Am. J. Hum. Genet. 64:886–893, 1999

A Parametric Copula Model for Analysis of Familial Binary Data

David-Alexandre Trégouët,[1] Pierre Ducimetière,[1] Valéry Bocquet,[1] Sophie Visvikis,[3] Florent Soubrier,[2] and Laurence Tiret[1]

[1]Institut National de la Santé et de la Recherche Médicale (INSERM) Unité 258, Hôpital Broussais, and [2]INSERM Unité 358, Hôpital Saint-Louis, Paris; and [2]Centre de Médecine Préventive, Vandoeuvre-lès-Nancy, France

Summary

Modeling the joint distribution of a binary trait (disease) within families is a tedious challenge, owing to the lack of a general statistical model with desirable properties such as the multivariate Gaussian model for a quantitative trait. Models have been proposed that either assume the existence of an underlying liability variable, the reality of which cannot be checked, or provide estimates of aggregation parameters that are dependent on the ordering of family members and on family size. We describe how a class of copula models for the analysis of exchangeable categorical data can be incorporated into a familial framework. In this class of models, the joint distribution of binary outcomes is characterized by a function of the given marginals. This function, referred to as a "copula," depends on an aggregation parameter that is weakly dependent on the marginal distributions. We propose to decompose a nuclear family into two sets of equicorrelated data (parents and offspring), each of which is characterized by an aggregation parameter (α_{FM} and α_{SS}, respectively). The marginal probabilities are modeled through a logistic representation. The advantage of this model is that it provides estimates of the aggregation parameters that are independent of family size and does not require any arbitrary ordering of sibs. It can be incorporated easily into segregation or combined segregation-linkage analysis and does not require extensive computer time. As an illustration, we applied this model to a combined segregation-linkage analysis of levels of plasma angiotensin I–converting enzyme (ACE) dichotomized into two classes according to the median. The conclusions of this analysis were very similar to those we had reported in an earlier familial analysis of quantitative ACE levels.

Received July 1, 1998; accepted for publication January 5, 1999; electronically published February 15, 1999.

Address for correspondence and reprints: Dr. Laurence Tiret, INSERM Unité de Recherche en Génétique Epidémiologique et Moléculaire des Pathologies Cardio-Vasculaires, 17 rue du Fer à Moulin, 75005 Paris, France. E-mail: tiret@idf.inserm.fr

0002-9297/99/6403-0026$02.00

Introduction

A large amount of medical research is directed toward characterization of genes involved in the predisposition to human diseases. Although considerable progress has been made for monogenic diseases, the identification of susceptibility genes for multifactorial diseases still poses numerous challenges, including the development of new statistical methodologies. Nonparametric methods have become increasingly popular for investigation of the genetic basis of multifactorial diseases, because they do not require the genetic model to be specified. However, the minimal assumptions made by these methods are at the expense of a limited power. In the case of a trait that is influenced by a so-called major gene—that is, a gene with an effect that is large enough to be distinguished from other sources of variability—parametric methods that specify a genetic model are much more powerful for the detection of genetic effects. These methods include segregation analysis, LOD-score analysis, and combined segregation-linkage analysis.

In the case of a quantitative phenotype, parametric models, in general, are based on the multivariate normal distribution, for the characterization of the joint distribution of the trait within a family. In the case of a binary phenotype, such as disease status, there does not exist a joint distribution with desirable properties similar to those of the multinormal distribution. Different formulations have been proposed. One assumes the existence of an underlying liability variable that is distributed normally, with individuals being affected if their liability exceeds a threshold, which may depend on covariates such as age and sex (Morton and MacLean 1974; Lalouel et al. 1983; Demenais 1991). However, this liability variable is a theoretical concept, the reality of which cannot be checked. Regressive logistic models have been proposed that model the joint distribution of familial binary traits, by conditioning each individual's phenotype on those of preceding relatives (Bonney 1986, 1987). These models require an arbitrary ordering of family members and, therefore, yield results that are dependent on family size and order. To overcome the problem of ordering the phenotypes, compound regressive models have been proposed (Bonney 1992), but their

practical use has not been demonstrated clearly. Conditional logistic regression models have been developed that allow specification of the joint distribution of correlated outcomes, without the need to order them (Connolly and Liang 1988; Tosteson et al. 1991; Abel et al. 1993). In these models, familial aggregation between two relatives is expressed in terms of an odds ratio (OR) conditional on the other family members. Therefore, the familial aggregation parameters are dependent on the marginal distributions of the data and on the size of the families. Moreover, the numerous computer calculations required by these models have limited their use.

