Environmental Health Perspectives
Vol. 87, pp. 27–35, 1990

Survival Models for Familial Aggregation of Cancer

by Wendy Mack,* Bryan Langholz,* and Duncan C. Thomas†

It has recently been shown that the relative risks of the order of 2 to 4 that are frequently found for cancer among relatives of affected cases are unlikely to be explainable by shared environmental risk factors. Classical methods of epidemiological analysis are not well suited to such analysis because they assume that the outcomes of each individual are independent. Classical methods of genetic analysis, on the other hand, are limited in their handling of environmental factors and variable ages of onset. The recent development of random effects models for survival analysis, however, appears to bridge this gap. Specifically, a proportional hazards model is postulated for the effects of measured covariates and of one or more components of frailty that are unmeasured but assumed to have some common distribution and known covariance structure within each family. From these assumptions, the posterior expectation of the hazard for each individual can be derived, given the covariate values and the observed and expected disease history of the family. These are then treated as known in a standard partial likelihood analysis; this is essentially a form of expectation-maximization algorithm. However, this does not provide a valid estimate of the covariance matrix because it fails to take account of the variability in the estimates of the frailties; an alternative approach using the imputation-posterior algorithm is suggested. This paper describes extensions of this approach to multivariate frailty distributions, modifications for application to pedigree and case-control studies, some simulation results, and applications to studies of breast cancer in twins and of lung cancer in relation to family smoking habits.

Introduction

Studies of familial aggregation of chronic disease pose considerable complexities for the data analyst. These include *a*) the correlation in outcomes resulting from the sharing of unmeasured genetic and/or environmental influences within families; *b*) the different degrees of sharing expected between different types of relatives; *c*) the censored survival time nature of the response variable, requiring the use of appropriate survival analysis techniques; and *d*) the need to consider multiple measured covariates (usually environmental but possibly including genetic markers); and *e*) the possibility of gene-environment interactions. Classical segregation and path analysis methods in genetics are designed specifically to address the first two issues, but are not well-suited to the latter three. Standard logistic and Cox regression techniques in epidemiology are well suited to dealing with the third and fourth issues but assume independent responses across individuals, and so they cannot be directly applied to the first two. The recent development of multivariate survival models (*1,2*) offers considerable promise to bridge the gap between these two approaches.

Basically the approach postulates the existence of an unobserved variable or variables (frailty) that take on the same or correlated values within a family and reflect unmeasured genetic and environmental influences on disease risk. Conditional on this frailty, the outcome is assumed to be independent between family members and to follow a proportional hazards model. In the analysis the unknown frailties are replaced by their posterior expectations, given the covariate values, periods of observation, outcomes of their family members, and current model parameters. These estimates are used as if they were known in a standard partial likelihood; this is essentially a form of expectation-maximization (EM) algorithm (*3*). In the next section we review this approach for the case where all family members are assumed to have the same frailty, and we discuss an approach to the problem of variance estimation using the imputation-posterior (IP) algorithm (*4*). The section "Multivariate Frailty Models" discusses some approaches to the case of correlated frailties. Our approach is closely related to that of Bonney (*5*), who proposes a logistic regression model for dichotomous outcomes using a regressive approach by ordering the subjects in the data set in such a way that each subject depends

*Department of Preventive Medicine, University of Southern California, Los Angeles, CA, 90033.

†MRC Biostatistics Unit, 5 Shaftsbury Road, Cambridge, CB2 2BW, England.

Address reprint requests to D. Thomas, MRC Biostatistics Unit, 5 Shaftsbury Road, Cambridge, CB2 2BW, England.

only on those preceding him or her in the data. A likelihood is constructed by assuming that the probability of each subject's outcome is given by a logistic function of his own observed covariates and unobserved genotype and then summing over all possible sets of genotypes, weighted by their prior probabilities under a particular inheritance model. Our approach generalizes this model to survival time data and does not require any ordering of the data.

The posterior estimate of the frailty is essentially a comparison of the observed number of cases in the family to an expected number based on the family members' ages and times at risk and their covariate values. The naturalness of this estimate can be seen by some calculations of familial relative risk (i.e., the risk of disease in relatives of an affected family member divided by that in relatives of unaffected family members) based on simple genetic models of single gene inheritance (*6*). These familial relative risks increase markedly with the number of affected family members but decrease only slowly with the number of unaffected members (Table 1). This occurs because if the disease is rare, each affected member considerably increases the posterior probability that the gene is present in the family, whereas each unaffected member only slightly reduces it, since the absence of a rare disease is not a particularly informative observation. An important corollary of this observation is that genetic relative risks (i.e., the risk of disease given the gene is present divided by that given the gene is absent) as large as 100 can easily produce familial relative risks of only 2 or 3. Conversely, since environmental risks this large are seldom observed, familial relative risks of 2 or 3 are difficult to explain by shared risk factors (*7*).

The method just described applies to the analysis of cohort data. A commonly used technique in genetics involves the ascertainment of affected probands and investigation of the disease history of their family members. This design is closely related to the case-control study in epidemiology in which affected cases are ascertained and matched with unaffected and unrelated controls, and their family histories are compared. These designs require some adaptation of the frailty analysis approach, which is discussed in the section "Modifications for Proband and Case-Control Designs."

