{(1)})\ln\{1+\alpha y_i\} \qquad (25)$$

subject to $\alpha \in \Re$.

The estimate at the ν-th cycle of θ_{ij} is given by

$$\theta_i^{(\nu)} = \frac{\hat{\theta}_i - \beta_1^{(\nu)} z_{1i}^{(1)} - \beta_2^{(\nu)} z_{2i}^{(1)}}{1 + \alpha^{(\nu)} y_i}.$$

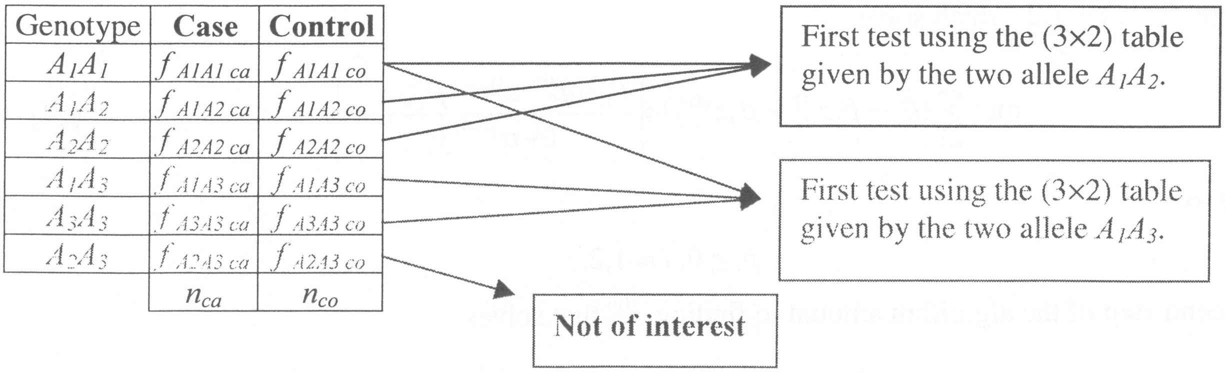
This iterative procedure is continued until sufficient accuracy is attained. What is remarkable here is that, at each step of the algorithm, a Newton Raphson technique can be used on each variable and it requires a very small number of iterations to converge.

Let $\theta^{(1)'}$ and $\theta^{(2)'}$ be the solutions corresponding (21) and (22) respectively. Then the overall maximum must be achieved at one of them and hence is the solution to (18).

3.5 Some Extensions of the Nonparametric Solution to Multivariate Problems

We consider the immediate extensions to the nonparametric solution illustrated in Sec. 3.1. Of course, we may have multiallelic loci as (A_1, A_2, A_3), where loci A_2 and A_3 can be both rare. In this case we can construct the previous nonparametric tests (by CMC or CEP) separately for locus (A_1, A_2) and (A_1, A_3), because the interest is on making comparisons between the rare alleles and that more common. We are not interested in knowing the association between two rare allele (may be one is of risk and the other of protection or one is neutral and the other of risk, …). Then it is sufficient to repeat the simple test for both the possible associations: rare1-common, rare2-common (see in Fig. 5).

Fig. 5: Multiallelic problem.



Genotype	**Case**	**Control**
A_1A_1	$f_{A1A1\ ca}$	$f_{A1A1\ co}$
A_1A_2	$f_{A1A2\ ca}$	$f_{A1A2\ co}$
A_2A_2	$f_{A2A2\ ca}$	$f_{A2A2\ co}$
A_1A_3	$f_{A1A3\ ca}$	$f_{A1A3\ co}$
A_3A_3	$f_{A3A3\ ca}$	$f_{A3A3\ co}$
A_2A_3	$f_{A2A3\ ca}$	$f_{A2A3\ co}$
	n_{ca}	n_{co}

First test using the (3×2) table given by the two allele A_1A_2.

First test using the (3×2) table given by the two allele A_1A_3.

Not of interest

More complicated is the situation where the association study involves more than two locus, as (*a*,*A*), where *A* is the rarest, and (*b*,*B*), where *B* is the rarest. We suppose the interest is in knowing the specific effect of all the multiple possible configurations (see Fig. 6).

Fig. 6: Multiloci extension (1).

Genotype	**Case**	**Control**
Aa, bb	$f_{aa,bb\ ca}$	$f_{aa,bb\ co}$
Aa, Bb	$f_{aa,Bb\ ca}$	$f_{aa,Bb\ co}$
Aa, BB	$f_{aa,BB\ ca}$	$f_{aa,BB\ co}$
Aa, bb	$f_{Aa,bb\ ca}$	$f_{Aa,bb\ co}$
Aa, Bb	$f_{Aa,b\ \ ca}$	$f_{Aa,b\ \ co}$
Aa, BB	$f_{Aa,BB\ ca}$	$f_{Aa,BB\ co}$
AA, bb	$f_{AA,\ \ ca}$	$f_{AA,\ \ co}$
AA, Bb	$f_{AA,B\ \ ca}$	$f_{AA,B\ \ co}$
AA, BB	$f_{AA,BB\ ca}$	$f_{AA,BB\ co}$
	n_{ca}	n_{co}

The main topic in this situation is to reconstruct the possible effect that may have one locus, given a specific configuration of the other locus. Then we use six different (3×2) contingency tables, one for any specific configuration (see Fig. 7):

Fig. 7: The possible configurations.

1) *aa*	Ca	Co
bb	...	...
Bb	...	...
BB	...	...

2) *Aa*	Ca	Co
bb	...	...
Bb	...	...
BB	...	...

3) *AA*	Ca	Co
bb	...	...
Bb	...	...
BB	...	...

4) *bb*	Ca	Co
aa	...	...
Aa	...	...
AA	...	...

5) *Bb*	Ca	Co
aa	...	...
Aa	...	...
AA	...	...

6) *BB*	Ca	Co
aa	...	...
Aa	...	...
AA	...	...

For example, in the first table, we make the nonparametric test (by CMC-Procedure or CEP) to study the association at locus (*b*,*B*) conditional to the genotype *aa* (more common) in the other locus. This procedure may seem a simple extension to the simple case, but if we consider the case of two loci with three alleles each one, then, in this case, we obtain two tables for each of the six configurations; and if we are more than two loci together the analysis of each type of association may be very difficult.

In the last situation, before executing the specific test for each configuration, it is helpful to make a single test to study if there is any type of significant association in one, at least, of all the configurations (it does not matter, for the moment, if it will be of risk or protection). Then, we may suppose that *k* polymorphic genes are jointly examined and that (with the usual notation) $\{(aa)_r, (Aa)_r, (AA)_r, r = 1,\ldots,k\}$ is the set of related genotypes. In this situation, we express the null hypothesis in terms of odd ratios, as:

$$H_0 : \left\{ \bigcap_{r=1}^{k} [(\theta_{Aa\,r} = 1) \cap (\theta_{aa\,r} = 1)] \right\},$$

which means that all *k* genes are jointly irrelevant for discrimination. The alternative of interest may assume two different expressions. The first is

$$H_1 : \left\{ \bigcup_{r=1}^{k} \begin{bmatrix} (\theta_{Aa\,r} \geq 1) \cap (\theta_{aa\,r} \geq 1) \\ XOR \\ (\theta_{Aa\,r} \leq 1) \cap (\theta_{aa\,r} \leq 1) \end{bmatrix} \right\},$$

where, of course, at least one inequality in each of the 2 × *k* lines is strict. The interpretation of this alternative is that there exists at least one gene which is relevant for discriminating cases with respect to controls. The aim of this alternative is not to know if all genes are of risk (XOR protection), but if we can admit that some genes may be of risk, others of protection, the remaining being neutral.

In order to solve this specific problem, let us suppose:

a) data are organized in a unit-by-unit representation: $\{Y_{jir}, r = 1,\ldots,k, i = 1,\ldots,n_j, j = case, control\}$, where Y_{jir} is the genotype of *r*-th gene on *i*-th subject of *j*-th group (that is Y_{jir} may assume one of the values: *aa*, *Aa*, *AA*);

b) permutations exchange units (by CMC or CEP) between groups, so that *k*-dimensional vectors are exchanged;

c) for each gene *r*, *r* = 1,...,*k*, calculate partial tests as $T^*_{r\,Aa} = f^*_{r\,AA\,case} f^*_{r\,Aa\,control} / (f^*_{r\,Aa\,case} f^*_{r\,AA\,control})$ and $T^*_{r\,aa} = f^*_{r\,Aa\,case} f^*_{r\,aa\,control} / (f^*_{r\,aa\,case} f^*_{r\,Aa\,control})$, *r* = 1,..., *k*, that are all significant for either large or small values;

d) within each gene calculate second order combined test and related p-value $\hat{\lambda}_r^{''}$, in accordance with the method previously discussed;

e) according to the nonparametric combination theory, we combine k second order transformed p-values $1-|2\hat{\lambda}_r^{''}-1|$ through any combining function ψ to obtain a third order overall combined test and related p-value $\hat{\lambda}^{'''}$;

f) if $\hat{\lambda}^{'''} \leq \alpha$, then reject the overall null hypothesis.

