Am. J. Hum. Genet. 61:223–227, 1997

Variable Age at Onset in Insulin-Dependent Diabetes Mellitus, by the Marker-Association-Segregation-χ^2 Method

Noël Bonneuil,[1] Antoine Clerget,[2] and Françoise Clerget-Darpoux[2]

[1]Institut National des Études Démographiques, and [2]INSERM U155, Paris

Summary

The marker-association-segregation-χ^2 (MASC) method with consideration of age, for nonaffected persons, and of age at onset, for affected persons, was applied to a sample of 308 HLA-typed families. Hazard rates modeling the instantaneous risk of catching the disease were estimated under the exponential distribution and with satisfactory goodness of fit. This class of models shows that the hypothesis of the absence of parental imprinting cannot be rejected for insulin-dependent diabetes mellitus.

Introduction

Cudworth and Woodrow (1975) showed that a genetic factor located in the HLA region is involved in the etiology of type 1 insulin-dependent diabetes mellitus (IDDM). However, Hodge et al. (1980), Risch (1984), Clerget-Darpoux et al. (1986), and Louis and Thomson (1986) agreed that a single allele of susceptibility was not enough to fit the observations.

Notably, Deschamps et al. (1990) found that 62% of their sample of 130 unrelated DR3DR4 patients without affected parents inherited the DR3 allele from their mothers. Therefore, Clerget-Darpoux et al. (1991) fit a model with maternal effect associated with the DR3 antigen and a "complementation" effect, defined as the presence of two alleles of susceptibility that are located at two loci of the HLA region.

Margaritte-Jeannin et al. (1995) concluded in favor of parental imprinting on a specific allele combination in the HLA region. They showed that, if maternal effect (Clerget-Darpoux et al. 1991) could not be retained, then parental imprinting, which reflects a different role of the same allele when transmitted by the father or by the mother, could be a good candidate for explaining the observed distributions of DR3 and DR4 among affected individuals. Undlien et al. (1995), using an independent set of data, found no such difference in the parental origin of susceptibility. Here we reconsider the data used by Margaritte-Jeannin et al. (1995) and introduce risks of expressing the disease as functions not only of genotypes but also of age.

Received September 13, 1996; accepted for publication April 11, 1997.

Address for correspondence and reprints: Dr. Noël Bonneuil, Institut National des Études Démographiques, 27, rue du Commandeur, 75675, Paris cedex 14, France. E:mail: bonneuil@cilaos.ined.fr

0002-9297/97/6101-0029$02.00

Material and Methods

Margaritte-Jeanin et al. considered a sample comprising (*a*) 390 French families followed by I. Deschamps and HLA typed in J. Hors's laboratory and (*b*) 94 Caucasian families from the Genetic Analysis Workshop 5 data (Clerget-Darpoux and Babron 1989; Spielman et al. 1989). From these 416 families, we retained only those for which age, for nonaffected individuals, or age at onset, for affected individuals, was known for all members of the families. The final reduced sample comprised 308 HLA-DR typed families with information on age. These families were classified into two familial configurations: 207 families in which no parent and no sib was affected (configuration *C*1) and 101 families in which no parent and at least one sib of the index was affected (configuration *C*2). All individuals were typed at HLA loci *A, B, C,* and *DR,* enabling determination of identity by descent (IBD) (0, 1, or 2).

The marker-association-segregation-χ^2 (MASC) method was introduced by Clerget-Darpoux et al. (1988, p. 248) "to take into account the simultaneous information of segregation and association of a marker and a disease," as well as the risk that some relatives of the patients would be affected. Patients are classified according to their familial configuration (one or no siblings affected), marker genotype, and degree of IBD with a randomly chosen sib. Then, a model of segregation is fit either by likelihood maximization, in which age is considered, or, alternatively, by χ^2 minimization, in which age is not considered. It is noteworthy that this model is based on the probability f_{ij} of being affected when having the genotype S_iS_j. This probability, called "penetrance," is usually assumed to be constant with age, which may not be a reasonable assumption for many diseases that are known to appear with age. We extended this method of MASC by introducing varying age at onset of the disease—that is, by making f_{ij} depend on age.