In this article, we present an alternative model for the analysis of familial binary data that is based on the copula theory. This model, first proposed by Meester and MacKay (1994) for the analysis of symmetric correlated categorical data, was applied here to nuclear families. In this model, the joint distribution of binary outcomes is characterized by a function of the given marginals. This function, referred to as a "copula," depends on a parameter that can be interpreted as a term of association between outcomes that is weakly dependent on the marginal parameters. The advantage of this model is that it does not require any ordering of sibs and provides estimates of association parameters that are independent of the size of the families. The model is described in Methods and is illustrated by a combined segregation-linkage analysis of high plasma levels of angiotensin I–converting enzyme (ACE).

Methods

Copula Theory

Let $y = y_1, ..., y_n$ be a cluster of n random variables with given marginal distribution functions $F_1, ..., F_n$. Suppose we are interested in modeling the joint-distribution function of y. If the true distribution is unknown, it is nevertheless possible to construct a joint distribution F for y that preserves the marginal distribution functions $F_1, ..., F_n$, by use of the following result.

If C is a distribution function on $[0, 1]^n$, with uniform univariate marginals, then $C[F_1(y_1), ..., F_n(y_n)]$ defines a joint-distribution function F for y that has the desired univariate marginals (Schweizer and Sklar 1983). The function C is called a copula. Note that, if the true joint distribution is known, there is (in general) a unique copula function that links it to its marginals. When the true joint distribution is unknown, the choice of a copula is not unique, and, consequently, the derived distribution F is not necessarily the true joint distribution of y.

Several examples of copula functions have been given by Genest and MacKay (1986) for the bivariate case ($n = 2$). Of particular interest is a family of copula functions known as "Frank's family," which was first introduced by Frank (1979) for $n = 2$ and was extended by S. G. Meester (personal communication) for $n > 2$. It is characterized by the following formulation of $F(y)$:

$$F(y) = C_\alpha[F_1(y_1), ..., F_n(y_n)]$$
$$= \tfrac{-1}{\alpha} \log\left\{1 + (e^{-\alpha} - 1)\prod_{i=1}^{n}\left[\frac{e^{-\alpha F_i(y_i)} - 1}{e^{-\alpha} - 1}\right]\right\}, \quad (1)$$

with

$$\lim_{\alpha \to 0}\{C_\alpha[F_1(y_1), ..., F_n(y_n)]\} = \prod_{i=1}^{n} F_i(y_i) .$$

From equation (1), this joint distribution for y appears to be modeled through the given marginals F_i ("mean" structure) and the dependence structure ("covariance" structure) characterized by the copula. This copula is a function of one parameter, α ($-\infty < \alpha < \infty$), which has the properties of a within-cluster association parameter (Genest 1987; Meester and MacKay 1994). Independence between the y_i's occurs if and only if $\alpha = 0$, and positive and negative within-cluster associations are characterized by $\alpha > 0$ and $\alpha < 0$, respectively. In the context of family data, negative associations are rarely encountered, although this possibility cannot be ruled out.

Note that, owing to the expression of the C_α function, which depends on a single association parameter α, the copula model described above only applies to equicorrelated (or symmetric) data. Last, since the joint distribution shown in equation (1) has the same form whatever the cluster size n, this copula model can easily deal with varying cluster sizes.

An important feature of this model is that, for continuous variables, α is an association parameter that has been shown to be independent of the marginal distributions and that is closely related to the Spearman correlation coefficient ρ_s (Schweizer and Wolff 1981). For binary variables, such an independence between α and the marginals is, in general, not true. However, it can be shown that the relationship between α and the pairwise OR between any two binary variables y_i and y_j, with marginal probabilities p_i and p_j, respectively, is as follows:

$$\mathrm{OR} = \frac{P(y_i = 1, y_j = 1)P(y_i = 0, y_j = 0)}{P(y_i = 1, y_j = 0)P(y_i = 0, y_j = 1)}$$
$$= \frac{(p_i + p_j - 1 - \Delta)(-\Delta)}{(1 - p_j + \Delta)(1 - p_i + \Delta)} , \quad (2)$$

with

$$\Delta = \frac{1}{\alpha}\log\left\{1 + \frac{[e^{-\alpha(1-p_i)} - 1][e^{-\alpha(1-p_j)} - 1]}{e^{-\alpha} - 1}\right\} .$$