A second kind of alternative of interest is:

$$H_1^{'}:\left\{\begin{matrix} \bigcup_{1\leq r\leq k}[(\theta_{Aa\,r}\geq 1)\bigcap(\theta_{aa\,r}\geq 1)] \\ XOR \\ \bigcup_{1\leq r\leq k}[(\theta_{Aa\,r}\leq 1)\bigcap(\theta_{aa\,r}\leq 1)] \end{matrix}\right\},$$

where again at least one inequality in each line is strict. It means that there is at least one gene which is of protection (XOR risk), whereas others are neutral.

Again, in order to solve the problem, we should modify steps e) and f) respectively into:

e') according to the nonparametric combination theory, combine k second order p-values $\hat{\lambda}_r^{''}$ through any suitable combining function to obtain a proper third order overall combined test and related p-value $\hat{\lambda}_1^{'''}$;

f) if $1-|2\hat{\lambda}_1^{'''}-1|\leq\alpha$, then reject the overall null hypothesis.

The third order combined tests and their p-values are always obtained by the same conditional simulation (by CMC-Procedure or CEP) results used for obtaining distributions of partial tests $T_{r\,h}^{*}$ and p-values $\hat{\lambda}_{r\,h}$ and $\hat{\lambda}_r^{''}$, $h = aa$, Aa, AA, r = 1,...,k.

3.6 Allelic Association Studies with Confounding Effects

Suppose now, to have a confounding factor that may have some effects on a particular pathology (for example, in individuals which live a contact to different levels of exposure to a specific agent). The study can involve one genotype (but the extension to more than one is immediate) that seems to be associated to that disease. We can represent the data as in Fig. 8.

Fig. 8: Confounding Factor with Two Levels.

Confounding Factor

	Exposure Less (-)		Exposure High (+)		
Genotype	Case	Control	Case	Control	
aa	f_1	F_2	f_3	f_4	n_{aa}
Aa	f_5	F_6	f_7	f_8	n_{Aa}
AA	f_9	f_{10}	f_{11}	f_{12}	n_{AA}
	$n_{Ca,-}$	$n_{Co,-}$	$n_{Ca,+}$	$n_{Co,+}$	

As a first situation we consider only two levels in the confounding factor. So, by using the usual odd ratios, we obtain: $\theta_{aa} = (f_3 f_2)/(f_4 f_1)$ that is the odd we use by referent, $\theta_{Aa} = (f_7 f_6)/(f_8 f_5)$ and $\theta_{AA} = (f_{11} f_{10})/(f_{12} f_9)$. These odds are estimated points of the level of association that is present in any genotype configuration between the different groups of exposure. Then, we may observe, if the single odds are different from one, where an association between the exposure levels and the genotype associated (or not) to the disease is present. Of course, this correspond to an usual stratification on a variable of stratum (in case-control association studies this is very common, as for factors like age, sex, ...).

We obtained three tests of hypothesis to testing if $\theta_x = 1$ or $\theta_x \neq 1$, where $x = aa, Aa, AA$.

In taking into consideration the problem of isotonic inference, we can continue in the analysis and make the second ratios by using the first odds: $K_{aa} = \theta_{aa} / \theta_{aa} = 1$, $K_{Aa} = \theta_{Aa} / \theta_{aa}$ and $K_{AA} = \theta_{aa} / \theta_{AA}$.

Then the hypotheses of interest are:

$$H_0 : \{K_{aa} = K_{Aa} = K_{AA} = 1\}$$

and

$$H_1 : \{(K_{aa} \leq K_{Aa} \leq K_{AA} \leq 1) \; XOR \; (K_{aa} \geq K_{Aa} \geq K_{AA} \geq 1)\},$$

where at least one inequality is strong. The alternative is constructed in a way such that it corresponds to the possible effects that the confounding factor may have jointly to the putative allele on the affected individuals with respect to the cases. In effect, if an higher level of exposure is such that it increases the effects of the putative rare allele (or it have some influence in other factors correlated to that one), the risk should increase, anyway, in homozygous rare subjects exposed to the factor, with respect to the heterozygous exposed individuals. So that, we do not admit the points that fall in the areas as $\{K_{aa} \geq K_{Aa} \leq K_{AA} \leq 1\}$, for example.

If we consider a more general environmental factor or a variable of stratum, where there are three different levels that we may order in some way, the data become as those in Fig. 9.

In this case, we need doing six partial odd ratios: $\theta_{aa} = f_3 f_2 /(f_4 f_1)$, $\theta_{Aa} = f_9 f_8 /(f_{10} f_7)$, $\theta_{AA} = f_{15} f_{14} /(f_{16} f_{13})$ and $\theta'_{aa} = f_5 f_4 /(f_6 f_3)$, $\theta'_{Aa} = f_{11} f_{10} /(f_{12} f_9)$, $\theta'_{AA} = f_{17} f_{16} /(f_{18} f_{15})$, that may be tested in the same way of the previous.

Fig. 9: Three Levels of Confounding.

Confounding Factor

	Exposure Level 1 (-)		Exposure Level 2 (+)		Exposure Level 3 (++)		
Genotype	Case	Control	Case	Control	Case	Control	
aa	f_1	f_2	f_3	F_4	f_5	F_6	n_{aa}
Aa	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	n_{Aa}
AA	f_{13}	f_{14}	f_{15}	f_{16}	f_{17}	f_{18}	n_{AA}
	$n_{Ca,-}$	$n_{Co,-}$	$n_{Ca,+}$	$n_{Co,+}$	$n_{Ca,++}$	$n_{Co,++}$	

The related second ratios become: $K_{aa} = \theta_{aa} / \theta_{aa} = 1$, $K_{Aa} = \theta_{aa} / \theta_{Aa}$, $K_{AA} = \theta_{aa} / \theta_{AA}$ and, for the second group, $K' = \theta' / \theta' = 1$, $K'_{Aa} = \theta'_{aa} / \theta'_{Aa}$, $K'_{AA} = \theta'_{aa} / \theta'_{AA}$. So, there are two sets of isotonic hypotheses:

$$H_0 : \{K_{aa} = K_{Aa} = K_{AA} = 1\}$$

against

$$H_1 : \{(K_{aa} \le K_{Aa} \le K_{AA} \le 1)\ XOR\ (K_{aa} \ge K_{Aa} \ge K_{AA} \ge 1)\},$$

and

$$H_0 : \{K'_{aa} = K'_{Aa} = K'_{AA} = 1\}$$

against

$$H_1 : \{(K'_{aa} \le K'_{Aa} \le K'_{AA} \le 1)\ XOR\ (K'_{aa} \ge K'_{Aa} \ge K'_{AA} \ge 1)\},$$

where all equations have a clear meaning.

Further, we may have an overall parameter that estimates the general association in the contingency table:

$$K^{FIN} = \frac{K'_{AA} K_{Aa}}{K'_{Aa} K_{AA}}.$$

If this value is equal to one, it means that there is no association in the data, if it is different, one type of association is present (now, it is not important which type).

To increase the difficulty, we can consider the case where there are two stratification or confounding factors with three levels each one and the genetic locus under study is three-allelic (*a*, the more common gene, *b*, *c*).

