BIOMETRICS 51, 542–551
June 1995

Score Test for Familial Aggregation in Probands Studies: Application to Alzheimer's Disease

Daniel Commenges, Hélène Jacqmin, Luc Letenneur, and Cornelia M. Van Duijn[1]

INSERM U330, 146 rue Leo Saignat, Bordeaux, 33076, France
[1]Department of Epidemiology and Biostatistics,
Erasmus University Medical School,
Rotterdam, The Netherlands

SUMMARY

When studying familial aggregation of a disease, the following two-stage design is often used: first select index subjects (cases and controls); then record data on their relatives. The likelihood corresponding to this design is derived and a score test of homogeneity is proposed for testing the hypothesis of no-aggregation. This test takes into account the selection procedure and allows adjustment to be made for explanatory variables. It appears as the sum of three terms: a pure test of homogeneity, a test of comparison of observed minus expected cases in the two groups, and a term which adjusts for the possible unequal probabilities of disease of the index subjects. Asymptotic efficiency and a simulation study show that the proposed test is superior to either the pure homogeneity test or tests based on the comparison of numbers of affected in the two groups. The test statistic, which has an asymptotically standard normal distribution, is applied to a study of familial aggregation of early-onset Alzheimer's disease for which a highly significant value (9.46) is obtained: this is the highest value among the three tests compared, in agreement with the simulation study. A logistic normal model is fitted to the data, taking account of the selection procedure: it allows to estimate the regression parameters and the variance of the random effect; the likelihood ratio test for familial aggregation seems less powerful than the score test.

1. Introduction

In many multifactorial diseases such as cancers, Alzheimer's disease, and chronic obstructive pulmonary disease (Cohen, 1980), a genetic determinant is suspected but it is difficult to specify the genetic mechanism. In a first stage, it is necessary to verify whether familial aggregation exists. Even this more modest aim is not easy to achieve, because the aggregation is weaker than in simpler genetic diseases and it is blurred by the influence of other factors. These factors may weaken the apparent aggregation or, on the contrary, they may be responsible for a familial aggregation of the disease. The relevant epidemiological issue is to assess whether there is a familial aggregation which cannot be explained by non-genetic factors.

Studies of familial aggregation of chronic diseases often have a particular design: they have been called proband studies (Tosteson, Rosner, and Redline, 1991) or case-control relatives studies (Commenges and Letenneur, 1992). The design is the following: in a first stage, a sample of index cases and a sample of index controls are constituted; then data on the relatives of these index subjects are collected. The simplest method of analysis is the comparison of the proportions of affected relatives in the case relatives group and in the control relatives group. Commenges and Letenneur (1992) have analyzed this simple situation in order to derive sample size formulas. However, at the stage of analysis it is important to use more sophisticated methods for both achieving increased power and adjusting for non-genetic factors.

Neuhaus and Jewell (1990) have stressed the importance of taking account of the sampling scheme when analyzing binary correlated data. Tosteson et al. (1991) have proposed an adaptation of the Rosner model (Rosner, 1984) for proband studies. Zhao and Lemarchand (1991) have proposed to estimate and test the odds-ratio of the group label using the GEE approach.

The aim of this paper is to derive a score test of familial aggregation which can be used in proband

Key words: Alzheimer's disease; Familial aggregation; Homogeneity; Score test.

studies and which can take into account explanatory variables. The statistic takes a particularly appealing form when a logistic model is chosen. A simulation study of the power of the test is presented. Finally, the test is applied to a study of familial aggregation of early-onset Alzheimer's disease. Also, the logistic normal model is fitted to the data, taking account of the selection procedure.

2. The Score Test

2.1 *The Random Effect Model*

We suppose the following random effect model which is an extension of the model used by Donald and Donner (1987) and by Liang (1987). The probability for family i is given by:

$$\mathrm{pr}(Y_i|\alpha_i) = \mathrm{pr}(Y_i;\ \alpha_i,\ X_i),$$

where $Y_i = \{y_{ij}, j = 1, \ldots, s_i\}$, $X_i = \{x_{ij}, j = 1, \ldots, s_i\}$, in which y_{ij} is the status of subject j, x_{ij} is the vector of explanatory variables for this subject and s_i the size of the family. The random effect α_i can be written:

$$\alpha_i = \alpha + \theta^{1/2} v_i,$$

where v_i has a distribution G and $E(v_i) = E(v_i v_j) = 0$; $E(v_i^2) = 1$. The distribution G need not be specified. The null hypothesis of no-aggregation, adjusted on the values of x_{ij} is "$\theta = 0$" and the alternative is "$\theta > 0$".

In order to obtain more specific results we shall later make the hypothesis of independence of the y_{ij} values conditionally on α_i and specify a model relating their distribution to the values of the x_{ij} values.