The MASC method basically compares observed and theoretical probabilities that two sibs, one of whom is the affected proband, are IBD for two haplotypes (IBD = 2), one haplotype only (IBD = 1), or no haplotype at all (IBD = 0). These probabilities are conditioned, first, by the familial configuration with respect to affected statuses and, second, on the information provided by the marker locus. In other words, we calculate X_{kl}^{Ci} = P(IBD = 2/Ci + ind = M_kM_l aff); Y_{kl}^{Ci} = P(IBD = 1/Ci + ind = M_kM_l aff); Z_{kl}^{Ci} = P(IBD = 0/Ci + ind = M_kM_l aff), where Ci denotes the familial configuration, M_k is the kth allele of the marker locus M, and "aff" and "ind" are abbreviations for "affected" and "index," respectively.

The candidate gene, whose recombination fraction with regard to the Marker locus is assumed to be negligible, is denoted S and its ith allelic form is denoted S_i. The penetrance of a pair S_iS_j, denoted as f_{ij}, is the probability that a zygote contracts the disease when he or she has S_iS_j in his or her genotype: f_{ij} = P(aff/S_iS_j). The probabilities U_{kl}^{Ci}, $U = X,Y,Z$, respectively, are then written as explicit functions of these penetrances f_{ij}, of the frequencies of the markers, of the probabilities for the parents to have given markers, of the probability of having n sibs, and of the familial configuration. The comparison with observed values of the U_{kl}^{Ci}, $U = X,Y,Z$, respectively, is obtained either through minimizing a χ^2 statistic with respect to the f_{ij}'s or maximizing an appropriate likelihood, which is the product of the $(U_{kl}^{Ci})^{n_{kl}^{Ci}}$, $U = X,Y,Z$, where n_{kl}^{Ci} is the number of individuals of marker kl and familial configuration Ci.

For example,

$$\frac{\sum_{i,j} \alpha_{ijkl} f_{ij} \sum_{s,t} \beta_{ijst} \sum_{n} (1-\phi_{ijst})^n \omega_{k\ln C1}(1-f_{ij})/[4(1-\phi_{ijst})]}{\sum_{i,j} \alpha_{ijkl} f_{ij} \sum_{s,t} \beta_{ijst} \sum_{n} (1-\phi_{ijst})^n \omega_{k\ln C1}}, \quad (1)$$

where ϕ_{ijst} = $(f_{ij}+f_{it}+f_{js}+f_{st})/4$, β_{ijst} = $q_sq_t(1-f_{si})(1-f_{tj})$ (q_s is the probability that one parent of the individual typed as S_iS_j has the haplotype S_s), α_{ijkl} = $P(S_iS_jM_kM_l)$, ω_{klnC1} is the probability of n sibs when the index is M_kM_l, the configuration is C1, and there exists at least one sib. For further detailed formulas, see the work of Clerget-Darpoux et al. (1988).

We extended this method to the case of age-dependent onset of disease. In this case, as commonly practiced in survival data analysis, each individual is considered as being at risk of contracting the disease. If the individual is already affected, then his or her age at onset is recorded as the defining endpoint to his or her period of risk. If the individual is not affected at the moment of the survey, we consider that this person is still at risk and will certainly become affected in the future, unless he or she dies from another cause before onset of the disease. At the time of survey, we record the individual's age as a minimal (censored) endpoint to his or her period of risk.