	V_1						V_2						V_3					
	U_1		U_2		U_3		U_1		U_2		U_3		U_1		U_2		U_3	
gen.	CA	CO	CA	CO	CA	CO	CA	CO	CA	CO	CA	CO	CA	CO	CA	CO	CA	CO
aa	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ab	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
bb	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ac	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
cc	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
bc	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

In this case we can use the following algorithm to obtain the *p*-values associated to the tests on the factor U conditioned to the levels of the factor V:

```
for s=1 to 3
 for j=1 to 6
   for k=1 to 2
5   Test1U(s,j,k)=MATRIX(j,s,k,2) MATRIX(j,s,k+1,1)/(MATRIX(j,s,k,1)MATRIX(j,s,k+1,2))
   next k
 next j
10 for k=1 to 2
    Test2Uab(s,k) = Test1U(s,1,k) / Test1U(s,2,k)
    Test2Ubb(s,k) = Test1U(s,1,k) / Test1U(s,3,k)
    Test2Uac(s,k) = Test1U(s,1,k) / Test1U(s,4,k)
    Test2Ucc(s,k) = Test1U(s,1,k) / Test1U(s,5,k)
   next k
11 Test3b = Test2Uab(s,2) Test2Ubb(s,1) / (Test2Uab(s,1) Test2Ubb(s,2))
12 Test3c = Test2Uac(s,2) Test2Ucc(s,1) / (Test2Uac(s,1) Test2Ucc(s,2))
next
```

where in the row 5 we obtain all the observed difference tests θ_{ij}, $i = 1$ or 2 because we have three levels and so we compare level U_1 against Level U_2 and U_2 against U_3, for all the genotypes of interest ($j = aa, ab, bb, ac, cc$); the 4 statements following the loop for in line 10 are the observed tests $K_{i1} = \theta_{i1} / \theta_{i2}$ and $K_{i2} = \theta_{i1} / \theta_{i3}$ for $i = 1, 2$. Finally, tests in row 11 and row 12 are K^{FIN} associated rare allele *b* and rare allele *c*. These results are repeated for all the levels of V. At the same way we operate by testing V|U.

3.7 The Hardy-Weinberg Equilibrium

In classical population case-control studies the researcher often is led to substitute the unrelated observed controls with the controls he obtains after assuming the Hardy-Weinberg equilibrium (1908). In brief, the Hardy-Weinberg law says that if we have a bi-allelic gene with a frequency of the rare allele A of p (obviously $q = 1 - p$ is the frequency of the more common allele a), then the expected genotype frequencies into the population (by assuming the typical conditions: diploid population, sexual reproduction, random mating, discrete generations, no mutation, no migration, very large population, no selection) are $f_{AA} = p^2$, $f_{Aa} = 2pq$ and $f_{aa} = (1 - p)^2$.
Therefore, supposing to have the following table of data:

Genotypes	Cases	Controls
AA	X_1	Y_1
Aa	X_2	Y_2
Aa	X_3	Y_3
Total	M	N

the observed controls can be changed with the expectative frequencies under H.-W. equilibrium:

Genotypes	Cases	Controls
AA	X_1	$Y_1^1 = N[(2Y_1 + Y_2)/(2N)]^2$
Aa	X_2	$Y_2^1 = 2N[(2Y_1 + Y_2)/(2N)][(2Y_3 + Y_2)/(2N)]$
aa	X_3	$Y_3^1 = N[(2Y_3 + Y_2)/(2N)]^2$

and the relative risks can be estimated by using the odds ratios obtained from this last table. This is done very frequently in literature and it showed produce more powerful results.

Furthermore, the parametric association studies by using the case-control tables may consider also another type of expected frequencies under Hardy-Weinberg equilibrium. As is done by Lathrop (1983) and also by Chiano and Clayton (1998), if we consider the null hypothesis of equal distributions in cases and controls, so that the odds ratios are both equal to one, we can assume the Hardy-Weinberg equilibrium exists in all the data of the table, and in controls and in cases. The expected controls under the null hypothesis are then constructed in this way:

Genotypes	Cases	Controls
AA	X_1	$Y_1^2 = N[(2X_1 + X_2 + 2Y_1 + Y_2)/(2M + 2N)]^2$
Aa	X_2	$Y_2^2 = 2N[(2X_1 + X_2 + 2Y_1 + Y_2)(2X_3 + X_2 + 2Y_3 + Y_2)/(2M + 2N)^2]$
aa	X_3	$Y_3^2 = N[(2X_3 + X_2 + 2Y_3 + Y_2)/(2M + 2N)]^2$

So that the likelihood ratio test may be constructed by considering the ratio between the frequencies under H_1, which are the frequencies observed from the second table, and the frequencies under H_1, which are the expected frequencies we obtain from the third table. This solution may be more powerful and robust with respect to spurious association (Lathrop, 1983, Chiano and Clayton, 1998).

In the permutation solution we consider the adjustment for controls on Hardy-Weinberg equilibrium based on the correction only for the population of controls individuals. This is done because the configuration of the data which involves the distribution equality between cases and controls, and so also the adjustment for the controls in H-W based on the global population, is only a (may be more than one) special permutation we obtain by changing the data in the table, therefore it is included in the analysis. Instead, at the moment, we do not have completed the likelihood based solution with this adjustment, because it need change the asymptotical distribution of the statistic.

4. Power and Sample Size Simulations

4.1 General Considerations

Many comparisons between population-based and family-based case-control studies have been done in the literature, furthermore in some cases also the differences between parametric and nonparametric solutions have been considered. Generally the results are not definitive because the relations do not show a strong dominancy of a solution with respect another one, but sometimes there is a better performance of one with respect the other, according with the sample size we consider, the frequency of rare allele in the population, the genetic model of the disease (dominant, codominant, recessive). In particular, the first type of comparison may be very difficult, because association studies which take in consideration related controls (usual TDT or the S-TDT) have take more assumptions.

Anyway, first we present some simulations for the nonparametric permutation solution illustrated in Sec. 3.1, by considering different types of population parameters and genetic models. The likelihood solution presented in Sec. 3.4 was not studied in-depth, because its asymptotical distribution now is not known very well.

Next, we extend the considerations to the comparisons between the nonparametric population-based method (Sec. 3.1) and the sibship transmission disequilibrium test (S-TDT).

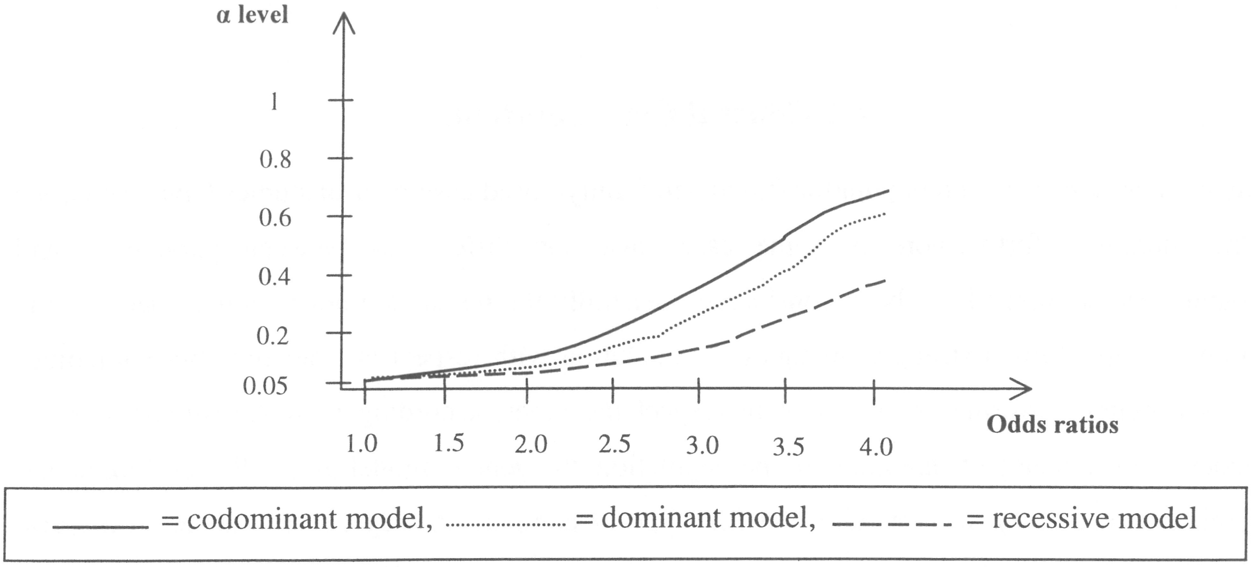
4.2 Simulations for Nonparametric Population-Based Solutions

We performed many simulations by considering different parameter types (allelic frequency in the population, the three genetic models for the allele effect, several values of the odds ratios) for the permutation solution. The number of simulation is always fixed to 1000 and the number of conditional Monte Carlo iterations (CMC-Iterations) is again 1000.

Simulations are performed by using one single locus with two alleles, one more common and one rare, as in the problem introduced in Sec. 1.4 and Sec. 3.1.