The model for the selection procedure will be that of single ascertainment: each affected (respectively, non-affected) subject in the population has the same probability τ (respectively, τ') of being selected as an index case (respectively, control); τ and τ' are supposed to be small.

2.2 *The Likelihood*

The data consist of the statuses of the subjects $Y_i = \{y_{ij}, j = 1, \ldots, s_i\}$ together with the explanatory variables X_i for $i = 1, \ldots, n$. The sample is divided into two subsamples: the subsample of n_1 families selected via a case index, and the subsample of n_0 families selected via a control index, so that $n = n_0 + n_1$. We shall use the group label function $k(i)$ which takes the value 1 if family i was selected via a case index, 0 if selected via a control index. Without loss of generality we rank the families so that $k(i) = 1$, $i = 1, \ldots, n_1$ and we give the rank 1 to the index subject in each family so that $y_{i1} = k(i)$.

It is necessary to condition on the event that the family has been selected in the case or control group. We introduce the selection variable ε_i which takes the value $k(i)$. It is also natural to condition on the event that among the members of the family, this is the subject $j = 1$ which is the index subject, event which we denote "$j_I = 1$". This latter conditioning does not affect the likelihood if all the subjects in the same family have the same probability of being a case, but it does in the general setup. The likelihood (marginal relatively to α_i) for family i of the case group is

$$L_i = \mathrm{pr}(Y_i|\varepsilon_i = 1, j_I = 1) = \frac{\mathrm{pr}(j_I = 1|\varepsilon_i = 1, Y_i)}{\mathrm{pr}(j_I = 1|\varepsilon_i = 1)} \frac{\mathrm{pr}(\varepsilon_i = 1|Y_i)}{\mathrm{pr}(\varepsilon_i = 1)} \mathrm{pr}(Y_i).$$

Under the hypothesis of single ascertainment (Elandt-Johnson, 1971) each affected subject has a small probability τ of being selected and we have

$$\mathrm{pr}(\varepsilon_i = 1|Y_i) = 1 - (1 - \tau)^{d_i} \approx \tau d_i$$

and

$$\mathrm{pr}(\varepsilon_i = 1) = E(\tau d_i) = \tau \sum_{j=1}^{s_i} \pi_{ij}.$$

where d_i is the number of affected in family i and $\pi_{ij} = \mathrm{pr}(y_{ij} = 1)$. We have also

$$\mathrm{pr}(j_I = 1|\varepsilon_i = 1) = \mathrm{pr}(\text{subject } \{i, 1\} \text{ be selected } |\varepsilon_i = 1)$$

$$= \pi_{i1} \Big/ \sum_{j=1}^{s_i} \pi_{ij} \text{ and } \mathrm{pr}(j_I = 1|\varepsilon_i = 1, Y_i) = 1/d_i.$$

Thus we obtain

$$l_i = \frac{\mathrm{pr}(Y_i)}{\pi_{i1}} = \mathrm{pr}(Y_i|y_{i1} = 1).$$

This result, although restricted to random effect models is stronger than the result obtained by Tosteson et al. (1990), since it says not only that this conditional likelihood is independent of the selection event but that it is the natural likelihood to consider.

Under the null hypothesis of independence we have $\mathrm{pr}(Y_i|y_{i1} = 1) = \mathrm{pr}(Y_{i-})$, where $Y_{i-} = \{y_{ij}, j = 2, \ldots, s_i\}$ is the observation of the relatives of the index.

For families of the control group, we obtain

$$L_i = \frac{\mathrm{pr}(Y_i)}{1 - \pi_{i1}} = \mathrm{pr}(Y_i|y_{i1} = 0).$$

In terms of conditional probability, we have $\mathrm{pr}(Y_i) = \int \mathrm{pr}(Y_i|v_i)\, dG(v_i)$ and $\pi_{ij} = \int p_{ij}(v_i) dG(v_i)$. Thus the total loglikelihood is

$$L = \sum_{i=1}^{n} \log \int \mathrm{pr}(Y_i|v_i)\, dG\,(v_i) - \sum_{i=1}^{n} \log \int p_{i1}^{k(i)}(v_i) q_{i1}^{1-k(i)}(v_i)\, dG(v_i). \tag{1}$$

Note that the second term in this likelihood is the correction for selection. If we assume a particular model and distribution G, this likelihood can be computed by numerical integration and it is possible to estimate all the parameters in the model by maximizing it. Also, a likelihood ratio statistic for the hypothesis "$\theta = 0$" can be computed using the difference between the maximized loglikelihood (1) and the maximized loglikelihood of the model under this hypothesis. Since the value of θ to be tested is on the boundary of its domain, the correct test is to consider the square root of the conventional likelihood ratio statistic; under the null hypothesis it has asymptotically a standard normal distribution and the null hypothesis is rejected for large positive values of this statistic (Self and Liang, 1987).