By use of equation (2), the correspondence between the OR and α can be derived for different values of (p_i, p_j) (fig. 1). For simplicity, we assumed that $p_i = p_j = p$. When $\alpha < 2.5$, the correspondence between the OR and α appears to be weakly dependent on the common marginal probability p. For example, values for α of 0.5, 1, 1.5, and 2 would correspond approximately to pairwise ORs 1.3, 1.6, 2.0, and 2.5, respectively. For higher values of α, the relationship between the OR and α is no longer independent of p. However, the range of variation of the OR, for a given α, remains relatively small (3.4–4.2 for $\alpha = 3$). A similar pattern was observed when $p_i \neq p_j$, with a slightly more pronounced dependency on the marginal probabilities (data not shown). These results indicate that a weak dependency between α and the marginal binary distributions may be expected, as was observed for real data by Meester and MacKay (1994).

Application to Familial Aggregation Analysis of a Binary Trait

We now consider nuclear families in which a binary trait (e.g., disease status) is measured. Let $\boldsymbol{y} = (y_F, y_M, y_1, ..., y_n)$ be the vector of the trait for the father (F), the mother (M), and the n children. Similarly, $\boldsymbol{x} = (x_F, x_M, x_1, ..., x_n)$ is the familial vector of measured covariates.

Under the assumption that, conditional on an individual's own covariates, an individual's status is independent of the covariates of the other family members, the joint probability of the trait, given the covariates, can be decomposed into two probabilities:

$$P(\boldsymbol{y}/\boldsymbol{x}) = P(y_F, y_M/x_F, x_M) \times P(y_1, ..., y_n/y_F, y_M, x_1, ..., x_n) .$$

These two probabilities then can be modeled by two different Frank's family copulas, since (y_F, y_M) and $(y_1, ..., y_n)$ can each be viewed as a set of equicorrelated data. Following S. G. Meester (personal communication), the joint distribution for parents can be written as

$$\begin{aligned} P(y_F, y_M/x_F, x_M) &= C_{\alpha_{FM}}[F_F(y_F/x_F), F_M(y_M/x_M)] \\ &\quad - C_{\alpha_{FM}}[F_F(y_F/x_F), F_M(y_M - 1/x_M)] \\ &\quad - C_{\alpha_{FM}}[F_F(y_F - 1/x_F), F_M(y_M/x_M)] \\ &\quad + C_{\alpha_{FM}}[F_F(y_F - 1/x_F), F_M(y_M - 1/x_M)] \\ &= \sum_{f=0,1}\sum_{m=0,1}\{(-1)^{(f+m)}C_{\alpha_{FM}}[F_F(y_F - f/x_F), \\ &\quad F_M(y_M - m/x_M)]\} , \end{aligned} \tag{3}$$

where α_{FM} is the parameter of association between the father and mother and where F_F and F_M are the marginal distribution functions characterizing each parent's status. These distribution functions can be modeled by use of a logistic representation; for example,

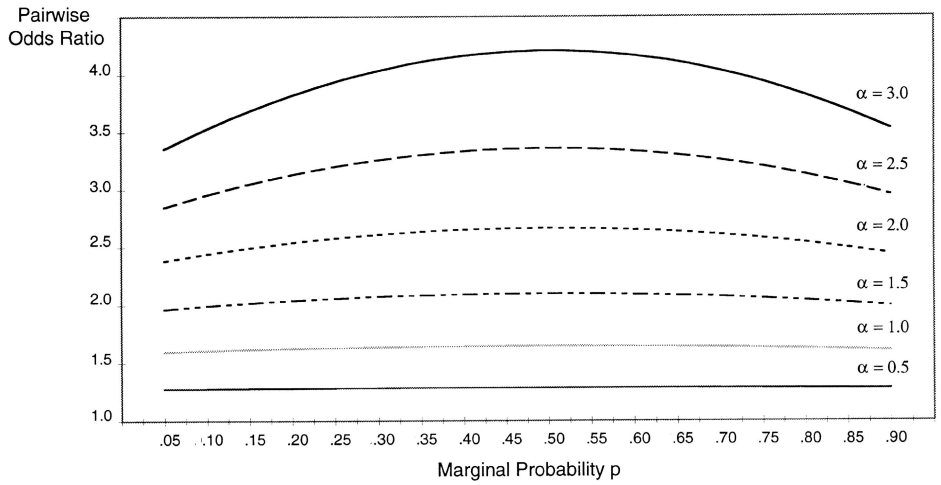

Figure 1 Correspondence between the pairwise OR and copula association parameter α, for two binary variables with common marginal probabilities $p_i = p_j = p$.