We hoped that the use of multivariate frailty models of the third section would allow the effects of unmeasured genotype and unmeasured environment to be separated. The "Simulation Results" section describes some simulation studies suggesting that, although the tendency toward familial aggregation can be easily detected by these methods, separation of separate components of frailty can be difficult. Some applications to a cohort study of breast cancer in twins and a case-control study of lung cancer in relation to smoking and disease histories are described in "Applications."

Table 1. Familial relative risks by number of affected and total siblings.[a]

	Number of siblings				
Number of diseased siblings	1	2	3	4	5
1	1.20	1.17	1.16	1.16	1.16
2		1.33	1.31	1.31	1.30
3			1.45	1.44	1.43
4				1.56	1.55
5					1.66

[a] Familial relative risks computed for a recessive single gene model, assuming a genetic relative risk of 10, a population proportion of genetically susceptible individuals of 0.01, and a disease probability of 0.00001 in nongenetically susceptible individuals.

Univariate Frailty Models for Familial Cohort Data

The univariate frailty model for proportional hazards can be written as

$$\lambda(t,z_{ij},\epsilon_i) = \lambda_0(t)\exp(\underline{z}_{ij}'\underline{\beta} + \epsilon_i)$$

where $i = 1, \ldots, I$ indicates the family, $j = 1, \ldots m_i$ the members of the family λ and $N = \Sigma m_i$. Note that in this model, the unobserved frailties ϵ_i are assumed to be the same for all members of the family; in the third section we consider extensions in which the ϵ_i are m_i-vectors having some known covariance structure. By integrating over possible values of the unknown frailty, Clayton and Cuzick (*2*) and Self and Prentice (*1*) show that if the ϵ_i are assumed to have a log-gamma distribution with zero mean and variance γ, then the posterior expectation of the hazard is given by

$$\lambda(t,z_{ij}) = \lambda_0(t)\exp(\underline{z}_{ij}'\underline{\beta})\ \mathcal{E}[\exp(\epsilon_i)] \qquad (1a)$$

where

$$\mathcal{E}[\exp(\epsilon_i)] = \frac{1+\gamma D_i}{1+\gamma E_i}\,. \qquad (1b)$$

D_i is the observed number of cases in family i and E_i is the expected number of cases, given their ages at risk and covariate values, under the null hypothesis of no shared frailty, i.e.,

$$E_i(t) = \sum_{k=1}^{m_i} \int_0^t Y_{ik}(u)\,\lambda_0(u)\exp(\underline{z}_{ik}'\underline{\beta})\,du \qquad (1c)$$

where $Y_{ik}(u)$ is an indicator function for whether subject ik is at risk at time u. If one wished to use a standard Cox regression program, Eq. (1*b*) could be approximated by $\exp[\gamma(D_i - E_i)]$ using a Taylor expansion of the posterior expectation of $\mathcal{E}[\exp(\epsilon_i)]$.

Fitting the model is easily accomplished by form of EM algorithm:

a) Given trial estimates of $\lambda_0(t)$, the $E_i(t)$ are computed using Eq. (1*c*) and the partial likelihood based on the general relative risk model (1*a*) and (1*b*) is maximized with respect to model parameters $\underline{\beta}$ and γ;

b) New estimates of $\lambda_0(t)$ are then computed using the Cox-Breslow-Oakes estimator (*8*),

$$\hat{\lambda}_0(t) = \frac{\Sigma_{ij}\, n(t_k)}{(t_k - t_{k-1})\, \Sigma_{ij}\, Y_{ij}(t) \exp(\underline{z}_{ij}'\underline{\beta})\ \mathcal{E}[\exp(\epsilon_i)]}$$

where $n(t_k)$ is the number of failures at t_k and $t \epsilon (t_{k-1},t_k)$.

We found that convergence was improved if the iteration in step *a* was taken to convergence before a new set of baseline rates was estimated in step *b*; usually, only two or three cycles were then needed to obtain overall convergence.

Self and Prentice (*1*) estimate the frailties conditional on F_t, the history of events, covariate values, and censoring up to time *t*, in order to avoid having the hazard at time *t* depend on events in the future. However, this approach fails to produce a true partial likelihood due to the dependence of the frailty estimates on γ, which is estimated from the entire data set including events in the future. It also has the undesirable result that the estimate of the frailty for each family varies as information about the family accumulates, even though the true frailty is assumed to be constant over time; thus, less information is used to predict the frailty effect for the early failures than for the later failures. For this reason, Clayton and Cuzick (*2*) prefer to base their frailty estimate on the lifetime history of the entire family, which should produce a more efficient estimator.