2.3 *The Score Statistic*

Applying L'Hospital's rule as in Liang (1987), we obtain the score statistic:

$$S = \frac{1}{2} \sum_{i=1}^{n} \{[\partial \log \mathrm{pr}\, (Y_i|\alpha_i)/\partial \alpha_i]^2 + \partial^2 \log \mathrm{pr}\, (Y_i|\alpha_i)/\partial \alpha_i^2\}$$

$$- \frac{1}{2} \sum_{i=1}^{n} \{[\partial \log \mathrm{pr}\, (y_{i1}|\alpha_i)/\partial \alpha_i]^2 + \partial^2 \log \mathrm{pr}\, (y_{i1}|\alpha_i)/\partial \alpha_i^2\},$$

where all the derivatives are taken at $\theta = 0$. If we make the hypothesis of conditional independence, we have $\mathrm{pr}(Y_i|\alpha_i) = \Pi \mathrm{pr}(y_{ij}|\alpha_i)$ and then:

$$\partial \log \mathrm{pr}(Y_i|\alpha_i)/\partial \alpha_i = \sum_{j=1}^{s_i} U_{ij} \quad \text{and} \quad \partial^2 \log \mathrm{pr}(Y_i|\alpha_i)/\partial \alpha_i^2 = -\sum_{j=1}^{s_i} V_{ij}$$

where

$$U_{ij} = \partial \log \mathrm{pr}(y_{ij}|\alpha_i)/\partial \alpha_i \quad \text{and} \quad V_{ij} = -\partial^2 \log \mathrm{pr}(y_{ij}|\alpha_i)/\partial \alpha_i^2.$$

S can then be written

$$S = \sum_{i}^{n} \sum_{j=1}^{s_i} \sum_{j'=j+1}^{s_i} U_{ij} U_{ij'} + \frac{1}{2} \sum_{i}^{n} \sum_{j=2}^{s_i} (U_{ij}^2 - V_{ij}).$$

The first term appears as a covariance term and is the numerator of the pair-wise correlation coefficient computed on all the subjects. The second term is the score statistic for overdispersion (Cox, 1983) applied to the relatives. Yet another way to write the formula is

$$S = S_U^R + \sum_{i=1}^{n} \sum_{j=2}^{s_i} U_{ij} U_{i1}$$

where S_U^R is the score statistic of homogeneity applied to the relatives and ignoring the selection, that is

$$S_U^R = \frac{1}{2} \sum_{i=1}^{n} \left[\left(\sum_{j=2}^{s_i} U_{ij} \right)^2 - \sum_{j=2}^{s_i} V_{ij} \right].$$

The second term then represents the information brought by the selection procedure.

The most natural model for $\mathrm{pr}(y_{ij})$ is the logistic model

$$p_{ij}(v_i) = \frac{e^{\alpha_i + \beta' x_{ij}}}{1 + e^{\alpha_i + \beta' x_{ij}}}$$

where β is a vector of regression coefficients. It is easy to verify that with this model

$$U_{ij} = y_{ij} - p_{ij} \quad \text{and} \quad V_{ij} = p_{ij} q_{ij},$$

where p_{ij} is the value of $p_{ij}(v_i)$ at $\theta = 0$. The statistic can then be written

$$S = \frac{1}{2} \sum_{i=1}^{n} \left\{ \left[\sum_{j=2}^{s_i} (y_{ij} - p_{ij}) \right]^2 - \sum_{j=2}^{s_i} p_{ij} q_{ij} \right\} + \sum_{i=1}^{n_1} \sum_{j=2}^{s_i} (y_{ij} - p_{ij}) - \sum_{i=1}^{n} p_{i1} \sum_{j=2}^{s_i} (y_{ij} - p_{ij}).$$

Note that the summations on j are on the relatives.

In general α and β are unknown and will be replaced by their maximum likelihood (ML) estimators under the null hypothesis. Under the null hypothesis the likelihood of family i is simply $\mathrm{pr}(Y_{i-})$. Thus the ML estimators of the regression parameters are obtained by conventional logistic regression applied to the relatives (that is, all the subjects excluding the index subjects). When the p_{ij} values are replaced by $\hat{p}_{ij}$, we have $\Sigma_i^n \, \Sigma_{j=2}^{s_i} (y_{ij} - \hat{p}_{ij}) = 0$ so that

$$\sum_{i}^{n_1} \sum_{j=2}^{s_i} (y_{ij} - \hat{p}_{ij}) = D_1 - E(D_1) = E(D_0) - D_0$$

where D_1 and D_0 are the total numbers of cases in the case and control relatives groups and $E(D_1)$ and $E(D_0)$ are their expectations.