$$F_F(U/x_F) = \begin{cases} 0 & \text{if } U = -1 \\ 1/(1 + e^{\lambda+\beta x_F}) & \text{if } U = 0 \\ 1 & \text{if } U = 1 \end{cases},$$

where λ is the baseline parameter and $\boldsymbol{\beta}$ is the vector of marginal regression parameters of covariates, which, for ease of notation, are assumed to be identical for all classes of individuals. Similarly, assuming that sibs are equicorrelated, $P(y_1, ..., y_n/y_F, y_M, x_1, ..., x_n)$ can be expressed as follows:

$$\sum_{j_1=0,1}, ..., \sum_{j_n=0,1} \left\{(-1)^{\Sigma_{i=1}^{n} j_i} C_{\alpha_{SS}}[F_1(y_1 - j_1/x_1, y_F, y_M), ..., F_n(y_n - j_n/x_n, y_F, y_M)]\right\}, \quad (4)$$

where α_{SS} is the parameter of association between sibs and where F_i is the marginal distribution function characterizing the ith offspring's status. As for the parents, a logistic representation can be used to describe the marginal distribution function:

$$F_i(U/x_i, y_F, y_M) = \begin{cases} 0 & \text{if } U = -1 \\ 1/(1 + e^{\lambda+\beta x_i+\gamma_{FO} y_F+\gamma_{MO} y_M}) & \text{if } U = 0 \\ 1 & \text{if } U = 1 \end{cases},$$

where γ_{FO} and γ_{MO} are the regression coefficients, for the parents' phenotypes, that characterize the familial aggregation between parents and offspring. According to this model, $\exp(\gamma_{FO})$ and $\exp(\gamma_{MO})$ are the classic pairwise ORs measuring the parent-offspring aggregation. In this formulation, α_{SS} characterizes the residual aggregation between sibs, after controlling for parent-offspring dependency. This sib-sib residual aggregation can be due to genes that have nonadditive effects and/or to shared environmental factors specific to the offspring. An estimate of the crude aggregation between sibs can be obtained by setting the γ coefficients equal to 0 in the logistic model. Finally, note that, since in this model the sibship is considered as a set of equicorrelated data, it is not possible to distinguish brother-brother, brother-sister, and sister-sister aggregations.

Extension to Segregation and Combined Segregation-Linkage Analyses of a Binary Trait

Let $\boldsymbol{g} = (g_F, g_M, g_1, ..., g_n)$ and $\boldsymbol{m} = (m_F, m_M, m_1, ..., m_n)$ be the familial genotypic vectors at an unobserved major locus influencing the trait and at a measured marker locus, respectively. The joint likelihood of the observations can be written as $L(\boldsymbol{y/x}) = \sum_g P(\boldsymbol{y/x}, \boldsymbol{g})P(\boldsymbol{g})$ for segregation analysis and as $L(\boldsymbol{y/x}, \boldsymbol{m}) = \sum_g P(\boldsymbol{y/x}, \boldsymbol{g})P(\boldsymbol{m/g})P(\boldsymbol{g})$ for segregation-linkage analysis, where the summation is over all the possible unobserved genotypes at the major locus. The first term of the likelihood function is the penetrance, which can be decomposed into two probabilities:

$$P(\boldsymbol{y/x}, \boldsymbol{g}) = P(y_F, y_M/x_F, x_M, g_F, g_M) \times P(y_1, ..., y_n/y_F, y_M, x_1, ..., x_n, g_1, ..., g_n).$$

These two probabilities are modeled by two different Frank's family copulas, as in equations (3) and (4). Genotypic effects at the major locus are defined, in a logistic scale, as the differences between the genotype-specific baseline parameters λ_g. The association parameters now describe the residual familial aggregation, after controlling for the major-gene effects. The genotypic probabilities are written as usually in segregation analysis or segregation-linkage analysis.