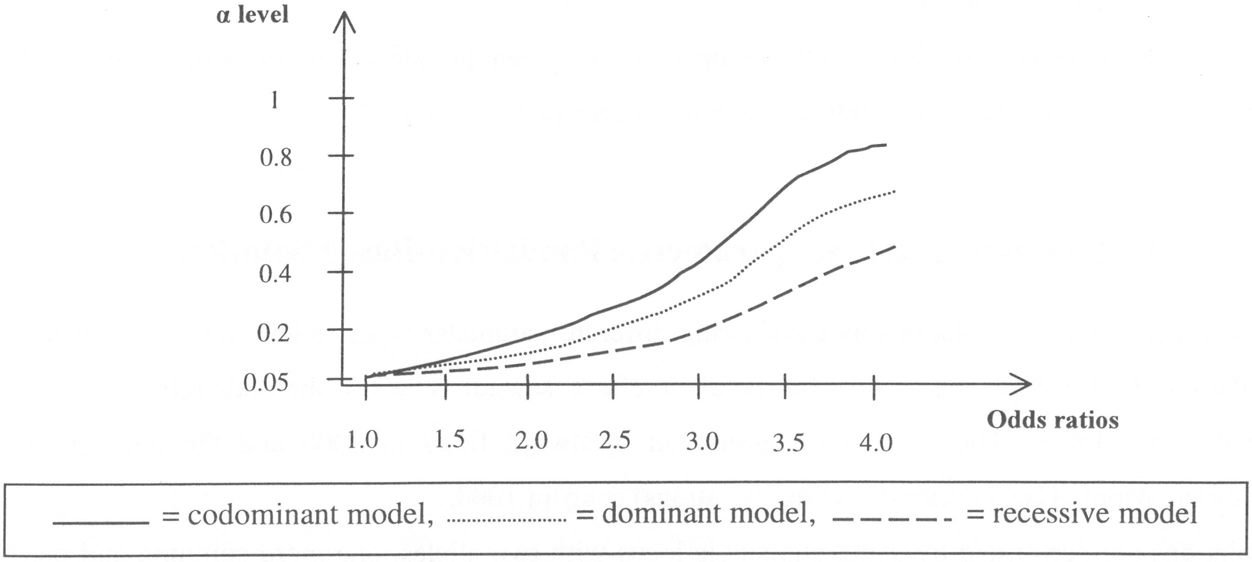
In Fig. 1 we show the power simulations for the nonparametric solutions with cases = controls = 50 and the frequency of rare allele is 0.05.

Fig. 1: Case = controls = 50, $f = 0.05$.



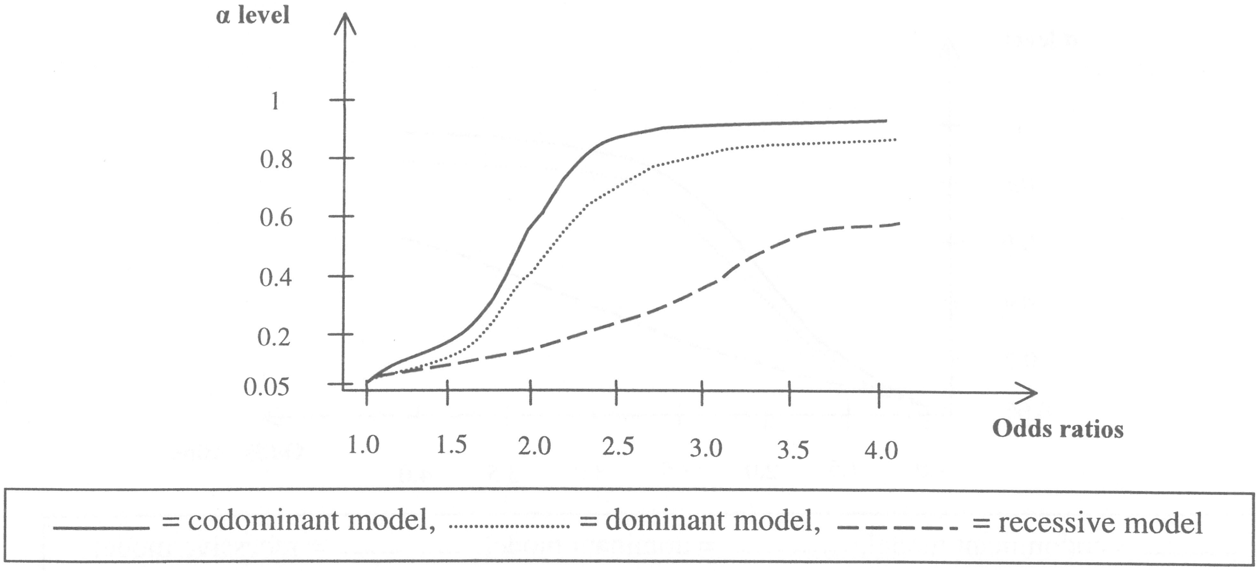
In Fig. 2 we show the power simulations for the nonparametric solutions with cases = controls = 100 and the frequency of rare allele is 0.05.

Fig. 2: Case = controls = 100, $f = 0.05$.



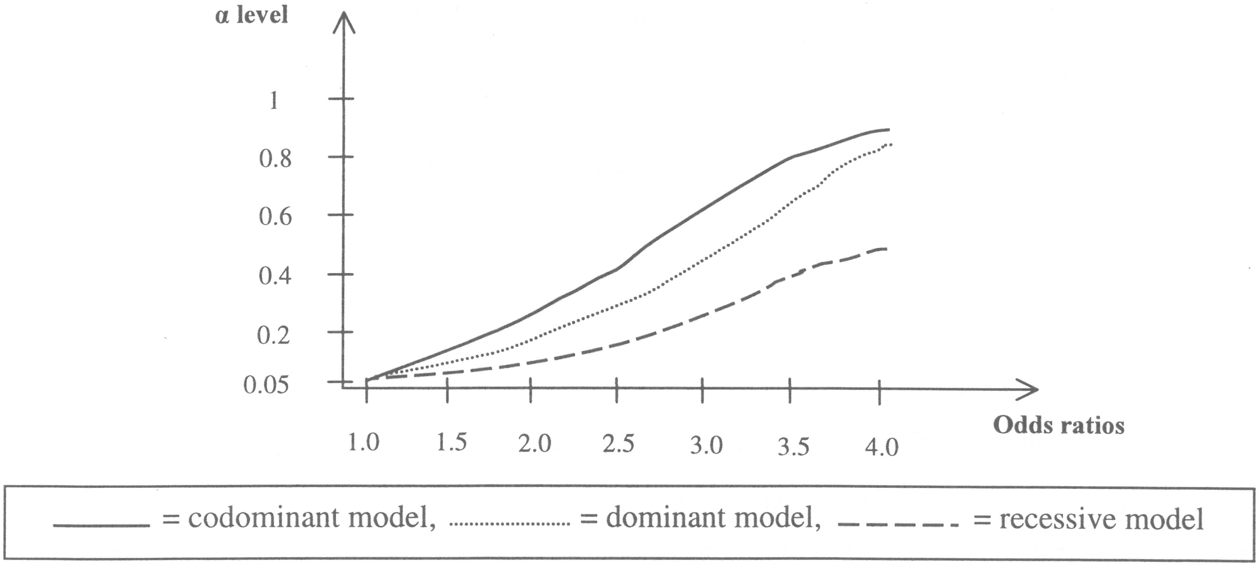
In Fig. 3 we show the power simulations for the nonparametric solutions with cases = controls = 500 and the frequency of rare allele is 0.05.

Fig. 3: Case = controls = 500, $f = 0.05$.