Finally the score statistic is

$$S = \frac{1}{2} \sum_{i=1}^{n} \left\{ \left[d_i - E(d_i) \right]^2 - \sum_{j=2}^{s_i} \hat{p}_{ij} \hat{q}_{ij} \right\} + D_1 - E(D_1) - \sum_{i=1}^{n} p_{i1} [d_i - E(d_i)], \qquad (2)$$

where $d_i = \Sigma_{j=2}^{s_i} y_{ij}$; $D_1 = \Sigma_{i=1}^{n_1} d_i$; $E(d_i) = \Sigma_{j=2}^{s_i} \hat{p}_{ij}$; $E(D_1) = \Sigma_{i=1}^{n_1} E(d_i)$ and the $\hat{p}_{ij}$ values are computed from the conventional logistic regression model.

If there are no explanatory variables, the last term is null and the term brought by the selection takes the form

$$D_1 - E(D_1) = N_1(\hat{\pi}_1 - \hat{p}) = \frac{N_1 N_0}{N} (\hat{\pi}_1 - \hat{\pi}_0)$$

where N_k, $k = 0, 1$ are the number of relatives in group k, $N = N_1 + N_0$, $\hat{\pi}_k = D_k / N_k$, $k = 0, 1$, and $\hat{p} = (D_1 + D_0)/N$. In this case the term brought by the selection involves the difference of proportions of affected relatives—the naive test statistic—weighted by $N_1 N_0 / N$; $\hat{\pi}_1$ and $\hat{\pi}_0$ are estimates of the marginal probabilities π_1 and π_0 in the two groups under the alternative hypothesis. The statistic takes the form

$$S = \frac{1}{2} \sum_{i=1}^{n} [(d_i - m_i \hat{p})^2 - m_i \hat{p} \hat{q}] + \frac{N_1 N_0}{N} (\hat{\pi}_1 - \hat{\pi}_0),$$

where m_i is the number of relatives ($m_i = s_i - 1$).

2.4 *Variance of the Score Statistic*

For more concise results denote

$$z'_{ij} = [1, x'_{ij}] \quad \text{and} \quad \gamma' = [\alpha, \beta'].$$

The variance of $\partial l/\partial\theta$ is

$$I = I_{\theta\theta} - I_{\theta\gamma} I_{\gamma\gamma}^{-1} I'_{\theta\gamma}$$

where $I_{\theta\theta} = \Sigma_{i=1}^{n} E(\partial l_i/\partial\theta)^2$; $I_{\gamma\gamma} = \Sigma_{i=1}^{n} E(\partial l_i/\partial\gamma)(\partial l_i/\partial\gamma)'$; and $I_{\theta\gamma} = \Sigma_{i=1}^{n} E(\partial l_i/\partial\theta)(\partial l_i/\partial\gamma)'$ where both $\partial l_i/\partial\theta$, $\partial l_i/\partial\gamma$ and the expectations are calculated at $\theta = 0$. After some computations (Appendix), we obtain

$$I_{\theta\theta} = \frac{1}{4}\sum_{i=1}^{n}\left[\sum_{j=2}^{s_i} p_{ij}q_{ij}(1 - 6p_{ij}q_{ij}) + 2\left(\sum_{j=2}^{s_i} p_{ij}q_{ij}\right)^2\right] + 2\sum_{i=1}^{n_1}\sum_{j=2}^{s_i} p_{ij}q_{ij}^2$$

$$- 2\sum_{i=1}^{n} p_{i1}\sum_{j=2}^{s_i} p_{ij}q_{ij}^2 - 2\sum_{i=1}^{n_1} p_{i1}\sum_{j=2}^{s_i} p_{ij}q_{ij} + \sum_{i=1}^{n} p_{i1}(1 + p_{i1})\sum_{j=2}^{s_i} p_{ij}q_{ij}$$

$$I_{\gamma\gamma} = \sum_{i=1}^{n}\sum_{j=2}^{s_i} p_{ij}q_{ij}z_{ij}z'_{ij}$$

$$I_{\theta\gamma} = \sum_{i=1}^{n}\sum_{j=2}^{s_i} p_{ij}q_{ij}\left(\frac{1}{2} - p_{ij}\right)z'_{ij} + \sum_{i=1}^{n_1}\sum_{j=2}^{s_i} p_{ij}q_{ij}z'_{ij} - \sum_{i=1}^{n} p_{i1}\sum_{j=2}^{s_i} p_{ij}q_{ij}z'_{ij}.$$

An estimate of I is obtained in replacing p_{ij} values by their ML estimators and the test statistic is $H_{sb} = S/I^{1/2}$.