Ascertainment Correction

In segregation or segregation-linkage analysis of a binary trait, families usually are selected through a particular scheme of ascertainment, not randomly. The likelihood therefore must be modified to incorporate an ascertainment correction, as follows:

$$L(\boldsymbol{y/x}, \boldsymbol{m}, A) = \frac{L(\boldsymbol{y/x}, \boldsymbol{m})P(A/\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{m})}{\sum_y [L(\boldsymbol{y/x}, \boldsymbol{m})P(A/\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{m})]},$$

where A is the ascertainment event. Several formulations for ascertainment corrections can be found in the reports by Cannings and Thompson (1979) and Elston and Sobel (1979).

Estimation of Parameters and Hypothesis Testing

Estimation of parameters is performed by maximization of the likelihood of the sample. Hypothesis testing is performed by means of the likelihood-ratio criterion. We developed our own program and linked it to the GEMINI maximization procedure (Lalouel 1981).

Results

Using combined segregation-linkage analysis, we had shown previously that quantitative plasma ACE levels were under the control of a major gene in complete linkage disequilibrium with a measured insertion (I)/deletion (D) polymorphism at the ACE locus (Tiret et al. 1992). As an illustration of the model proposed above, we performed a new segregation-linkage analysis of ACE levels, dichotomized for the purpose of application into two classes according to the median of the distribution.

The sample included 95 nuclear families that had volunteered for a free health examination and that comprised both natural parents ($n = 190$) and at least two

Table 1

Mean Values (SDs) and Ranges for Age and for Quantitative Plasma-ACE Levels, in the Study Population

	Fathers (n = 95)	Mothers (n = 95)	Sons (n = 120)	Daughters (n = 82)
Age (in years):				
Mean (SD)	41.4 (4.2)	39.4 (3.5)	14.3 (3.0)	14.4 (3.2)
Range	32–58	32–49	7–20	7–21
Plasma ACE (in IU/liter):				
Mean (SD)	89.5 (29.5)	84.4 (27.5)	125.6 (45.1)	106.7 (39.4)
Range	31–150	28–165	43–252	27–239

offspring aged ⩾9 years (n = 222). Since the relation between quantitative ACE levels and age of offspring had been shown to be nonlinear, adjustment for age and age^2 was made prior to analysis, separately for sons and daughters. For parents, no adjustment for age was necessary. A binary variable then was defined by dichotomizing the age- and sex-adjusted ACE distribution, according to the median.

The segregation-linkage analysis assumed a marker I/D in linkage disequilibrium with a putative functional polymorphism at the ACE locus, which had two alleles, a and A, with A associated with high ACE levels. The parameters of the model were the marker allele I frequency; the frequencies of major allele a conditional on I (π_I) and D (π_D); the major genotype effects $\delta_{Aa} = \lambda_{Aa} - \lambda_{aa}$ and $\delta_{AA} = \lambda_{AA} - \lambda_{aa}$; and the four residual association parameters (α_{FM}, γ_{FO}, γ_{MO}, and α_{SS}). Since the families were selected randomly, no correction for ascertainment was performed.

The mean ages and quantitative plasma ACE levels for the sample are reported in table 1. The I/D polymorphism was in Hardy-Weinberg equilibrium, and allele I frequency was .43 ± .02. Our earlier report (Tiret et al. 1992) had shown that there was no correlation of quantitative ACE levels in parents and that the parent-offspring and the sib-sib correlations were not significantly different (r = .24 ± .04 for the common correlation).

Results of the familial analysis of the binary phenotype are reported in table 2. The model including familial aggregation (model 1) was better supported than model 0, which assumed no familial aggregation (χ^2 = 15.78, with 4 df; $P < .005$). The association between spouses was not significantly different from 0 (χ^2 = 0.09, with 1 df). The father-offspring and mother-offspring association parameters were not different (χ^2 = 0.04, with 1 df). The crude aggregation between sibs, estimated from a model that set the γ coefficients equal to 0 (data not shown), was significantly different from 0 (α_{SS} = 1.86 ± 0.86). This estimate would approximately correspond to a sib-sib OR of 2.40. After controlling for the parent-offspring dependency, the residual aggregation between sibs was no longer significant (α_{SS} = 1.44 ± 0.88; χ^2 = 2.97, with 1 df). The most parsimonious model of familial aggregation (model 2) indicated, for high ACE levels, a common parent-offspring OR of 2.10 (95% confidence interval [CI] 1.38–3.20) and no residual aggregation between sibs, after accounting for the parent-offspring dependency.