Self and Prentice do not give a variance estimator for $\underline{\beta}$ and γ, although Clayton and Cuzick do. The problem with naive use of the information matrix from the partial likelihood is that it assumes the covariates, including ϵ_i, are known. The obvious correction would be to compute derivatives of the log likelihood with respect to $\underline{\beta}$ and γ including terms for the dependence of $\hat{\epsilon}_i$ on $\underline{\beta}$, γ, and λ_0. But the λ_0 in turn depend on $\underline{\beta}$, γ, and $\hat{\epsilon}_i$, thus leading to an infinite recursion, which seems intractable. A more promising approach would be to use the full survival likelihood and invert the full information matrix with respect to $\underline{\beta}$, γ, and $\underline{\lambda}_0$. Representing the full survival likelihood as

$$L^F(\underline{\beta},\underline{\lambda}_0,\epsilon_i) = \prod_{ij} [\lambda_0(t_{ij}) \exp(\underline{z}_{ij}'\underline{\beta} + \epsilon_i)]^{D_{ij}} \exp[-\Lambda_0(t_{ij}) \exp(\underline{z}_{ij}'\underline{\beta} + \epsilon_i)],$$

a potential approach would be to use the full expected likelihood, L^{FE}, obtained by integrating out the unknown frailties,

$$\begin{aligned} L^{FE} &= \Pi_i \int \Pi_j\, L^F_{ij}(\underline{\beta},\underline{\lambda}_0,\epsilon_i)\, P(\epsilon_i \mid \gamma)\, d(\epsilon_i \mid \gamma) \\ &= \Pi_i \{[\Pi_j\, \lambda_0(t_{ij})^{D_{ij}} \exp(\underline{z}_{ij}'\underline{\beta} D_{ij})] \\ &\quad \times \left[\Sigma_j\, \Lambda_0(t_{ij}) \exp(\underline{z}_{ij}'\underline{\beta}) + \frac{1}{\gamma}\right]^{-D_{i.}-1/\gamma} \\ &\quad \times \Gamma\left(D_{i\,.} + \frac{1}{\gamma}\right) / \gamma^{1/\gamma}\, \Gamma(1/\gamma)\}. \end{aligned}$$

A more attractive approach is to use the IP algorithm (*4*), which provides a Monte Carlo estimate of the entire posterior distribution of model parameters. Essentially one would proceed as follows. Given the current estimates of $\underline{\beta}$, γ, and $\underline{\lambda}_0$, one would draw a single random sample of ϵ_i for each family from their posterior distributions, which is also gamma with parameters $1/\gamma + D_i$ and $(1/\gamma + E_i)^{-1}$. Then treating these as known, one would draw a single random sample of $\lambda_0(t)$, $\underline{\beta}$, and γ from their respective posterior distributions, given the current estimates of the other parameters. The process continues indefinitely, and after a sufficient number of iterations, it settles down to produce successive random samples from the posterior distributions of each of the parameters. Details of this approach are given in Appendix 1.

Multivariate Frailty Models

The assumption of a common frailty ϵ_i for all members of the family is simplistic in that family members share different degrees of genetic and environmental influences. This suggests that one might consider replacing ϵ_i by an m_i-vector, whose elements have some covariance structure determined by their relationships to each other. The log-gamma distribution is not easily generalized to multivariate settings, but a multivariate normal distribution provides a close approximation. By fitting the first two derivatives of the posterior distribution of ϵ_i, Mack (*9*) has developed approximate expressions for the posterior expectation of ϵ_i which can be used as described above. For example, suppose there are two components of frailty, one a genetic effect and one an environmental effect; let ϵ_{ijp} $(p = 1,2)$ denote these two unobserved variables and let ρ_{ijkp} denote the correlation in $\epsilon_{i \cdot p}$ between members *j* and *k* of family *i* (e.g., 1 for the genetic factor for monozygotic twins, ½ for first degree relatives, ¼ for second degree relatives, etc.; c_1 for the environmental factor between spouses, c_2 between sibs, c_3 between parents and offspring, etc.). Then we would estimate ϵ_{ijp} by

$$\sum_{k=1}^{m_i} (D_{ik} - E_{ik})\, \rho_{ijkp}\ .$$

As before, we take the regression coefficient for frailty to be unity and estimate the variance of each component of the frailty distribution and any unknown parameters in their correlation matrices.

Unfortunately, for simple family structures, estimates of the variance and correlation in frailty are virtually colinear. Thus, a strong familial aggregation can be equally well explained either by a large variance with a small correlation or by a small variance with a high correlation (Fig. 1). Thus, it would be necessary to arbitrarily fix at least one of the unknown correlations and hope to estimate the other correlations relative to it, leaving the variance free; with sufficiently rich family structures this may be feasible, although the simulations described later are not encouraging.

We have also considered a two-point frailty distribution based on single gene models of inheritance. These are expressed in terms of two parameters, the gene frequency and the genetic relative risk, with the association between family members determined by Men-

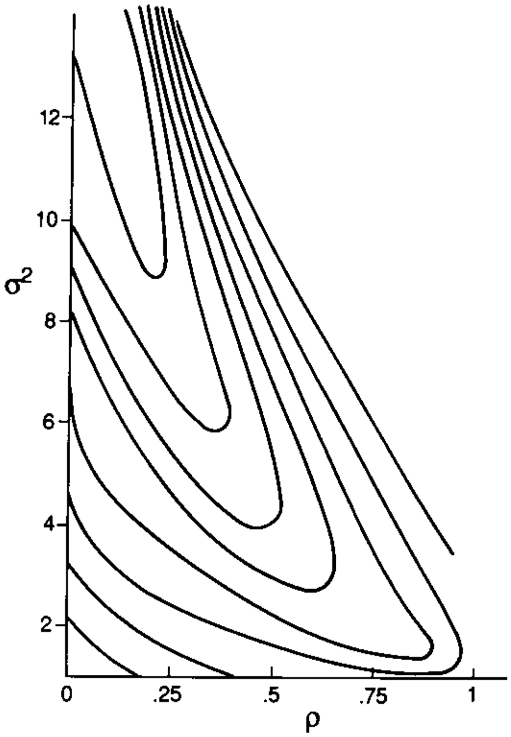

FIGURE 1. Likelihood surface for bivariate frailty.