3. Asymptotic Efficacy

Using the fact that $(S - \theta I)/I^{1/2}$ is asymptotically distributed as $N(0, 1)$ when θ tends toward zero and n_1 and n_0 tend toward infinity, it can be seen that the Pitman asymptotic efficacy (PAE) (Zacks, 1985) of S is just I/N. In the case where $n_1 = n_0$ and $m_i = m$ and there is no explanatory variable ($p_{ij} = p$), it reduces to

$$\text{PAE}(S) = \frac{1}{4}pq + \frac{1}{2}(m - 1)p^2q^2.$$

Using the same argument for S_U^R (which is the score statistic for a marginal likelihood ignoring the groups labels), we find $\text{PAE}(S_U^R) = \frac{1}{2}(m - 1)p^2q^2$. The PAE of $\hat{\pi}_1 - \hat{\pi}_0$ can be computed noting that its expectation is equal to the intracluster correlation coefficient ρ (Commenges and Letenneur, 1992) and its variance to $4pq$. Using the relation between ρ and θ and making use of L'Hospital's rule, we find that $\partial\rho/\partial\theta$ computed at $\theta = 0$ is equal to pq and thus, $\text{PAE}(\hat{\pi}_1 - \hat{\pi}_0) = pq/4$. Thus we have the relation

$$\text{PAE}(S) = \text{PAE}(\hat{\pi}_1 - \hat{\pi}_0) + \text{PAE}(S_U^R)$$

and the increase in PAE of S relative to that of $\hat{\pi}_1 - \hat{\pi}_0$ is $2(m - 1)pq$.

4. Simulation Study of the Power of the Test

We studied the power of the proposed test in the case of no covariate for different values of θ and for three different distributions of v: i) normal; ii) double exponential; iii) a discrete distribution. This last distribution is defined as follows

$$v = \begin{cases} -\sqrt{\dfrac{\omega}{1 - \omega}} & \text{with probability } 1 - \omega \\[2ex] \sqrt{\dfrac{1 - \omega}{\omega}} & \text{with probability } \omega \end{cases}$$

We used the value $\omega = .10$, which, for instance, is a plausible value of the proportion of people presenting a defective allele which increases the risk of late-onset Alzheimer's disease. We give the correspondence between θ and the intracluster coefficient ρ (computed by simulation) for the three distributions with a choice of the intercept α such that the probability of the disease is .05 for $v = 0$ (Table 1). The intracluster correlation coefficient is particularly attractive in proband studies because of the relation $\pi_1 - \pi_0 = \rho$. Very different values of ρ are achieved for the same value of θ in the three distributions.

Table 1

Correspondence between the value of θ and the value of the intracluster coefficient ρ in three different distributions of v.

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9	1
Discrete distribution	.010	.026	.048	.073	.102	.133	.166	.199	.233	.266
Double exponential	.006	.013	.023	.033	.044	.053	.075	.077	.093	.105
Normal distribution	.005	.011	.016	.024	.031	.039	.049	.057	.064	.076

For performing the simulations we should normally simulate a population and draw at random index cases and controls. We avoided this stage by using a rejection sampling algorithm. The distributions of the disease probability P in the two groups are given by (Commenges and Letenneur, 1992): $dG_{P|\varepsilon=1}(p) = pdG(p)/E(P)$ and $dG_{P|\varepsilon=0}(p) = (1 - p)dG(p)/[1 - E(P)]$. We generate a variable from $G_{P|\varepsilon=1}$ using the comparison function $dG_P/E(P)$: that is, we generate p from G_P and we accept it with probability p. For $\varepsilon = 0$ we accept p with probability $1 - p$.

We compared the power of H_{sb} at level .05 for the same value of ρ, .026, obtained with values of θ equal to .1, .35, .45 for the discrete, double exponential, and normal distributions, respectively; the estimated powers were .508, .501, and .538, respectively. Thus the power depends highly on the distribution for given value of θ but seems to be fairly independent of the distribution for given value of ρ.

Since the results depend essentially on the value of ρ, we performed the main simulations only for the discrete distribution (for simplicity of computations), for values of $\theta = 0, .1, .2, .3, .4, .5$ (corresponding to values of $\rho = 0, .01, .026, .048, .073, .102$), for sample sizes $n_1 = n_0 = n/2 = 100$ and number of relatives in each family $m_i = m = 2; 4; 6$. Each case was replicated 1,000 times. Table 2 gives the powers for the three statistics z (the conventional test for comparison of proportions) based on $\hat{\pi}_1 - \hat{\pi}_0$, H_b based on S_U^R, and H_{sb} based on S. It appears that the power of H_{sb} is always the largest; as expected the difference of power between H_{sb} and z increases with m, while H_b becomes more competitive.

Table 2

Power for the discrete distribution of the score test, H_{sb}, compared to the power of the test z based on the comparison of proportions, $\hat{\pi}_1 - \hat{\pi}_0$, and the pure test of homogeneity H_b based on S_U^R. The powers are for $n_1 = n_0 = 100$ and have been estimated with 1000 replications.