Following Cox and Oakes (1984), Elston (1973), Bonney (1986), and Abel and Bonney (1990), we take each penetrance f_{ij} relative to the genotype S_iS_j, as the instantaneous risk $h_{ij}(a)$ of being affected at age a, as $h_{ij}(a) = \lim_{\Delta\to 0+}[P(a \leqslant T < a + \Delta/a \leqslant T)]/\Delta$, where T is the random variable "age at onset." The probability of surviving in the healthy state until age a is $F_{ij}(a)$ = $P(T > a)$; hence, $[dF_{ij}(a)/da] = -h_{ij}(a)F_{ij}(a)$. After integration, $F_{ij}(a) = \exp[-\int_0^a h_{ij}(b)db]$. The notion of instantaneous risk supersedes the notion of penetrance. Each individual is now viewed as being susceptible to the disease, whatever his or her affected status. In the case of an unaffected person, at the moment of the survey, we consider this observation as being censored. An affected person will contribute to the likelihood through a term $F_{ij}(a)h_{ij}(a)$, where a is the age at onset in that person, whereas an unaffected person will contribute to the likelihood through a term $F_{ij}(a)$, where a is the age of the person at the moment of the investigation. The main difficulty arising in the introduction of survival data analysis into the MASC method is in accounting for the contributions brought by the various members of a given family. A simple parameterization is the exponential family, because penetrances f_{ij} are replaced one-to-one by constant instantaneous risks $h_{ij}(a) = h_{ij}$ for all values of a. Subsequently, in the exponential specification, the survival function is $F_{ij}(a) = \exp(-h_{ij}a)$. The ages or ages at onset of parents a_f and a_m, for father and mother, respectively, are introduced, as well as the ages or the ages of onset in the n sibs, a_{sk}, $k = 1, \ldots, n$. The probabilities U_{kl}^{Ci}, $U = X,Y,Z$, respectively, are rewritten to account for the parameterization of penetrances by use of instantaneous risks; for example, the β_{ijst} term in equation (1) is rewritten as $\beta_{ijst}(a_f,a_m)$ = $\{q_sq_t[F_{tj}(a_f)F_{is}(a_m) + F_{tj}(a_m)F_{is}(a_f)]/2\}$ in the case of one affected parent; the $(1-\phi_{ijst})^n$ are replaced by $\Pi_{k=1}^n\{1 - [h_{ij}F_{ij}(a_{sk}) + h_{it}F_{it}(a_{sk}) + h_{js}F_{js}(a_{sk}) + h_{st}F_{st}(a_{sk})]\}/4$, where k denotes the kth sib. Similarly, for each family (Ci, ind aff, a,a_f,a_m,a_s), the equivalent forms of the U_{kl}^{Ci}, $U_{kl}^{Ci}(a,a_f,a_m,a_{s1}, \ldots, a_{sn})$, $U = X,Y,Z$ are computed. Finally, the expected number of individuals in each class (familial configuration × marker genotype × degree of IBD with a randomly chosen sib) can be derived, and a χ^2 test of goodness of fit is conducted.

We simulated several (three) sets of 100 samples of 308 families of C1 and C2, in which the risk of contracting the disease, $f_{ij}(a)$, has an exponential distribution of intensity h_{ij}, $i = 1,2$, $j = 1,2$. (The study to do would be to make the three parameters h_{11}, h_{12}, and h_{22} describe the whole space of possible values. For each point of this three-dimensional space, we would have to

simulate 100 samples to run the models. Because of the computation time required, however, we were content with trying three different sets of parameters, choosing different situations—[1] $h_{11} = .01$, $h_{12} = h_{21} = .01$, and $h_{22} = .001$; [2] $h_{11} = .05$, $h_{12} = h_{21} = .02$, and $h_{22} = .01$; and [3] $h_{11} = .006$, $h_{12} = h_{21} = .005$, and $h_{22} = .00$—and the coupling frequencies $c_{11} = .25$ between S_1 and M_1 and $c_{22} = .75$ between S_2 and M_1, as well as the frequency of the genes, .2 for S_1 [and .8 for S_2].) Although the models with age-dependent penetrances and with penetrances constant with age are not statistically comparable, because they are not nested models, we also ran, on these same 100 samples, the model with penetrances constant with age. We saw that the model with penetrances constant with age can produce goodness of fit comparable to that of the model with age-dependent penetrances. Goodness of fit, of course, is not a sufficient criterion by which to assess the appropriateness of a model, and, for diseases with a delayed age at onset, the model with age-dependent penetrances is preferred from a biological standpoint. We also simulated three series of 100 samples with penetrances constant with age, on which we ran models based on age-dependent penetrances, as well as models with penetrances constant with age. Similarly, the criterion of goodness of fit does not permit us to distinguish a "better" model, since the models with age-dependent penetrances and the models with penetrances constant with age do not give significantly different goodness of fit.