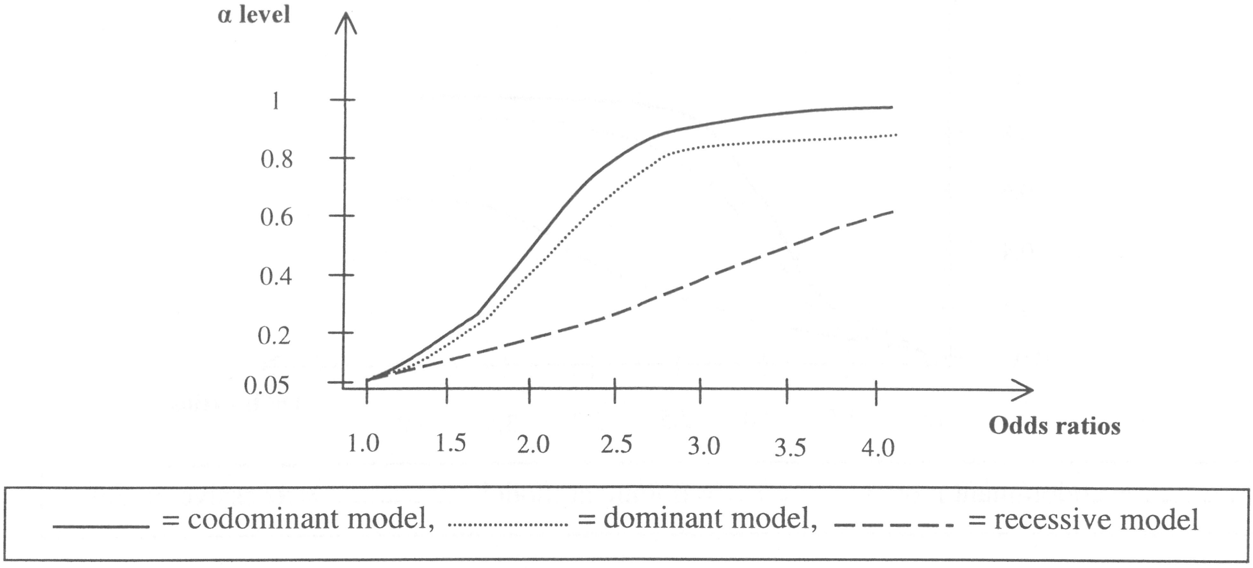
In Fig. 4 we show the power simulations for the nonparametric solutions with cases = controls = 50 and the frequency of rare allele is 0.10.

Fig. 4: Case = controls = 50, $f = 0.10$.



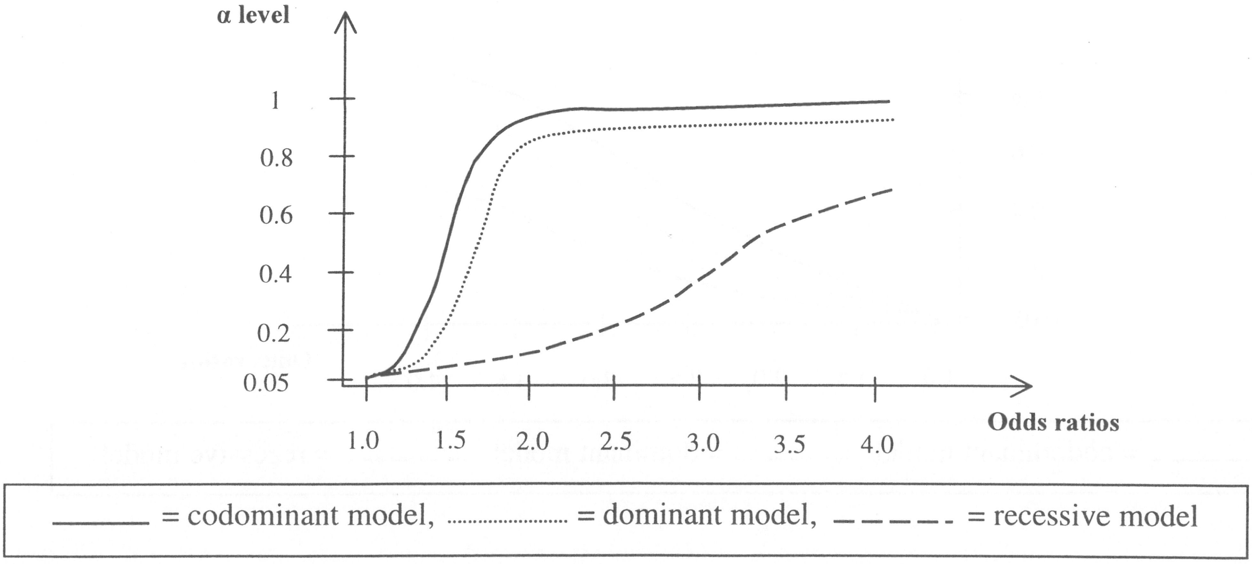
In Fig. 5 we show the power simulations for the nonparametric solutions with cases = controls = 100 and the frequency of rare allele is 0.10.

Fig. 5: Case = controls = 100, f = 0.10.



In Fig. 6 we show the power simulations for the nonparametric solutions with cases = controls = 500 and the frequency of rare allele is 0.10.

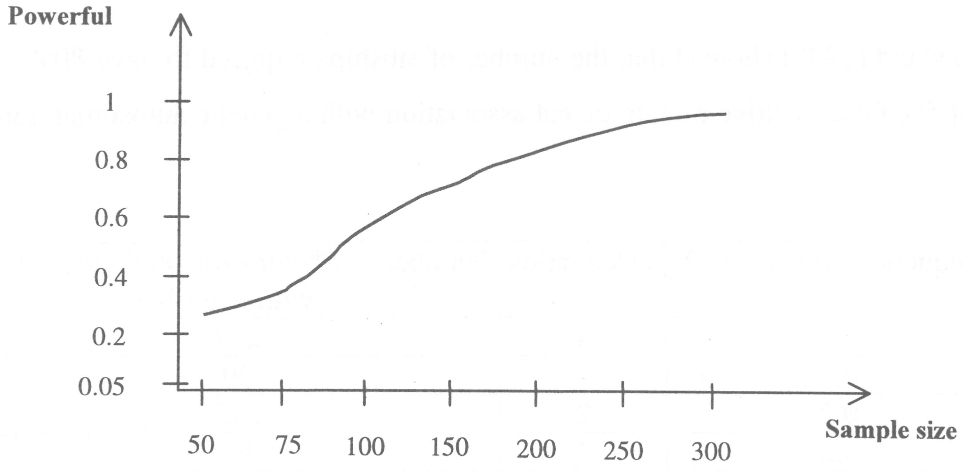
Fig. 6: Case = controls = 500, f = 0.10.



As we can observe from the previous figures, the power of nonparametric solution is very good also for small sample size, and for low frequency of rare allele. Of course, the situation where the rare allele is recessive is the worst, and generally, in this case, the number of subjects, we need to perform an analysis of association, is very large.

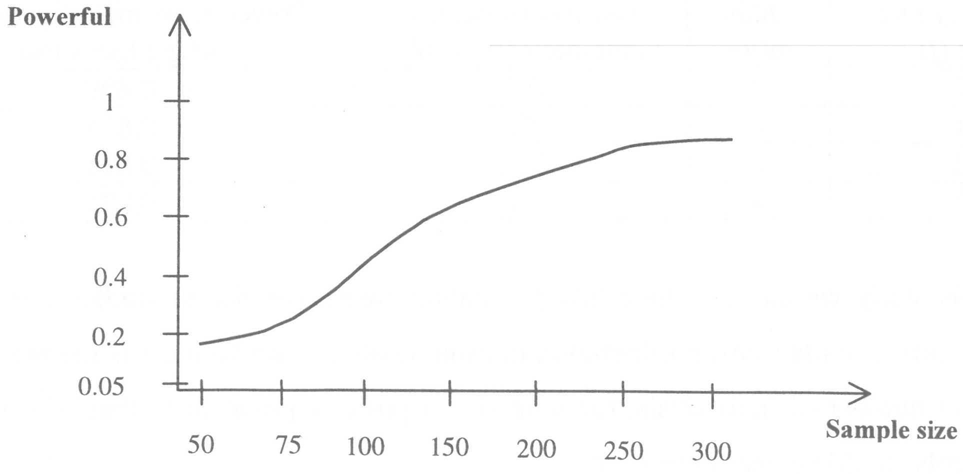
In Fig. 7, we consider a simulation study for the permutation test with a frequency of 0.10 for the rare allele and significance level α fixed to 0.05 for the case where the rare allele is codominant and the odds are 2.

Fig. 7: Codominant model with odds ratios = 2, $f = 0.10$, $\alpha = 0.05$.



In Fig. 8, we consider simulations with a frequency of 0.10 for the rare allele and significance level α = 0.05 for the case where the rare allele is dominant and the odd is 2.

Fig. 8: Dominant model with odds ratios = 2, $f = 0.10$, $\alpha = 0.05$.



How we can note from the figures, the nonparametric permutation solution have very good power behaviour.

A priori, we cannot know what is the best type of test for a specific case, because it may depends from the specific table of data we have: furthermore in real dataset the samples of cases and controls are not generally equal and with those sample sizes the solutions may be quite different about their power.

Anyway, generally, for small sample sizes, nonparametric tests should be preferred to their parametric counterparts.

4.3 Comparison Between S-TDT and Population-Based Methods

Shaid and Rowland (1998) showed that the number of sibships required to have 80% power, under assumption of 5% false-positive rate, to detect association with a genetic autosomal dominant locus disease are:

Frequency rare allele (*f*)	Odds ratios	Number of sibships for single locus test with one sib
0.1	2	257
0.1	4	59
0.2	2	212
0.2	4	57

We performed the simulations with the permutation test by using the same parameters for allele frequencies, odds ratios, significance level and assuming the dominant model. The results are shown in the following table.