θ	0	.1	.2	.3	.4	.5
$m = 2$						
z	.042	.118	.301	.518	.795	.948
H_b	.066	.095	.152	.272	.405	.558
H_{sb}	.047	.129	.326	.581	.839	.966
$m = 4$						
z	.060	.155	.424	.784	.950	.997
H_b	.055	.133	.363	.686	.880	.986
H_{sb}	.059	.184	.576	.914	.994	1
$m = 6$						
z	.050	.191	.579	.891	.994	.999
H_b	.050	.180	.578	.919	.995	.999
H_{sb}	.052	.279	.782	.987	1	1

5. Familial Aggregation of Alzheimer's Disease

We applied the test to a study of familial aggregation of early-onset Alzheimer's disease (Hofman et al., 1989; Van Duijn et al., 1993). Although the disease is clearly hereditary in some families, familial aggregation in other families may result from clustering of longevity and the high risk of disease at old age. The design of the study we used is that of a case-control relatives study. The index cases were subjects who were diagnosed with Alzheimer's disease before the age of 70 years. The diagnosis required was that of probable Alzheimer's disease according to NINCDS-ADRDA (McKahnn et al., 1984) criteria. Each index case was paired to an index control matched for age (within 5 years), gender, and place of residence. Detailed data on family history were collected by interviewing a next of kin of the index case or control, and the information was verified by a sibling. The informants were asked specifically about the occurrence of dementia due to Alzheimer's disease in all first-degree relatives; subjects with a history of neurologic, psychiatric, or metabolic disorders other than Alzheimer's disease that may also lead to dementia were considered as unaffected. The age of each subject at the time of the study or at death was collected. The required information was available for 193 index cases and 194 index controls and for a total of 2421 relatives.

There were 32 affected among the 1142 relatives of index controls and 121 affected among the 1279 relatives of index cases. We computed the statistic H_{sb} in two models: i) unadjusted model, ii) model adjusted for age and sex. For both models we computed also the pure homogeneity statistic H_b based on S_R^U and a test based on an ordinary logistic regression (OLR) model in which the status of the index was entered as an explanatory variable; in the unadjusted model, this latter approach is equivalent to a simple comparison of proportions of affected relatives in the two groups; thus OLR is the extension of the test z.

The results are shown in Table 3: all the test statistics are much larger than the .05 critical value (1.64). However, it is necessary to adjust for age and sex, since both variables are significant. The statistic H_{sb} is higher than either H_b or the OLR statistic in both the adjusted and non-adjusted model; this is in agreement with the simulation study. For the adjusted model, the three terms of S in equation (2) are, respectively, equal to 29.8, 44.5, and $-.09$; thus both the difference between observed and expected numbers of cases and the homogeneity statistic S_R^U are important, while the third term is negligible.

Table 3

Values of four statistics for testing familial aggregation of Alzheimer's disease: H_b pure statistic of homogeneity; OLR ordinary logistic regression; H_{sb} proposed score statistic; RLR square-root of the likelihood ratio statistic from the logistic normal model. For the adjusted ordinary logistic regression model and for the logistic normal model, the odds ratios and (values of the Wald test statistic) are given; for age, these are the odds ratios for a difference of ten years; for the logistic normal model, the estimated value of θ is also given.

	Non-adjusted model	Adjusted model	Logistic normal model
Sex		1.62 (2.75)	1.82 (3.10)
Age		1.82 (7.82)	2.16 (8.27)
θ			1.66
H_b	4.70	6.06	
OLR	6.34	6.81	RLR = 8.59
H_{sb}	8.15	9.46	

We fitted also the logistic normal model proposed by Stiratelli, Laird, and Ware (1984) by maximizing the likelihood (1) specialized to this model; numerical integration with Gaussian quadrature was used to compute the likelihood as in Anderson and Aitkin (1985) and the Newton–Raphson algorithm was used for maximization. The likelihood ratio test was computed as twice the difference between the value obtained for (1) and the loglikelihood of an OLR model applied to the relatives. The estimated values of the regression coefficients and of their standard deviations (sex: .598 (.193); age: .0769 (.00930)) are slightly higher than those obtained by OLR (sex: .482 (.175); age: .0596 (.00762)); in Table 3 the more meaningful odds ratios and Wald statistics are given. The square root of the likelihood ratio test is larger than either H_b or the OLR statistic but smaller than H_{sb}.

We conclude that there is evidence of familial aggregation of early-onset Alzheimer's disease. It would be interesting to adjust for other explanatory variables which may also be aggregated in families. The increased power of H_{sb} should prove useful in studies of familial aggregation of late-onset Alzheimer's disease in which the familial aggregation is weaker.

6. Discussion

The first point to be noted is that there is a significant increase in power using the score test H_{sb} rather than methods which ignore the clustering of the data or which treat it as a nuisance like in

Zhao and LeMarchand (1992). This has an implication on the choice of the sample size to achieve a given power. Commenges and Letenneur (1992) recommended a formula giving a slightly larger size than usually required when comparing proportions. If the score test is used however, the power is higher for a given difference of marginal probabilities between the two groups $\pi_1 - \pi_0$, than would be obtained with independent samples; in this case, the usual sample size formulae are rather conservative and can be used.