For the IDDM data, the distribution of HLA DR alleles was considered to be DR3 (12%) and DR4 (13%) (Baur et al. 1984). Alleles different from DR3 and DR4 are denoted "DRX." In the model with complementation effect (Clerget-Darpoux et al. 1991), the susceptibility to the disease comes from two specific alleles, denoted "α_0" and "β_0," located at two closely linked loci *A* and *B* in the HLA region. The recombination fractions between *A*, *B*, and the HLA markers are assumed to be negligible. Only individuals having at least one α_0 allele or at least one β_0 allele can develop the disease. We denote as "α" and "β" all other alleles different from α_0 and β_0 at loci *A* and *B*, respectively. This model with complementation effect is equivalent to a one-locus model with four alleles, where $S_1 = \alpha_0\beta_0$, $S_2 = \alpha_0\beta$, $S_3 = \alpha\beta_0$, and $S_4 = \alpha\beta$. The model involves the instantaneous risks h_{ij}, $i, j = 1, \ldots, 4$, and the 12 coupling probabilities c_{ij} between the disease alleles S_i, $i = 1, \ldots, 4$ and the marker alleles DR3, DR4, and DRX. Note that, for all j, $\Sigma_i c_{ij} = 1$ and that the frequency of the ith disease allele is $\Sigma_{j\in\{3,4,X\}} c_{ij} P(\mathrm{DR}j)$.

Since affected individuals have either α_0 or β_0, the instantaneous risks of genotypes lacking in α_0 and β_0 are constrained to 0. As already suggested by Clerget-Darpoux et al. (1991), taking advantage of the study of Khalil et al. (1990), we assume that susceptibility is due to a specific heterodimer on the cell surface. Subsequently, the coupling probabilities of DR3 with β and those of DR4 with α are set to 0, as are those of DRX with $\alpha_0\beta_0$.

Table 1

Model 1 of Hazard-Rates Matrix, with Imprinting

	Maternal Allele			
Paternal Allele	$\alpha_0\beta_0$	$\alpha_0\beta$	$\alpha\beta_0$	$\alpha\beta$
$\alpha_0\beta_0$	*E*	*F*	*G*	*H*
$\alpha_0\beta$	*I*	0	*J*	0
$\alpha\beta_0$	*K*	*L*	0	0
$\alpha\beta$	*M*	0	0	0

Table 2

Model 2 of Hazard-Rates Matrix, with Imprinting

	Maternal Allele			
Paternal Allele	$\alpha_0\beta_0$	$\alpha_0\beta$	$\alpha\beta_0$	$\alpha\beta$
$\alpha_0\beta_0$	*H*	*L*	*H*	*M*
$\alpha_0\beta$	*H*	0	*H*	0
$\alpha\beta_0$	*L*	*L*	0	0
$\alpha\beta$	*M*	0	0	0

Results

Under the hypothesis of parental imprinting, maternal and paternal effects do not have the same effect. The probability of being affected, given an individual who has inherited a disease allele from his or her mother, will not be equal to the probability in the presence of paternal inheritance. Thus, the penetrances depend not only on the genotype but also on the parental inheritance of each allele (Margaritte-Jeannin et al. 1995). This specification leads us to estimate the hazard-rates matrix presented in table 1, where uppercase letters represent hazard rates. To have a clearer view of the hazard-rates matrices, we use the notation $h_{11} = H$, $h_{12} = L$, and $h_{14} = M$. Since Margaritte-Jeannin et al. (1995) claimed the existence of a maternal effect, we can restrict the hazard-rates matrix in table 2 by imposing $h_{14} = h_{41} = M$, implying the absence of paternal effect.