delian laws. Detailed expressions are given elsewhere (*9*) and summarized in Appendix 2. Unfortunately, this approach also seems to suffer from similar problems of colinearity: there is no obvious way to simultaneously estimate both the gene frequency and the genetic relative risk. It appears that, without other information to estimate the gene frequency, it would be difficult to estimate the two parameters simultaneously, although complex family structures may be more informative.

Again, it is tempting to consider an approach using the IP algorithm. For each individual, one would draw a random ϵ_{ij} from its posterior distribution given the subject's disease status D_{ij}, his expected incidence E_{ij} [Eq. (1*c*) except for the summation over family members], and the current values of ϵ_{ik} for his family members. This is simplified by knowing the conditional independence structure of the family: for example, for a genetic model, ϵ_{ij} is dependent only on the ϵ_{ik} of his or her parents, spouse, and offspring. The rest of the iteration proceeds as described above. Details are given in Appendix 1.

Modifications for Proband and Case-Control Designs

In pedigree analysis, cases of a disease are ascertained from some disease registry (these are called the probands). All their family members are identified and their disease status determined (*10*). Thus, probands have the disease with probability one by virtue of the sampling scheme. The standard approach in genetic analysis is to exclude these subjects and determine whether the occurrence of disease in the rest of the family is consistent with Mendelian laws given the presence of the affected proband. We propose to use the same principle in fitting the frailty model. Thus, the cohort would consist only of the nonprobands, who would be viewed as a birth cohort. However, the probands would be counted as observed events in Eq. (1*b*) for estimating the frailties of the nonprobands, calculating their E_{ij} by applying the estimated baseline rates from the cohort of nonprobands to their time from birth to diagnosis.

This can be viewed as a purely internal analysis, which does not take advantage of the information that the cohort may have a higher incidence rate than the population at large. Such a higher incidence rate would presumably be a reflection of the same phenomenon that leads to clustering within families and an analysis that takes advantage of that information should be more powerful, particularly if the number of nonproband cases in the family is small. Assuming a set of population rates were available and thought to be applicable to the cohort (i.e., comparable ascertainment of cases within the family and in the population and no selection biases other than as a result of sampling by probands), then in principle one could use the population rates to derive an alternative estimate of the baseline rates. The difficulty comes in adjusting the population rates for covariates, because their distributions in the population may be different from those in the cohort. If the disease is rare and we are prepared to assume that the covariate distribution in the nonaffected members of the family is representative of the population, then we can approximately estimate the baseline rates as

$$\lambda_0(t) = \bar{\lambda}(t) / \mathscr{E}[\exp(\underline{z}_{ij}'\underline{\beta}) \mid D_{ij} = 0]$$

where $\bar{\lambda}(t)$ is the population rate. An important advantage of the case-control design is that such assumptions are unnecessary as the controls essentially provide an estimator of the baseline rates.

In the typical case-control study, cases of the disease of interest are ascertained and matched with one or more controls drawn from the population at risk. For the purpose of this discussion, the sample of controls might be drawn from population lists, by canvassing the neighborhood of each case, by random digit dialing, from lists of the cases' friends, from hospitalized or dead patients with other diseases, but not from relatives of the cases. In the standard approach, a family history covariate, such as the presence or absence of a positive family history, the number of affected family members, or the number affected weighted by their degree of relatedness, is computed for each case and control and used in standard conditional logistic regression analyses. This is inefficient and potentially biased because it does not take into account the number of events expected in each family. For example, if cases tend to have larger families (perhaps through some socioeconomic correlate), then they will be more likely to have positive family histories even if there is no shared genetic or environmental risk. It is natural, therefore, to compute a family history covariate as the difference of observed and expected events. The problem then arises as to how the expected events are to be computed.

One approach would be to treat all the family members other than the sampled cases and controls themselves as a birth cohort and to analyze these data in the same way as just described for the proband design. A simpler approach would be to analyze only the cases and controls using conditional logistic regression, treating the frailty estimate as a known covariate. For this purpose, baseline rates could be estimated either from the cohort of family members or from population rates as previously described, and the cases and controls themselves would not contribute to the frailty estimation (otherwise the case families would have higher expected frailties than the control families under the null hypothesis).