The last term of the formula of S depends on p_{i1}: it will give a positive contribution if the p_{i1} are lower in the case-relatives group than in the control relatives group. Choosing index cases with low probability and index controls with high probability of the disease would probably increase the power of the design: for instance, in a proband study of Alzheimer's disease one could choose older control than case index subjects. However, epidemiologists might prefer pairing the case and control index subjects: since the likelihood is conditional on the data of the index subjects, the statistic does not need to be modified.

Finally, it is possible to fit the logistic normal model using the modified likelihood derived here to take into account the design of the study.

Acknowledgements

This work was essentially done during a stay of the first author at the Biostatistics Department of University of Washington in Seattle, supported by scholarship from NATO 22C91FR.

Résumé

Lorsque l'on étudie la concentration familliale d'une maladie, on utilise souvent le plan à deux étapes suivant: premièrement, selection de sujets index (case et témoins); deuxièmement, receuil des données concernant leurs apparentés. La vraisemblance correspondant à ce plan est donnée et un test du score d'homogénéité est proposé pour tester l'hypothèse de nonconcentration. Le test tient compte de la procédure de sélection et permet d'ajuster sur des variables explicatives. Il apparaît comme la somme de trois termes: un test d'homogénéité pur, un test de comparaison des effectifs observés dans les groupes par rapport à l'effectif attendu et un terme qui ajuste sur les possibles différences entre sujets index. L'éfficacité asymptotique et une étude de simulation montre que le test proposé est supérieur à la fois au test d'homogénéité pur ou au test basé sur la comparaisons des proportions d'affectés dans les deux groupes. La statistique de test, qui a asymptotiquement une répartition normale centrée réduite, est appliquée à l'étude de la concentration familliale de la maladie d'Alzheimer à début précoce et une valeur hautement significative (9.46) est atteinte: cette valeur est la plus grande parmi les trois tests utilisés, en accord avec l'étude de simulation. Un modèle logistique normal est ajusté aux données, en tenant compte de la procédure de sélection: il permet d'estimer les paramètres de régression et la variance de l'effet aléatoire; le test du rapport de vraisemblance semble moins puissant que le test du score.

References

Anderson, D. A. and Aitkin, M. (1985). Variance component models with binary response: Interviewer variability. *Journal of the Royal Statistical Society, Series B* **2,** 203–210.

Cohen, B. H. (1980). Chronic obstructive pulmonary disease: A challenge in genetic epidemiology. *American Journal of Epidemiology* **112,** 274–288.

Commenges, D. and Letenneur, L. (1992). Comparison of the proportions of affected relatives of cases and controls: Analysis and minimum sample size formula. *Statistics in Medicine* **11,** 1767–1776.

Commenges, D., Letenneur, L., Jacqmin, H., Moreau, T., and Dartigues, J.-F. (1994). Test of homogeneity of binary data with explanatory variables. *Biometrics* **50,** 613–620.

Cox, D. R. (1983). Some remarks on overdispersion. *Biometrika* **70,** 269–274.

Donald, A. and Donner, A. (1987). Adjustment to the Mantel–Haenszel chi-square statistic and odds ratio variance estimator when the data are clustered. *Statistics in Medicine* **6,** 491–499.

Elandt-Jonson, R. (1971). *Probability Models and Statistical Methods in Genetics*. New York: Wiley & Sons.

Hofman, A., Schulte, W., Tanja, T. A., et al. (1989). History of dementia and Parkinson's disease in 1st-degree relatives of patients with Alzheimer's disease. *Neurology* **39,** 1589–1592.

Liang, K. Y. (1987). A locally most powerful test for homogeneity with many strata. *Biometrika* **74,** 259–264.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Rice, D., and Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group. *Neurology* **34,** 939–944.

Neuhaus, J. M. and Jewell, N. P. (1990). The effect of retrospective sampling on binary regression models for clustered data. *Biometrics* **46,** 977–90.

Rosner, B. (1984). Multivariate methods in ophtalmology with application to other paired-data situations. *Biometrics* **40,** 1025–35.

Self, S. G. and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82,** 605–610.

Stiratelli, R., Laird, N. M., and Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics* **40,** 961–971.

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

Van Duijn, C. M., Farrer, L. A., Cupples, L. A., and Hofman, A. (1993). Genetic transmission of Alzheimer's disease among families in a dutch population-based study. *Journal of Medical Genetics* **30,** 640–646.

Zacks, S. (1985). Pitman efficiency. In *Encyclopedia of Statistical Sciences*, S. Kotz and N. L. Johnson (eds). New York: Wiley & Sons.