Frequency rare allele (*f*)	Odds ratios	Number of cases = Number of controls	Power of permutation test for single locus test
0.1	2	257	0.829
0.1	4	59	0.821
0.2	2	212	0.811
0.2	4	57	0.831

From this study we can conclude that population-based association studies could be more powerful with same sample sizes. Furthermore, in using related controls, there is the big problem to find a sufficient number of individuals, for achieving a power equivalent to that of methods with unrelated controls, could be very expensive.

5. Case Study

5.1 Background

High blood levels of coagulation factor VII gene represent a risk factor for ischaemic cardiovascular disease. This is supported by many studies during the past decade. The Northwick Park Heart Study group reported that high levels of factor VII were independently associated with an increase in the risk of coronary events in middle-aged men (Meade *et Al.*, 1986) and more recently they showed that the level of factor VII was predictive of the risk of fatal but not nonfatal myocardial infarction (Meade *et Al.*, 1993). The same trend was observed in the Prospective Cardiovascular Münster Study (Heinrich *et Al.*, 1994). However, other investigations failed to find such associations (Cortellaro *et Al.*, 1992, Vaziri *et Al.*, 1992). Factor VII blood levels are influenced by both environmental and genetic factors as age, body-mass index, oral contraceptive use, etc. (see for reference Iacoviello *et Al.*, 1998). Here, we refer to works of Iacoviello *et Al.*, 1998, and Di Castelnuovo *et Al.*, 2000, where they investigate whether the risk of myocardial infarction is associated with polymorphisms in the factor VII gene and whether these polymorphisms are associated with factor VII levels (Iacoviello *et Al.*, 1998); furthermore they consider the role of the decanucleotide insertion/deletion functional polymorphism in the promoter region of factor VII gene and of possible interactions of promoter with HVR4 intron polymorphism (see for reference Di Castelnuovo *et Al.*, 2000).

5.2 Study Population

They studied patients with a family history of thrombosis. Case patients were persons over 45 years who had a myocardial infarction and who were reported to have at least one first-degree relative who had a myocardial infarction or a stroke (or both) before the age of 65 years. The patients were selected from the "Gruppo Italiano per lo Studio della Sopravvivenza nell'Infarto Miocardico" (GISSI) trial population on the basis of an interview regarding their family history of thrombosis.
Controls were consecutive patients over the age of 45 years who were hospitalized for any clinical reason except myocardial infarction, stable or unstable angina, stroke, or transient ischaemic

attacks. Furthermore they could not report a personal or family history of thrombosis or have definite defects of the haemostatic system. For more details, refer to Iacoviello *et Al.*, 1998.

5.3 Laboratory Measurements and Techniques

The cases were assessed five to seven months after their most recent ischaemic event. Blood was collected from the patients between 8 and 10 a.m. after an overnight fast and after a 20-minute rest in the supine position. Blood was not collected from patients who were receiving oral anticoagulant drugs. Blood samples for DNA and biochemical analyses were available for 165 of 171 cases and 225 of 272 controls.

So they studied three polymorphisms for the factor VII gene:

1) Hypervariable region 4 of intron 7 of the factor VII gene was amplified (see Iacoviello *et Al.*, 1998) and three alleles were identified: a common allele (H6) of 443 bp with six monomers, a less frequent allele (H7) of 480bp with seven monomers of 37bp and a very rare allele (H5) of 406 bp with five monomers.
2) Two fragments were detected for the R353Q polymorphism: the more common (the R allele) of 205 bp and the rarest (the Q allele) of 272 bp.
3) For the promoter region they considered the normal allele (1) and the rare allele (2) with the deletion.

For more details, refer to Iacoviello *et Al.*, 1998, and Di Castelnuovo *et Al.*, 2000.

5.4 Results of Statistical Analyses

From the data, we obtained the following case-control tables for the polymorphisms of factor VII loci.

Table 1

R353Q	Cases	Controls	Total
QQ	1	10	11
RQ	49	76	125
RR	114	138	252
Total	164	224	388

Table 2

Promoter region	Cases	Controls	Total
22	1	4	5
12	42	75	117
11	119	140	259
Total	162	219	381

Table 3

Hypervariable Region 4	Cases	Controls	Total
H7H7	12	31	43
H6H7	60	97	157
H6H6	84	94	118
H7H5	4	1	5
H6H5	5	2	7
Total	165	225	390

In table 1, we see the data for the R353Q polymorphism, where R is the more common allele: here, there is one missing in both cases and controls. In table 2, we see the data for the promoter region polymorphism of the factor VII gene: in this case the more common allele is the 1 allele and there are three missing data in the cases and six in the controls. In table 3, there are the data for the hypervariable region 4, where H6 is the more common allele whereas H7 and H5 are the rarest alleles.

Only the data for table 2 are not in Hardy-Weinberg equilibrium, whereas, in table 1 and 3, they follow the Hardy-Weinberg law; furthermore, it was proven that R353Q and hypervariable region 4 polymorphisms are in linkage disequilibrium (Iacoviello, 1998).

Consider the odds ratios for myocardial infarction with each of the three table (see tables 4, 5 and 6), by imposing too "perfect" Hardy-Weinberg equilibrium in the controls.

Table 4

R353Q	Estimated Relative Risks
Odd QQ/RQ	0.155
Odd QQ/RQ in H-W	0.153
Odd RQ/RR	0.780
Odd RQ/RR in H-W	0.797

Table 5

Promoter	Estimated Relative Risks
Odd 22/12	0.446
Odd 22/12 in H-W	0.199
Odd 12/11	0.659
Odd 12/11 in H-W	0.759

Table 6

Hypervariable region 4	Estimated Relative Risks
Odd 77/67	0.626
Odd 77/67 in H-W	0.729
Odd 67/66	0.692
Odd 67/66 in H-W	0.644
Odd 55/65	Infinity
Odd 55/65 in H-W	Infinity
Odd 65/66	2.798
Odd 65/66 in H-W	2.738

Now, if we perform the simple permutation test illustrated in Sec. 3.1, by looking at the odds ratios of table 4 and 5, we obtain (by using 10000 CMC-iterations and the Liptak nonparametric combination) the *p*-values that are represented in Fig. 1. At the same way, in Fig. 1 are represented the results with the multiallelic permutation test of Sec. 3.5.

Fig. 1: *p*-values of simple nonparametric test.

Test	***p*-value**
R353Q:{(Odd QQ/RQ<1)∩(Odd RQ/RR<1) xor (Odd QQ/RQ>1)∩(Odd RQ/RR>1)}	0.013
R353Q in exact H-W equilibrium	0.009
Promoter:{(Odd 22/12<1)∩(Odd 12/11<1) xor (Odd 22/12>1)∩(Odd 12/11>1)}	0.129
Promoter in exact H-W equilibrium	0.036
Hypervariable:{(Odd 77/67<1)∩(Odd 67/66<1) xor (Odd 77/67>1)∩(Odd 67/66>1)}	0.015
Hypervariable region for 7 allelic form in exact H-W equilibrium	0.016
Hypervariable:{(Odd 55/65<1)∩(Odd 65/66<1) xor (Odd 55/65>1)∩(Odd 65/66>1)}	0.250
Hypervariable region for 5 allelic form in exact H-W equilibrium	0.250

Results are significant for the Q allele of the R353Q polymorphism at the level $\alpha = 0.05$, so that we can say that Q allele has a protective effect (Odd < 1) against the myocardial infarction; furthermore, by looking at the two estimated relative risks, we may consider this as an effect of codominant type. The equivalence of odds ratios between observed and H-W adjusted controls supports the above observation that controls are in H-W equilibrium.

The p-value for the protective region polymorphism is not significant for the observed data, but is 0.036 with the controls adjusted by H-W equilibrium. Anyway, we cannot infer the 2 allele produce a protective effect. Finally, we can note the strong protective effect of the 7 allelic form into the hypervariable region 4 polymorphism of the factor VII region. Also this allele is protective against the myocardial infarction and the effect seems to be codominant. Allele 5 does not present any particular effect.

Now, we consider the analysis of data by using the multiloci extension of permutation test illustrated in Sec. 3.5. The three following tables represent the analysis of haplotypes: hypervariable & promoter region polymorphisms (table 7), hypervariable region & R353Q polymorphisms (table 8), promoter region & R353Q polymorphisms (table 9).