The difference between models 1 and 2 lies only in the complexity—nine parameters for the former versus three parameters for the latter. The MASC model with no parental imprinting is specified by requiring that hazard rates be symmetrical; that is, $h_{ij} = h_{ji}$, which means $I = F$, $K = G$, $M = H$ and $J = L$ in table 1 and $H = L$

Table 3

Statistics of Models with Age-Dependent Penetrances and with Penetrances Constant with Age, with and without Imprinting

	Model 1			Model 2				
	−2 Log Likelihood	χ^2	df	−2 Log Likelihood	χ^2	df	Likelihood-Ratio Test	df
With age-dependent penetrances:								
With imprinting	2,206.2	24.42	16	2,218.2	29.98	22	12.0	6
Without imprinting	2,205.7	30.69	20	2,222.0	35.09	23	16.3	3
Likelihood-ratio test	−0.5[a]		4	3.8		1		
With penetrances constant with age:								
With imprinting	675.9	23.91	16	676.3	36.13	22	.43	6
Without imprinting	677.0	30.95	20	677.3	42.39	23	.32	3
Likelihood-ratio test	1.2		4	1.0		1		

[a] Negative value is due to imperfect numerical convergence.

in table 2. With parental imprinting, models 1 and 2 have, respectively, five and two parameters.

Table 3 presents the maximum-likelihood estimates and χ^2 associated with these models, with age-dependent penetrances and with penetrances constant with age, for our sample of 308 families. On the basis of the results in table 3, we can test two hypotheses: (1) model 1 against model 2 and (2) imprinting against absence of imprinting, using likelihood-ratio tests. With penetrances constant with age, models 1 and 2 are not significantly different. With age-dependent penetrances and no parental imprinting, model 1 must be preferred. With age-dependent penetrances and parental imprinting, the likelihood-ratio test between model 1 and model 2 equals 12.0, whereas the 5% level of χ^2_6 is 12.592: model 2 can be said to be not significantly different from model 1. Moreover, model 2 is more parsimonious.

Similarly, the likelihood-ratio test can be used to test the presence of parental imprinting. Table 3 shows that there is no significant difference, regardless of whether parental imprinting is considered in model 1 and in model 2, whether penetrances are constant or varying with age. The significance of age dependency cannot be tested formally, since models with age-dependent penetrances and with penetrances constant with age are not nested within one another.

Among model 2 with and without imprinting and model 1 with imprinting, model 2 without imprinting is the most parsimonious. Model 1 without imprinting is significantly different and must be preferred.

Table 4 presents the estimated hazard rates for model 1 with age-dependent penetrances and without imprinting, with its coupling matrix presented in table 5. Unfortunately, no inference of these coefficients can be computed with MASC yet (we will attempt this task in a future paper).

Discussion

The consideration of age at onset in genetic diseases appearing along the life cycle should better reflect the process of disease expression than does consideration of penetrances constant with age. Focusing on the instantaneous risk makes the probability of expressing the disease depend on age.

We have presented here an extension, with age, of the MASC method first introduced by Clerget-Darpoux et al. (1988), which we have used to reexamine IDDM

Table 4

Model 1 without Parental Imprinting: Best-Fit Values for Hazard-Rates Matrix

	Maternal Allele			
Paternal Allele	$\alpha_0\beta_0$	$\alpha_0\beta$	$\alpha\beta_0$	$\alpha\beta$
$\alpha_0\beta_0$	.017	.011	.014	.012
$\alpha_0\beta$	.011	0	.047	0
$\alpha\beta_0$	.014	.047	0	0
$\alpha\beta$	.012	0	0	0

Table 5

Model 1 with Age-Dependent Penetrances and without Parental Imprinting: Best-Fit Values for Matrix of Coupling Frequencies c_{ij}

	Marker Allele		
Haplotype	DR3	DR4	DRX
$\alpha_0\beta_0$	.25	.34	.02
$\alpha_0\beta$	0	.67	0
$\alpha\beta_0$	.75	0	0
$\alpha\beta$	0	0	.98

data. For the selected 308 families in which age or age at onset are known, the age dependent–penetrance model that we selected gives no significant role to parental imprinting in IDDM. Although no inference of the hazard rates has yet been made, the data in table 4 imply that the risk is highest (h_{ij} = .047) for individuals with genotype $\alpha_0\beta$ from one parent and with $\alpha\beta_0$ from the other parent; this statement remains to be tested, however. With the same reserve with regard to inference, the risk would be quite the same for other combinations involving at least one α_0 and one β_0 (.011, .012, and .017). To illustrate the consequence of these numbers, let us mention that a 20-year-old individual at risk .047 has a probability of $1 - \exp(-.04720) = .61$ of having the disease, whereas the same probability for a 20-year-old individual at risk .011 is only .20. These probabilities become .85 and .36, respectively, for 40-year-old individuals. To determine the difference between individuals of genotype $\alpha_0\beta$ $\beta_0\alpha$, with each disease allele coming from a different parent, and individuals of other genotypes, the expectancy that an individual will live without the disease is a good indicator: it is, respectively, 21.3 and 90.9 years at birth and is 8.3 and 73.0 years at age 20 years. In a future paper, we hope to be able to test the significance of this important difference.