Model 3, which assumed a major gene in complete association with the I/D polymorphism, allowed us to test the effects associated with the I/D polymorphism. These effects were highly significant (χ^2 = 64.70, with 2 df; $P < .0001$). The genotype effects were compatible with an additive effect of the D allele (model 4 vs. model 3; χ^2 = 1.28, with 1 df). When an additive model was assumed, the D allele was associated with an OR of 3.68 (95% CI 2.59–5.23) for high ACE levels. After controlling for the effects of the I/D polymorphism, the parent-offspring aggregation was no longer significant (model 5 vs. model 4; χ^2 = 3.14, with 1 df). Because of convergence problems, we were unable to estimate π_I and π_D simultaneously. Model 6, which relaxed the constraint π_I = 1, was not better supported than model 5 (χ^2 = 0.94, with 1 df). On the other hand, model 7, which relaxed the constraint π_D = 0, had a better likelihood than model 5 (χ^2 = 5.94, with 1 df; $P < .02$). Finally, model 8, which assumed the existence of a major gene in linkage equilibrium with the I/D polymorphism ($\pi_I = \pi_D$), had a worse likelihood than model 7, with the same df.

The most parsimonious model, therefore, was a model specifying a major gene in complete linkage disequilibrium with the I/D polymorphism, since the A allele associated with high ACE levels was always carried by marker allele D. In this model, the frequency of major allele A was .36 ± .07, and this allele was associated with an OR of 12.62 (95% CI 4.27–37.29) for high ACE levels, compared with the OR of 3.68 associated with the D allele. The lack of residual familial aggregation after controlling for the major-gene effect suggested that the segregation of this major gene was the only source of familial resemblance. The inferences made from this analysis were very similar to those obtained

Table 2

Familial Analysis of Dichotomized ACE Levels

Parameter	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Frequency of marker allele I	.434	.434	.434	.434	.434	.434	.434	.434
Frequency of major gene *a:*								
Conditional on allele I (π_I)	[1]	[1]	[1]	[1]	[1]	.818	[1]	.745
Conditional on allele D (π_D)	[0]	[0]	[0]	[0]	[0]	[0]	.356	(.745)
Major genotype effect:								
δ_{Aa}	[0]	[0]	1.642	1.303	1.359	1.721	2.535	3.051
δ_{AA}	[0]	[0]	2.742	(2.606)	(2.718)	(2.442)	(5.070)	(6.102)
Residual aggregation:								
Spouses (α_{FM})	−.260	[0]	[0]	[0]	[0]	[0]	[0]	[0]
Father-offspring (γ_{FO})	.704	.741	.403	.415	[0]	[0]	[0]	[0]
Mother-offspring (γ_{MO})	.795	[.741]	[.403]	[.415]	[0]	[0]	[0]	[0]
Residual sib-sib (α_{SS})[a]	1.436	[0]	[0]	[0]	[0]	[0]	[0]	[0]
$-\text{Log}_e L$	577.23	578.80	546.45	547.09	548.66	548.19	545.69	576.52
Alternate model	0[b]	1	2	3	4	5	5	…
df	4	3	2	1	1	1	1	…
χ^2	15.78	3.14	64.70	1.28	3.14	.94	5.94	…

NOTE.—Square brackets indicate that the parameter is fixed to the value given, and parentheses indicate that the parameter is constrained.

[a] Aggregation when controlling for parent-offspring dependency.

[b] Model 0 is without familial aggregation

from the analysis of quantitative ACE levels (Tiret et al. 1992).

Discussion

We have described an extension of the Frank's family copula model proposed by Meester and MacKay (1994) for correlated categorical data. The model of Meester and MacKay (1994) applies to equicorrelated data, and the correlation between binary outcomes is expressed through a single association parameter.