To fully exploit the power of the case-control design, one needs to have covariate values, not just for the cases and controls but also for all their family members. This can be quite difficult to obtain, particularly for those who have died. A reasonable approximation may be to impute values for missing covariate information by randomly sampling from the posterior distribution of covariate values given *a*) the covariate value of their matched case or control, *b*) the assumed value of the intra-family correlation of the covariate, *c*) the age-specific population distribution of covariate values, *d*) whether they and their matched case or control were affected or not, and *e*), the current fitted value of the regression coefficient for the covariate. If covariate values are known for some family members, this information will allow the intra-family correlation to be estimated. Detailed examples of such imputation rules are described elsewhere (*9*). For example, suppose for unaffected subjects at age *t*, the z_{ij} were normally distributed with mean μ(t), variance σ^2(t) and correlation ρ. Then for affected subjects, the marginal distribution of z_{ij} would also be normal with mean $\mu(t) + \beta'\sigma^2(t)$ and variance of $\sigma^2(t)$ (*11*). Then conditional on z_{i0}, D_{ij}, and D_{i0}, z_{ij} is normally distributed with mean

$$\mu(t) + \beta\sigma^2(t)D_{ij} + \rho[z_{io} - \mu(t) - \beta\sigma^2(t)D_{i0}]$$

and variance $\sigma^2(t)(1-\rho^2)$, where *D* is an indicator for whether the subject is affected or not.

Simulation Results

In order to assess the feasibility of disentangling genetic and environmental influences using multivariate frailty models, we carried out a number of simulations. In most of these simulations, 25 four-generation cohorts of family members were generated; in relation to a subject in the third generation, the possible relatives would include siblings, parents, grandparents, offspring, aunts/uncles, and nieces/nephews; other family members might also have spouse and in-law relations. Each simulation comprised a total of 418 to 471 members. Two measured environmental covariates (one continuous and one dichotomous) and two multivariate normally distributed frailties (one for genetic and one for environmental influences, with a known covariance structures that were chosen to be as different as possible) were randomly assigned to each family member. Exponentially distributed failure times, conditional on the measured and unmeasured covariates, were generated for the cause of interest as well as for competing causes. Parameter values were adjusted to produce from 27 to 47 cases. Each simulation was analyzed using the method of Self and Prentice including various combinations of genetic and environmental frailty estimates. Because of the amount of computing required for the fitting, it was not feasible to replicate the simulations, so we are unable to describe the test size or power, but the consistency of our findings across simulations suggests that the results are unlikely to be due to chance.

Table 2 summarizes the results of a portion of the simulations we conducted. In simulations 1 through 4, a highly significant familial aggregation was detected, but in none were we able to fit both a genetic and an environmental component simultaneously. Indeed, each time we tried, the estimate of the environmental vari-

Table 2. Simulation results.

	Simulated frailty variances		LR chi-square tests (df)[a]	
Simulation	Genetic	Environmental	Genetic variance	Environmental variance
1	5	1	20.92 (1)	14.70 (1)
2	3	1	24.54 (1)	16.88 (1)
3	1	3	20.50 (1)	17.50 (1)
4	0	3	3.72 (1)	7.06 (1)
5	3	1	7.44 (1)	5.00 (1)
6	(See Table 3 for results)			
	Covariate effect (Wald's)	Genetic variance (Wald's)		
7	43.39 (1)	—		
	—	2.66 (1)		
	45.06 (1)	0.67 (1)		

[a] Likelihood ratio (LR) tests for frailty components fitting each component separately, adjusted only for measured covariates.

ance was negative and the likelihood for the genetic-only model was always slightly larger than for the environment-only model (even for simulation 3 in which the true environmental variance was larger than the genetic variance).

Simulations 2 and 5 are identical except that simulation 2 includes both first- and second-degree relatives (25 families, 469 total subjects, 46 cases), whereas simulation 5 is restricted to first-degree relatives (50 families, 418 total subjects, 34 cases). This was done to assess the relative value of trying to obtain larger pedigrees versus a larger number of small pedigrees. Intuitively, one would imagine that the relative informativeness of the two designs would be approximately proportional to the number of cases from multiple-case families. In simulation 2, these numbered 40, whereas in simulation 5 they numbered 24. Furthermore, the average number of cases in families with more than one case was 4.1 in simulation 2 but only 2.25 in simulation 5. The pattern of likelihood ratio statistics was very similar between the two simulations, but the chi-square values were about 3.3 times larger for simulation 2. Given that the two simulations had roughly the same total number of subjects, we would conclude that the larger pedigrees were more informative per subject; even scaling the statistics by the number of cases, we would also conclude that the larger pedigrees were more informative per case (presumably owing to the larger number of affected family members per case). However, this conclusion is based on the assumption that the quality of the data on second degree relatives is as good as that of the first degree relatives, which is unlikely to be true in practice.

Simulation 2 was also used to compare the results of the frailty analysis with what might be expected, using simpler family history covariates in standard methods of analysis (Table 3). Somewhat to our surprise, fitting the simulated model using frailty methods did not produce more significant results than a simple binary covariate (presence or absence of other affected family members). Also treating this binary covariate, or the number of affected family members, or the observed minus expected number as fixed covariates (i.e., using all times but excluding the index subject) consistently produced larger chi squares than treating the same covariates as time-dependent (i.e., using only events prior to the current time, but including the index subject); to assess whether this might be because of some liberality in the procedure under the null hypothesis, simulation 6 was generated with the true frailty variance being zero. All chi squares in this case were trivial.

Table 3. Fixed versus time-dependent family history covariates.