Table 7

Hypervariable & promoter	Cases	Controls	Total
66-11	75	74	149
66-12	7	18	25
66-22	0	0	0
67-11	37	55	92
67-12	21	38	59
67-22	1	1	2
77-11	2	8	10
77-12	10	19	29
77-22	0	3	3
65-11	3	2	5
65-12	2	0	2
65-22	0	0	0
55-11	0	0	0
55-12	0	0	0
55-22	0	0	0
75-11	2	1	3
75-12	2	0	2
75-22	0	0	0
Total	162	219	381

We studied the association of each locus by conditioning on the other locus, and we use the same isotonic system of hypotheses that we did above. We performed a permutation test for obtaining combined p-values which test each pair of odds ratios for one locus conditional to a specific genotype of the other locus. Therefore, we performed permutation tests by using 10000 CMC-iterations and the combining function of Liptak. In this case we did not use to adjust the controls by the Hardy-Weinberg law, because there are linkage disequilibrium between the two loci and we are studying for their possible associated effect.

From the data in table 7 we obtained six *p*-values (we do not report the results that include the hypervariable region 4 polymorphism with the allelic form 5, because the relative sample size is too low):

1) the first *p*-value tests jointly the two odds ratios of the promoter region polymorphism conditionally to the genotypic form 66 of the hypervariable region 4 polymorphism, we denotes it by the name *p*-value (66-..) and it result 0.054, not significant at the level $\alpha = 0.05$, so we can not say that the 2 allele in the promoter region polymorphism present one particular variation of effect if it is associated to the form 66 of the hypervariable region 4 polymorphism.
2) *p*-value (67-..) is 0.887, we accept the null hypothesis that there is not any variation of effect with the allele 2 in the promoter region polymorphism conditionally to the genotypic form 67 of the hypervariable region 4 polymorphism.
3) *p*-value (77-..) is 0.984, we accept the null hypothesis that there is not any variation of effect with the allele 2 in the promoter region polymorphism conditionally to the genotypic form 77 of the hypervariable region 4 polymorphism.
4) *p*-value (7.-11) is 0.024, we have a significant result that there is an increasing protective effect (the two odds are 0.37 and 0.66) with the 7 allele in the hypervariable region 4 polymorphism conditionally with the genotypic form 11 of the promoter region polymorphism.
5) *p*-value (7.-12) is 0.621, we accept the null hypothesis that there is not any variation of effect with the 7 allele in the hypervariable region 4 polymorphism conditionally with the genotypic form 12 of the promoter region polymorphism.
6) *p*-value (7.-22) is 0.603, we accept the null hypothesis that there is not any variation of effect with the 7 allele in the hypervariable region 4 polymorphism conditionally with the genotypic form 22 of the promoter region polymorphism.

Table 8

Hypervariable & R353Q	Cases	Controls	Total
66-RR	77	78	155
66-RQ	6	15	21
66-QQ	0	1	1
67-RR	28	50	78
67-RQ	32	45	77
67-QQ	0	2	2
77-RR	3	7	10
77-RQ	8	16	24
77-QQ	1	7	8
65-RR	4	2	6
65-RQ	1	0	1
65-QQ	0	0	0
55-RR	0	0	0
55-RQ	0	0	0
55-QQ	0	0	0
75-RR	2	1	3
75-RQ	2	0	2
75-QQ	0	0	0
Total	164	224	388

From the data in table 8 we obtained six *p*-values (we do not report the results that include the hypervariable region 4 polymorphism with the allelic form 5, because it is not interesting and its sample size is too low):

1) *p*-value (66-..) is 0.049, it is slightly significant at the level $\alpha = 0.05$, so we could say that the Q allele in the R353Q polymorphism present an increasing protective effect (odds ratios are 0 and 0.40) if it is associated to the form 66 of the hypervariable region 4 polymorphism. Anyway, by looking also to the small sample size of 66-RR haplotype, we need further analyses to confirm this result.
2) *p*-value (67-..) is 0.877, we accept the null hypothesis that there is not any variation of effect with the Q allele in the R353Q polymorphism conditionally to the genotypic form 67 of the hypervariable region 4 polymorphism.
3) *p*-value (77-..) is 0.383, we accept the null hypothesis that there is not any variation of effect with the Q allele in the R353Q polymorphism conditionally to the genotypic form 77 of the hypervariable region 4 polymorphism.

4) *p*-value (7.-RR) is 0.043, we have a slightly significant result that there is an increasing protective effect (odds ratios are 0.76 and 0.56) with the 7 allele in the hypervariable region 4 polymorphism conditionally with the genotypic form RR of the R353Q polymorphism.
5) *p*-value (7.-RQ) is 0.841, we accept the null hypothesis that there is not any variation of effect with the 7 allele in the hypervariable region 4 polymorphism conditionally with the genotypic form RQ of the R353Q polymorphism.
6) *p*-value (7.-QQ) is 0.734, we accept the null hypothesis that there is not any variation of effect with the 7 allele in the hypervariable region 4 polymorphism conditionally with the genotypic form QQ of the R353Q polymorphism.

Table 9

Promoter & R353Q	Cases	Controls	Total
11-RR	107	126	233
11-RQ	12	14	26
11-QQ	0	0	0
12-RR	6	7	13
12-RQ	35	62	97
12-QQ	1	6	7
22-RR	0	0	0
22-RQ	1	0	1
22-QQ	0	4	4
Total	162	219	381

The six *p*-values for table 9 are: *p*-value (11-..) = 0.998, *p*-value (12-..) = 0.158, *p*-value (22-..) = 0.000, *p*-value (..-RR) = 0.981, *p*-value (..-RQ) = 0.723, *p*-value (..-QQ) = 0.076. Only the *p*-value that tests for a possible variation of effect with the Q allele in the R353Q polymorphism conditionally with the genotypic form 22 of the promoter region polymorphism, shows a significant evidence for that, but its result is strongly conditioned to the small sample size (see lines 22-RR, 22-RQ and 22-QQ in table 9). From these results we can say that allele 7 into the hypervariable region 4 polymorphism and allele Q into the R353Q polymorphism of the factor VII region have a protective (and codominant) effect against the myocardial infarction. Furthermore, the protective effect of allele 7 may increase if into the haplotype of the factor VII region we have the genotypic form 22 of the promoter region polymorphism. These results agree with those obtained in Iacoviello *et Al.*, 1998. We performed also the permutation test illustrated in Sec. 3.6 by considering the confounding factor represented by the smoke for testing if there is a difference of interaction between the subjects under different levels of the confounding factor: in fact people into the data are subdivided as smokers and not smokers. So if we look at the table 10, we see the case-control data for the R353Q polymorphism jointly to the smoking stratum variable. Therefore, we test if there is an interaction effect between smoke and R353Q polymorphism, such that there is a different level of protection for R353Q polymorphism among the two levels: smoke = No and Smoke =Yes. We obtained a combined *p*-value of 0.000 with 10000 CMC-iterations and by using the Liptak combining function. This significant result mains that the allelic variant Q of the R353Q polymorphism has a protective effect that **increases** in people who are smokers with respect people are not smokers.

Table 10

	Smoke = NO	Smoke = NO	Smoke = YES	Smoke = YES
R353Q	Cases	Controls	Cases	Controls
RR	27	60	87	53
RQ	17	42	32	25
QQ	0	1	1	6
Total	44	103	120	84

At the same way, in table 11 we see the joined data for the hypervariable region 4 polymorphism and the smoking stratum variable. Here we had a combined *p*-value of 0.002, so that we conclude the protective effect of allelic variant 7 of hypervariable region 4 polymorphism has an increasing protective effect when people are smokers. Tests on 5 allelic variant are not reported, because it is not interesting.

Table 11

	Smoke = NO	Smoke = NO	Smoke = YES	Smoke = YES
Hypervariable	Cases	Controls	Cases	Controls
66	24	41	60	39
67	17	50	43	29
77	4	12	8	15
65	0	0	5	1
55	0	0	0	0
75	0	1	4	0
Total	45	104	120	84

Table 12

	Smoke = NO	Smoke = NO	Smoke = YES	Smoke = YES
Promoter	Cases	Controls	Cases	Controls
11	33	59	86	58
12	10	41	32	23
22	0	1	1	2
Total	43	101	119	83

Result on the table 12 for the promoter region are not significant.

We can repeat the previous analyses for tables 10, by looking at the association between the presence or not of the rare allele in case and controls. We performed these analyses with table 13 and we obtained more differences between case and control groups into the smokers than into not smokers.