Acknowledgments

We are grateful to Patricia Margaritte-Jeanin for helpful discussions.

References

Abel L, Bonney GE (1990) A time-dependent logistic hazard function for modelling variable age of onset in analysis of familial diseases. Genet Epidemiol 7:391–407

Baur MP, Neugebauer M, Deppe H, Sigmund M, Luton T, Mayr WR, Alber ED (1984) Population analysis on the basis of deduced haplotypes from random families. In: Albert ED, Bauer MP, Mayr WR (eds) Histocompatibility testing 1984. Springer, New York, Tokyo, pp 677–755

Bonney GE (1986) Regressive logistic models for familial disease and other binary traits. Biometrics 42:611–625

Clerget-Darpoux F, Babron MC (1989) Testing genetic models for IDDM by the MASC method. Genet Epidemiol 6:59–64

Clerget-Darpoux F, Babron MC, Deschamps I, Hors J (1991) Complementation and maternal effect in insulin-dependent diabetes. Ann Hum Genet 49:42–48

Clerget-Darpoux F, Babron MC, Prum B, Lathrop GM, Deschamps I, Hors J (1988) A new method to test genetic models in HLA associated diseases: the MASC method. Ann Hum Genet 42:247–258

Clerget-Darpoux F, Dizier MH, Bonaïté-Pellié C, Babron MC, Hochez J, Martinez M (1986) Discrimination between genetic models for insulin dependent diabetes mellitus. Genet Epidemiol 3 Suppl 1:313–318

Cox DR, Oakes D (1984) Analysis of survival data. Chapman & Hall, London and New York

Cudworth AG, Woodrow JC (1975) Evidence for HLA linked in juvenile diabetes mellitus. BMJ 3:133–135

Deschamps I, Hors J, Clerget-Darpoux F, Gardais E, Robert J, Marcelli-Barge A, Lestradet H, et al (1990) Excess of maternal HLA-DR3 antigens in HLA DR3, 4 positive type I (insulin-dependent) diabetic patients. Diabetologia 33:425–430

Elston RC (1973) Ascertainment and age of onset in pedigree analysis. Hum Hered 23:105–112

Hodge SE, Rotter JI, Lange K (1980) A three-allele model for heterogeneity of juvenile-onset insulin-dependent diabetes. Ann Hum Genet 43:349–409

Khalil I, d'Auriol L, Gobet M, Morin L, Lepage V, Deschamps I, Sik Park M, et al (1990) A combination of HLA-DQβ Asp57 negative and HLA Arg52 confers susceptibility to IDDM. J Clin Invest 85:1315–1319

Louis EJ, Thomson G (1986) Three allele synergistic mixed model for insulin-dependent diabetes mellitus. Diabetes 35: 958–963

Margaritte-Jeannin P, Clerget-Darpoux F, Hors J, Deschamps I (1995) Testing parental imprinting in insulin-dependent diabetes mellitus by the marker-association-segregation-χ^2 method. Am J Hum Genet 56:1080–1087

Risch N (1984) Segregation analysis incorporating linkage markers. I. Single-locus models with an application to type I diabetes. Am J Hum Genet 36:363–386

Spielman RS, Baur MP, Clerget-Darpoux F (1989) Genetic analysis of IDDM: summary of GAW5 IDDM results. Genet Epidemiol 6:43–58

Undlien DE, Akselsen HE, Joner G, Dahl-Jørgensen, Aagenæs Q, Søvik O, Thorsby E, et al (1995) No difference in the parental origin of susceptibility HLA class II haplotypes among Norwegian patients with insulin-dependent diabetes mellitus. Am J Hum Genet 57:1511–1514