	Likelihood ratio, fixed (df)	Likelihood ratio, time-dependent (df)
Simulation 2		
Binary	27.04 (1)	18.42 (1)
Number observed	20.20 (1)	13.02 (1)
Observed − expected	20.98 (1)	14.16 (1)
Simulation 6 (null case)		
Binary	0.44 (1)	0.20 (1)
Number observed	0.62 (1)	0.23 (1)
Observed − expected	0.02 (1)	0.002 (1)

Simulation 7 addressed the case of a measured covariate being intermediate on a causal path from unmeasured genotype to disease (e.g., hormones as mediators of a genetic effect for breast cancer). As expected, the addition of the measured covariate considerably reduced the estimate of the genetic variance, but addition of the genetic frailty did not affect the measured covariate effect.

Finally, in simulation 8 we considered a design that ought to be optimal for separating genetic and environmental influences: monozygotic (MZ) and dizygotic (DZ) twins reared together and apart. Even in this case, the two frailty components could not be fitted simultaneously.

Applications

Swedish Cohort Study of Breast Cancer in Twins

The details of this application of the Self and Prentice model for univariate frailty are described elsewhere (*12*). Basically, a cohort of 11581 female twin pairs was assembled from the Swedish registry of twin births from 1886 to 1958, consisting of all those for whom both members responded to a questionnaire in 1961 (if born before 1925) or 1971 (if born after 1925) and for whom zygosity could be determined. This was then linked with the Swedish cancer registry that was created in 1958 to identify cancer cases and with the national death registry to assess vital status (*13*).

To illustrate the methods, we constructed a sample of the cohort consisting of all disease concordant pairs, a random 10% sample of discordant pairs, and a random 1% sample of nondiseased pairs. Because both the observed and expected number of disease concordant pairs are overestimated by the same factor, these overestimates will cancel in computing relative risks. Results of these analyses are presented in Table 4.

In the absence of measured covariates, a significant frailty variance was found with an estimate of 1.37 (SE = 0.75). This estimate was reduced only slightly by

Table 4. Swedish twin frailty analysis.

Parameter	Estimate	SE	LR chi-square (df)
Variance of frailty distribution (γ)	1.37	0.75	4.95 (1)
N − E (all twins)	0.70	0.26	6.59 (1)
(MZ twins)	0.98	0.39	
N − E			7.64 (2)
(DZ twins)	0.55	0.31	
N − E *zyg[a]	1.02	0.35	7.61 (1)

[a] Zygosity coded 1 for monozygosity (MZ); ½ for dizygosity (DZ).

adjustment for birth cohort, cigarette smoking, and relative weight. (It would have been desirable to have more relevant covariates for breast cancer, but the study was not designed with this disease as its primary focus, and the relevant questions were not asked.)

The approximate frailty covariate, observed minus expected cases, produced an estimate of 0.70 for all twins, 0.98 for MZ and 0.55 for DZ twins. The best fit was obtained by taking observed minus expected cases weighted by 1 for MZ and ½ for DZ twins. This can be seen as an approximation to the bivariate genetic frailty model.

There was a significant interaction between attained age and this genetic frailty covariate (χ^2_1 = 5.44), such that the frailty effect was stronger at younger ages. This is consistent with the suggestion that the genetic effect is strongest for premenopausal breast cancer.

Case-Control Study of Adenocarcinoma of the Lung and Familial Smoking

A population-based case-control study of adenocarcinoma in Los Angeles females was done to assess risk factors, including personal and passive smoking and family history. Details of the study design and the major findings can be found in the publication by Wu et al. (*14*). In particular, a highly significant effect of a family history of lung cancer was found, even after adjusting for personal smoking and other risk factors. In our analysis, we sought to determine whether some of this familial relative risk could be explained by correlation of family members' smoking habits.

For the analyses of passive smoking effects, each case and control was asked questions about the smoking habits of her parents, siblings, spouse, and other cohabitants. We also knew which of the subjects' first-degree relatives had had lung cancer and if so, whether or not they smoked. Finally, we knew how many brothers and sisters the subject had. Because of the design of the questionnaire, however, we did not know the lifetime smoking histories for the subjects' parents (only their status during the subjects' childhood and at diagnosis if they had lung cancer) nor which of the sibs smoked. Using the information that we did have on each family, we therefore tried to impute values for the unknown smoking histories to arrive at a random decision as to whether each family member smoked and if so, his age at starting and quitting and average number of cigarettes per day. This imputation applied the age-specific distributions of variables for cases and controls to affected and unaffected subjects, respectively, in the spirit of the section "Modifications for Proband and Case-Control Designs." The various decision rules are described elsewhere (*9*).

The analysis is based on the conditional likelihood for the cases and their matched controls, taking as a family history covariate the expectation of the frailty given the lifetime covariate, disease, and censoring histories of the family members. (We have not attempted a cohort-style analysis because of the large size of the resulting cohort.) In calculating this frailty estimate, the baseline hazards were initially estimated from a lifetable analysis of the cohort of family members (excluding the index cases and controls) adjusted only for measured covariates, and then a single cycle of the two step process that was described in the second section was done.

Table 5. Lung cancer frailty analysis.