Since a nuclear family cannot be viewed as a set of symmetric data, we proposed to decompose the family into two sets of equicorrelated data (parents and offspring), each characterized by a within-cluster association parameter (α_{FM} and α_{SS}, respectively). The dependency between parents and offspring was modeled through regression logistic coefficients γ_{FO} and γ_{MO}. This formulation has the advantage that it does not require an arbitrary ordering of sibs, as do regressive models (Bonney 1986, 1987). However, it does require an ordering between parents and offspring, but this ordering appears to be quite natural. Another advantage of this formulation is that association parameters are independent of the marginal distributions—in particular, the size of the families—unlike the conditional logistic models proposed by Connolly and Liang (1988) and Abel et al. (1993).

Some limitations of this model should be discussed: One is the absence of symmetry between the α's and the γ's, which precludes testing of their equality. However, both parameters can be easily interpreted in terms of the classic OR, making comparison of their magnitudes possible. Another limitation of this model is that, owing to the implicit symmetric nature of the data within a copula, the α_{SS} parameter cannot be split to allow for sex difference between sibs. Last, extension to multigenerational pedigrees, although possible, would rapidly become complex, since each class of relatives would have to be considered as a distinct copula. However, the problem of extension to pedigrees is not specific to this modeling and arises with most other models for binary data. In our view, the model proposed here mainly applies to the study of multifactorial traits, characterized by frequent genes and strong environmental correlations, which, in general, are investigated through samples of nuclear families rather than through extended pedigrees.

If one's main interest is to assess familial aggregation without performing segregation analysis, the copula model probably is not the most appropriate model, owing to the limitations mentioned above. The estimating equations (EE) technique, which allows specification of familial aggregation through the marginal OR, by only modeling the first- and second-order moments, is better suited. In addition, the EE method can accommodate any kind of familial dependency, whereas the proposed copula model assumes an equicorrelation between sibs. Several EE applications have been proposed for the analysis of correlated binary data (Liang and Beaty 1991; Zhao and Le Marchand 1992; Hsu and Zhao 1996; Trégouët et al. 1997). However, if the EE technique has proved to be very efficient for the analysis of correlated data, the numerous calculations required in segregation analysis cause its application to be less practical for

this type of analysis (Whittemore and Gong 1994), and a maximum-likelihood method seems to be more appropriate.

A major interest in the copula model is its computer tractability, which makes it more attractive than other, previously proposed models (Bonney 1992; Abel et al. 1993) for segregation or segregation-linkage analysis. Further studies are needed to explore the properties of this copula model. In particular, there is a close connection between bivariate copulas (Genest and MacKay 1986) and the method of modeling association in bivariate frailty models for survival data (Oakes 1989). This connection suggests that incorporation of a copula model in survival analysis methods for familial diseases with variable age at onset should be possible (Abel and Bonney 1990). Several other copulas have been described (Clayton 1978; Hougaard 1986), and investigation of how they can be incorporated into a familial analysis framework would be interesting. Simulation studies are also required, to compare this copula model to other models (Bonney 1992; Abel et al. 1993) in terms of power to detect a major gene and type I error. However, it should be kept in mind that segregation and segregation-linkage analyses have a relatively low power to detect susceptibility genes that have a modest effect, whatever the parameterization used for the familial aggregation.

We applied the copula model to real data on high ACE levels and compared the results with those previously obtained from a combined segregation-linkage analysis of quantitative ACE levels (Tiret et al. 1992). The main conclusions of both analyses were very similar, namely, the existence of a major gene in complete linkage disequilibrium with the I/D polymorphism and strongly influencing the "risk" of having high ACE levels. As in our earlier analysis (Tiret et al. 1992), there was no aggregation of the trait between spouses, and the interpretation that the aggregation between sibs was no longer significant after the parent-offspring dependency was considered is consistent with the equal parent-offspring and sib-sib correlations observed in the analysis of quantitative ACE levels. This result is also consistent with similar magnitudes in the parent-offspring and the crude sib-sib ORs (2.10 and 2.40, respectively). This pattern of familial resemblance is compatible with the absence of shared environmental factors specific to offspring that influence ACE levels. It also suggested the lack of dominance genetic variance, a feature that was confirmed by the additive allele effects inferred at the major locus and that accounted for the entire heritability. In our previous analysis, the residual familial resemblance for quantitative ACE levels was still significant after controlling for the I/D polymorphism and disappeared only after a major-gene effect was introduced. This slight difference from the analysis described here probably can be explained by a loss of power consecutive to the truncation of the continuous phenotype. This loss of power can be assessed roughly by comparison of the χ^2 values for the testing of the same specific hypotheses in both analyses. For example, in the analysis of quantitative ACE levels, the χ^2 value for the testing of residual familial resemblance after controlling for the I/D polymorphism was 9.39 (1 df), and that for the testing of the existence of a major gene in complete linkage disequilibrium with the I/D polymorphism was 10.48 (1 df). In the analysis of dichotomized ACE levels, the corresponding χ^2 values were 3.14 and 5.94, respectively, with the same number of df, indicating a clear loss of power.