Table 13

	Smoke = NO	Smoke = NO	Smoke = YES	Smoke = YES
R353Q	Cases	Controls	Cases	Controls
RR	27	60	87	53
RQ & QQ	17	43	33	31
Total	44	103	120	84

The *p*-values for permutation tests relative to the odds ratios of data groups for Smoke = No and Smoke = Yes are, respectively, 0.082 (odds = 0.878) and 0.001 (odds = 0.648), which suggests the protective effect of allele R353Q is active when we consider people that smoke.

These nonparametric results, which used every time complex hypothesis systems and the isotonic alternatives, support and add further confirm to those obtained by Iacoviello, Di Castelnuovo *et Al.* (1999) with different parametric and simple approaches.

6. Conclusions

6.1 Statistical Methodology

In concluding, we would like to emphasize the role of nonparametric combination be a flexible methodology for solving complex problems. These complex testing problems are not adequately taken into consideration in the standard literature. This is in spite of the fact that they are very frequently encountered in a great variety of practical applications. These problems emphasize the versatility and effectiveness of the nonparametric combination methodology.

It should also be stressed that, because permutation tests are conditional with respect to a set of sufficient statistics, the nonparametric combination, in very mild conditions, frees the researcher from the necessity to model the dependence relations among responses. Furthermore, several Monte Carlo experiments have showed that the unconditional power behaviour of combined tests is similar to that of their best parametric counterparts, in the conditions for the latter.

The nonparametric combination of dependent permutation partial tests is a method for the combination of significance levels or rejection probabilities. Conversely, the way generally followed by most parametric tests, based for instance on likelihood ratio behaviour, essentially corresponds to the combination of discrepancy measures usually expressed by distance of points in sample space χ. In this sense, this method appears as a substantial extension of standard parametric approaches. Further, in the presence of a stratification variable, the nonparametric combination, through a multi-phase procedure, allows for flexible solutions. For instance, we may first combine partial tests with respect to variables within each stratum and then combine the *combined* test with respect to strata. Alternatively, we may first combine partial tests related to each variable with respect to strata, and then combine the *combined* tests with respect to the variables.

As a final remark, in very mild conditions, the nonparametric combination method may be considered as a way of reducing the degree of complexity for most testing problems. We saw these last characteristics in the genetic problem discussed, when we perform partial tests on the different odds ratios obtained from the same case-control data, and then calculate the combined test by using the first p-values. For case-control association studies we used these permutation methods but we also introduced other parametric likelihood methods. We did not perform power comparisons between these two types of tests, because we cannot obtain a good approximation of the

References

Boehnke M, Langefeld CD (1998), **Genetic association mapping based on discordant sib pairs: the discordant-alleles test**. *Am. J. Hum. Genet. 62: 950-961.*

Cappuccio FP, Sagnella GA, MacGregor GA (2000), **Association studies of genetic polymorphisms and complex disease** (Correspondence). *The Lancet 355: 1278.*

Cheung BMY, Kumana CR (2000), **Association studies of genetic polymorphisms and complex disease** (Correspondence). *The Lancet 355: 1277.*

Chiano MN, Clayton DG (1998), **Genotypic relative risks under ordered restriction**. *Genet. Epidemiol. 15: 135-146.*

Chowdhury TA (2000), **Association studies of genetic polymorphisms and complex disease** (Correspondence). *The Lancet 355: 1277-1278.*

Cortellaro M *et Al.* (1992), **The PLAT Study: hemostatic function in relation to atherothrombotic ischemic events in vascular disease patients: principal results**. *Arterioscler. Thromb. 12: 1063-1070.*

Devlin B, Roeder K (1999), **Genomic control for association studies**. *Biometrics 55: 997-1004.*

Di Castelnuovo A *et Al.* (2000), **The decanucleotide insertion/deletion polymorphism in the promoter region of the coagulation factor VII gene and the risk of familial myocardial infarction**. *Thrombosis Research 98: 9-17.*

Di Castelnuovo A, Mazzaro D, D, Pesarin F, Salmaso L (1999), **Multidimensional permutation testing for isotonic inference: an application to genetics**. Volume of Abstracts. *International Biometric Society Italian Region, 3^o Congresso nazionale*, Roma (Italy), *July 7-9.*

Di Castelnuovo A, Mazzaro D, Pesarin F, Salmaso L (2000), **Permutation tests for isotonic inference with applications in genetics**. Volume of Abstracts, *Challenges and opportunities in genetic epidemiology,* Trento (Italy), *Sept. 7th*.

Elston RC (1998), **Linkage and association**. *Genet. Epidemiol. 15: 565-576.*

Foppa I, Spiegelman D (1997), **Power and sample size calculations for case-control studies of gene-environmental interactions with a polytomous exposure variable**. *Am. J. Epidemiol. 146: 596-604.*

Gambaro G, Angiani F, D'Angelo A (2000), **Association studies of genetic polymorphisms and complex disease** (Viewpoint). *The Lancet 355: 308-311.*

Haines JL, Pericak-Vance MA (1998), **Approaches to Gene Mapping in Complex Human Diseases**. *Wiley-Liss, Inc. 323-332.*

Heinrich J *et Al.* (1994), **Fibrinogen and factor VII in the prediction of coronary risk: results from the PROCAM study in healthy men.** *Arterioscler. Thromb. 14: 54-59.*

Hosmer DW, Lemeshow S (1992), **Confidence interval estimation of interaction**. *Epidemiology. 3: 452-456.*

Khoury MJ, Beaty TH (1994), **Applications of the case-control method in genetic epidemiology**. *Epidemiol. Rev. 16: 134-150.*

Iacoviello L, Donati MB (1999), **Gene-environment interactions: implications for cardiovascular disease**. *Cardiologia. 44: 227-232.*

Iacoviello L, Di Castelnuovo A *et Al.* (1999), **Cigarette smoking doubles the risk of myocardial infarction of a protective polymorphism in the blood coagulation factor VII gene**, *Thromb. Haemost. 81: 658.*

Iacoviello L *et Al.* (1998), **Polymorphisms in the coagulation factor VII gene and the risk of myocardial infarction.** *N. Engl. J. Med. 338: 79-85.*

Lathrop, GM (1983), **Estimating genotype relative risks**, *Tissue Antigens 22: 160-166.*

Lazzeroni LC, Lange K (1998), **A conditional inference framework for extending the transmission/disequilibrium test**. *Hum. Hered. 48: 67-81.*

Meade TW *et Al.* (1986), **Haemostatic function and ischaemic heart disease: principal results of the Northwick Park Heart Study**. *Lancet 2: 533-537.*

Meade TW *et Al.* (1993), **Fibrinolytic activity, clotting factors, and long-term incidence of ischaemic heart disease in the Northwick Park Heart Study**. *Lancet 342: 1076-1079.*

Pesarin F (2001), **Multivariate Permutation Tests with Applications to Biostatistics**. *JOHN WILEY & SONS*, in press.

Pritchard JK, Rosenberg NA (1999), **Use of unlinked genetic markers to detect population stratification in association studies**. *Am. J. Hum. Genet. 65: 220-228.*

Risch N, Merikangas K (1996), **The future of genetic studies of complex human diseases**. *Science. 273: 1516-1517.*

Sasieni PD (1997), **From genotypes to genes: doubling the sample size**. *Biometrics. 53: 1253-1261.*

Schaid JD, Rowland C (1998), **Use of parents, sibs, and unrelated controls for detective of associations between genetic markers and disease**. *Am. J. Hum. Genet. 61: 1492-1506.*

Sharma AM (2000), **Association studies of genetic polymorphisms and complex disease** (Correspondence). *The Lancet 355: 1277.*

Smith PG, Day NE (1984), **The design of case-control studies: the influence of confounding and interaction effects**. *Int. J. Epidemiol. 13: 356-365.*

Spielman RS, Ewens WJ (1998), **A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test**. *Am. J. Hum. Genet. 62: 450-458.*

Spielman RS, McGinnis RE, Ewens WJ (1993), **Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM)**. *Am. J. Hum. Genet. 52: 506-513.*

Terwilliger JD, Göring HHH (2000), **Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design**. *Hum. Biol. 72: 63-132.*

Vaziri ND *et Al.* (1992), **Coagulation, fibrinolytic, and inhibitory proteins in acute myocardial infarction and angina pectoris**. *Am. J. Med. 93: 651-657.*

Wacholder S, Silverman DT, McLaughlin JK, Mandel JS (1992), **Selection of controls in case-control studies. Types of controls groups**. *Am. J. Epidemiol. 135: 1029-1041.*