Parameter	Estimate	SE	LR chi-square (df)
Initial unadjusted frailty variance	2.49	1.05	22.66 (1)
Frailty variance adjusted for personal smoking: Baseline rates estimated by average cohort rates; E_i not incorporating smoking covariate	1.99	0.94	14.38 (1)
Frailty variance adjusted for personal smoking: Baseline rates adjusted for smoking; E_i incorporating smoking covariate	1.59	0.78	11.82 (1)

The frailty variance was initially estimated at 2.49 (LR χ^2_1 = 22.66) with no smoking effects in the model (Table 5). Addition of personal smoking reduced this estimate to 1.99 (LR χ^2_1 = 14.38); to obtain this estimate, the average rates for the entire cohort were used as $\lambda_0(t)$ and the smoking covariate was not used in estimating the E_i terms in the frailty. In the next iteration, the smoking-adjusted baseline rates and family members' smoking habits were used to obtain smoking-adjusted $\hat{E}_i$ and $\hat{\epsilon}_i$, the resulting variance estimated reduced to 1.59 on the first iteration, but was still highly significant (LR χ^2_1 = 11.82). Thus, we would conclude that the familial aggregation of lung cancer was only partially explained by familial aggregation of smoking. Although this conclusion can only be tentative in view of the probable high degree of misclassification of family members' smoking habits, we designed the imputation rules in such a way as to maximize the smoking × lung cancer association, thereby giving familial smoking the largest possible opportunity for explaining the association.

Discussion

The methods we have described provide a means of analyzing survival data for families, taking into account their interrelationships and any measured covariates. The latter could include environmental exposures, genetic markers, or variables on a causal pathway from genotype to outcome (such as hormones or reproductive events in breast cancer). Thus, they appear to address the major limitations of classical genetic and epidemiologic methods, as enumerated at the beginning. Numerous details remain to be resolved, however, including the development of a tractable variance estimator, the identifiability of the multivariate models, and the validity of the proposals for applications to noncohort designs. Although we have developed a feasible program for the univariate frailty model (but not the correct variance estimator), it is highly computer-intensive and the proposed extensions to multivariate frailty models and the IP algorithm are likely to be even more so. The simulations suggest that simple approximations may perform quite well. Thus, the development of practical procedures and the study of the power to distinguish various alternatives remain high priorities.

The incorporation of genetic markers into such an analysis will be addressed in a separate paper. Such markers m_{ij} could simply be included as measured covariates in the models we have described above. However, an approach that would be more in the spirit of linkage analysis (*15*) would assume that m is informative about ϵ_{i} and that conditional on ϵ_{ij}, D_{ij} is independent of m. Assuming a logistic dependence of m_{ij} on ϵ_{ij}, the contribution of subject ij to the likelihood would be of the form

$$\int\Pi_{ij}\, P(D_{ij} \mid E_{ij},\underline{\mathrm{z}}_{ij},\epsilon_{ij})\, P(m_{ij} \mid \epsilon_{ij})\, P(\underline{\epsilon}_i)\, d\underline{\epsilon}_i\ .$$

This work was supported in part by National Cancer Institute grant numbers CA42949 and CA14089. We are grateful to David Clayton for many helpful suggestions and to Birgitta Floderus and Anna Wu for making available the data that were used for illustration.

Appendix 1

Formulation of the IP Algorithm for Frailty Models

For univariate gamma frailty models, one would proceed as follows:

Step 0. One could obtain initial estimates of $\underline{\lambda}_0$ at $\underline{\beta} = \gamma = 0$ by a simple Kaplan-Meier survival analysis on the entire cohort, or better, by applying a Cox regression analysis to obtain the maximum likelihood estimate (MLE) of $\underline{\beta}$ and $\underline{\lambda}_0$ at $\gamma = 0$. Better yet, one could use the EM-algorithm approach described in the text to obtain the MLE of $\underline{\beta}$, $\underline{\lambda}_0$, and γ.

Step 1. For each family, one would randomly sample a single value of the frailty ϵ_i from the posterior distribution of $e^{\epsilon_i} \mid F_i\ \underline{\beta},\ \gamma,\ \underline{\lambda}_0$ for the current values of these parameters. Assuming the prior distribution of e^{ϵ_i} is gamma with shape parameter $1/\gamma$ and scale parameter γ, then the posterior distribution of e^{ϵ_i} is also gamma with parameters $D_i + (1/\gamma)$ and $[E_i + (1/\gamma)]^{-1}$. Thus, it suffices to randomly sample frailties from this gamma distribution.

Step 2a. Now treating the $\hat{\epsilon}_i$ as known, one randomly samples values of $\lambda_0(t_k) = \lambda_k\delta(0)$ for each failure time t_k from their posterior distributions given F, $\underline{\beta}$ and $\underline{\epsilon}$. Assuming λ_k has a flat prior on $[0,\infty]$, then the probability density function (pdf) for λ_k is given by the likelihood, $\lambda_k\, e^{-\lambda_k \mathrm{S}_k}\, d\lambda_k$ where $S_k = \Sigma_{ij\epsilon R_k} \exp(\mathrm{z}_{ij}'\underline{\beta} + \hat{\epsilon}_i)$. Thus the cumulative distribution function (cdf) is simply $e^{-\lambda_k S_k}$, so it suffices to draw a uniform [0,1] deviate F and compute $\hat{\lambda}_k$ as $-\ln F/S_k$.