In conclusion, this copula model based on Frank's family provides a flexible model for the analysis of familial binary data. Its main attractive features are that the association parameter is independent of marginal distributions, that varying cluster sizes can be accommodated easily, and that the technique is computationally tractable.

Acknowledgments

We wish to deeply thank S. G. Meester for providing his unpublished Ph.D. thesis. We also are deeply grateful to two anonymous reviewers for providing helpful suggestions on earlier drafts of this article.

References

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

Abel L, Golmard JL, Mallet A (1993) An autologistic model for the genetic analysis of familial binary data. Am J Hum Genet 53:894–907

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

——— (1987) Logistic regression for dependent binary observations. Biometrics 43:951–973

——— (1992) Compound regressive models for family data. Hum Hered 42:28–41

Cannings C, Thompson E (1979) Ascertainment in the sequential sampling of pedigrees. Clin Genet 12:208

Clayton DG (1978) A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika 65: 141–151

Connolly M, Liang K (1988) Conditional logistic regression models for correlated binary data. Biometrika 75:501–506

Demenais F (1991) Regressive logistic models for familial diseases: a formulation assuming an underlying liability model. Am J Hum Genet 49:773–785

Elston R, Sobel E (1979) Sampling considerations in the gathering and analysis of pedigree data. Am J Hum Genet 31: 62–69

Frank M (1979) On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$. Aequationes Math 19:194–226

Genest C (1987) Frank's family of bivariate distributions. Biometrika 74:549–555

Genest C, MacKay J (1986) The joy of copulas: bivariate distributions with uniform marginals. Am Stat 40:280–283

Hougaard P (1986) A class of multivariate failure time distributions. Biometrika 73:671–678

Hsu L, Zhao LP (1996) Assessing familial aggregation of age at onset, by using estimating equations, with application to breast cancer. Am J Hum Genet 58:1057–1071

Lalouel J (1981) GEMINI: a computer program for optimization of general nonlinear functions. Tech rep 14, Department of Medical Biophysics and Computing, University of Utah, Salt Lake City

Lalouel J, Rao D, Morton N, Elston R (1983) A unified model for complex segregation analysis. Am J Hum Genet 35: 816–826

Liang KY, Beaty TH (1991) Measuring familial aggregation by using odds ratio regression models. Genet Epidemiol 8: 361–370

Meester SG, MacKay J (1994) A parametric model for cluster correlated categorical data. Biometrics 50:954–963

Morton N, MacLean C (1974) Analysis of family resemblance. III. Complex segregation analysis of quantitative traits. Am J Hum Genet 26:489–503

Oakes D (1989) Bivariate survival models induced by frailties. J Am Stat Assoc 84:487–493

Schweizer B, Sklar A (1983) Probabilistic metric spaces. North Holland, New York

Schweizer B, Wolff EF (1981) On parametric measures of dependence for random variables. Ann Stat 9:879–885

Tiret L, Rigat B, Visvikis S, Breda C, Corvol P, Cambien F, Soubrier F (1992) Evidence, from combined segregation and linkage analysis, that a variant of the angiotensin I–converting enzyme (ACE) gene controls plasma ACE levels. Am J Hum Genet 51:197–205

Tosteson T, Rosner B, Redline S (1991) Logistic regression for clustered binary data in proband studies with application to familial aggregation of sleep disorders. Biometrics 47: 1257–1265

Trégouët DA, Ducimetière P, Tiret L (1997) Testing association between candidate-gene markers and phenotype in related individuals, by use of estimating equations. Am J Hum Genet 61:189–199

Whittemore AS, Gong G (1994) Segregation analysis of case-control data using generalized estimating equations. Biometrics 50:1073–1087

Zhao LP, Le Marchand L (1992) An analytical method for assessing patterns of familial aggregation in case-control studies. Genet Epidemiol 9:141–154