Zhao, L. P. and LeMarchand, L. (1992). An analytical method for assessing patterns of familial aggregation in case-control studies. *Genetic Epidemiology* **9,** 141–54.

Received March 1993; *revised November* 1993; *Accepted March* 1994.

APPENDIX:

Variance of the Score Statistic for the Logistic Model

Computation of $I_{\theta\theta}$

Noting that the score statistic can be written $S = S_U + \Sigma_{i=1}^{n} w_i U_i$ with $w_i = k(i) - 1$, where $U_i = \Sigma_{j=2}^{s_i} U_{ij}$ and $S_U = S_U^R$ is the score statistic for randomly chosen families applied to the relatives, we shall make use of the result obtained by Commenges et al. (1994):

$$E(S_U)^2 = \frac{1}{4}\sum_{i=1}^{n}\left[\sum_{j=2}^{s_i} p_{ij}q_{ij}(1 - 6p_{ij}q_{ij}) + 2\left(\sum_{j=2}^{s_i} p_{ij}q_{ij}\right)^2\right].$$

We have, using $E(U_i) = E(U_iU_j) = 0$

$$E(S)^2 = E\left[S_U^2 + 2S_U\left(\sum_{i=1}^{n} w_iU_i\right) + \left(\sum_{i=1}^{n} w_iU_i\right)^2\right]$$

$$= E(S_U^2) + E\left[\left(\sum_{i=1}^{n} U_i^2\right)\left(\sum_{i=1}^{n} w_iU_i\right)\right] + \sum_{i=1}^{n} w_i^2\text{var}(U_i)$$

$$= E(S_U^2) + \sum_{i=1}^{n} w_iE(U_i^3) + \sum_{i=1}^{n} w_i^2 \text{ var}(U_i).$$

We have also $E(U_i^3) = E[(\Sigma_{j=2}^{s_i} U_{ij})^3] = \Sigma_{j=2}^{s_i} E(U_{ij}^3)$ and $E(U_{ij}^2) = p_{ij}q_{ij}$, $E(U_{ij}^3) = p_{ij}q_{ij}(1 - 2p_{ij})$. Finally we find

$$E(S^2) = E(S_U^2) + \sum_{i=1}^{n} w_i \sum_{j=2}^{s_i} p_{ij}q_{ij}(1 - 2p_{ij}) + \sum_{i=1}^{n} w_i^2 \sum_{j=2}^{s_i} p_{ij}q_{ij}.$$

With $w_i = k(i) - p_{i1}$ we obtain

$$E(S^2) = E(S_U^2) + 2\sum_{i=1}^{n_1}\sum_{j=2}^{s_i} p_{ij}q_{ij}^2 - 2\sum_{i=1}^{n} p_{i1}\sum_{j=2}^{s_i} p_{ij}q_{ij}^2$$

$$- 2\sum_{i=1}^{n_1} p_{i1}\sum_{j=2}^{s_i} p_{ij}q_{ij} + \sum_{i=1}^{n} p_{i1}(1 + p_{i1})\sum_{j=2}^{s_i} p_{ij}q_{ij}.$$

Computation of $I_{\gamma\gamma}$

The information on γ comes from the likelihood valid for the null hypothesis pr(Y_-). Thus $I_{\gamma\gamma}$ has the same value than for the unselected problem

$$I_{\gamma\gamma} = \sum_{i=1}^{n} \sum_{j=2}^{s_i} p_{ij} q_{ij} z_{ij} z'_{ij}$$

Computation of $I_{\theta\gamma}$

We have

$$I_{\theta\gamma} = \sum_{i=1}^{n} E(\partial l_i/\partial\theta)(\partial l_i/\partial\gamma)' = E\left[\left(S_U + \sum_{i=1}^{n} w_i U_i\right)(\partial l_i/\partial\gamma)'\right].$$

The term $E[S_U(\partial l_i/\partial\gamma)']$ was given by Commenges et al. (1994) as being equal to $\Sigma_{i=1}^{n}$ $\Sigma_{j=2}^{s_i} p_{ij} q_{ij}$ $(½ - p_{ij}) z'_{ij}$. The additional term is

$$\begin{aligned} E\left[\left(\sum_{i=1}^{n} w_i U_i\right)\left(\sum_{i=1}^{n} \sum_{j=2}^{s_i} U_{ij} z'_{ij}\right)\right] &= \sum_{i=1}^{n} w_i \sum_{j=2}^{s_i} z'_{ij} E(U_{ij}^2) \\ &= \sum_{i=1}^{n} w_i \sum_{j=2}^{s_i} z'_{ij} p_{ij} q_{ij} \\ &= \sum_{i=1}^{n} \sum_{j=2}^{s_i} z'_{ij} p_{ij} q_{ij} - \sum_{i=1}^{n} p_{i1} \sum_{j=2}^{s_i} z'_{ij} p_{ij} q_{ij}. \end{aligned}$$