Step 2b. Assuming a flat prior on R^K for $\underline{\beta}$, the posterior distribution of $\underline{\beta}$ given $\underline{\epsilon}$ and $\underline{\lambda}_0$ is again given by the likelihood function, which can be written as

$$\ln L(\underline{\beta}) = \underline{\beta}'\Sigma_{ij}\underline{z}_{ij}D_{ij} - \Sigma_{ij}\Lambda_0(t_{ij})\exp(\underline{z}_{ij}'\underline{\beta} + \epsilon_i).$$

This can be expanded in a Taylor series around the current estimate of $\underline{\beta}$ to obtain

$$\ln L(\underline{\beta}) = \ln L(\underline{\hat{\beta}}) + (\underline{\beta} - \underline{\hat{\beta}})'A_1 + (\underline{\beta} - \underline{\hat{\beta}})'\underline{\underline{A}}_2(\underline{\beta} - \underline{\hat{\beta}})$$

where $A_1 = \Sigma\Lambda_0(t_{ij})\,\underline{z}_{ij}\exp(\underline{z}_{ij}'\underline{\hat{\beta}} + \hat{\epsilon}_i)$ and $\underline{\underline{A}}_2 = \Sigma\Lambda_0(t_{ij})$ $\underline{z}_{ij} \text{ ø } \underline{z}_{ij} \exp(\underline{z}_{ij}'\underline{\hat{\beta}} + \hat{\epsilon}_i)$.

Thus, we can draw the next value of $\underline{\beta}$ from a multivariate normal distribution with mean $\underline{\hat{\beta}} - \underline{\underline{A}}_2{}^{-1}A_1$ and covariance matrix $\underline{\underline{\mathrm{A}}}_2{}^{-1}$.

Step 2c. One randomly samples γ from the posterior distribution of

$$\gamma \mid \underline{\epsilon} = \mathrm{e}^{-\mathrm{C}/\gamma} / \gamma^{1/\gamma}\,\Gamma(1/\gamma)$$

where $\mathrm{C} = \Sigma\hat{e}_{\mathrm{i}} - \Sigma\hat{\epsilon}_i$.

Steps 1 and 2 are repeated and the simulated values of $\underline{\hat{\beta}}$ and $\hat{\gamma}$ (after an initial run-in period if the process is started at $\gamma = 0$ rather than the MLE) are tabulated as the joint posterior distribution.

For multivariate frailty models, steps 2*a* and 2*b* are unchanged and steps 1 and 2*c* are revised according to the form of the assumed prior distribution of ϵ_i. For example, for a polygenic inheritance model (omitting the subscripts ij and letting e_M, e_F, e_S, e_{0_m} denote respectively the frailties of the mother, father, spouse, and offspring $m = 1, \ldots, M_{ij}$ of subject ij), then the prior distribution of ϵ_{ij} is $\mathrm{N}[\mu_{ij}, \sigma^2_{ij}]$ where

$$\hat{\mu}_{ij} = [\hat{\epsilon}_M + \hat{\epsilon}_F + \Sigma_m(\hat{\epsilon}_{o_m} - \hat{\epsilon}_S)]/2,$$

and

$$\hat{\sigma}^2{}_{ij} = \frac{\hat{\sigma}^2}{1 + M_{ij}/4}$$

and σ^2 is an unknown parameter assumed to be constant for all subjects. Then the posterior density of ϵ_{ij} is proportional to

$$\exp(\epsilon_{ij}D_{ij} - E_{ij}e^{\epsilon ij})\, N(\epsilon_{ij}\,|\mu_{ij}, \sigma^2_{ij})$$

which can be approximated by a normal density with mean $\hat{\mu}_{ij} + (\mathrm{D}_{ij} - \mathrm{E}_{ij})/\hat{\sigma}^2$ and variance $\hat{\sigma}^2_{ij}/(1 + \hat{\sigma}^2_{ij}E_{ij})$. In step 1, one would therefore simply sample $\hat{\epsilon}_{ij}$ from these approximating normal distributions. In step 2*c*, $\hat{\sigma}^2$ would be sampled from the posterior distribution of σ^2 given the set of residuals $\hat{\epsilon}_{ij} - \hat{\mu}_{ij}$. Assuming a flat prior for $\ln\sigma$, this has density $\Sigma(\hat{\epsilon}_{ij} - \hat{\mu}_{ij})^2\,(1 - \mathrm{M}_{ij}/4)$ $/\,\chi^2_{\text{N-I}}$. Thus, one only computes this sum of squares and divides it by a random chi square deviate.

Appendix 2

Two-Point Model for Frailty among Sibs

Let $\gamma = gg$, gG, and GG denote the possible genotypes, π the prevalence of allele G, and R_γ the relative risk associated with genotype γ (1 of gg or gG and R for GG in a recessive model; 1 for gg and R for gG and GG in a dominant model). We assume a proportional hazards model of the form

$$\lambda(t,\underline{z},\gamma) = \lambda_0(t)e^{\underline{z'\beta}}R_\gamma.$$

For a pair of sibs, the prior probability of their joint genotype is given by Mendelian laws, e.g., $P_{gg,gg} = (1 - \pi)^4$, $P_{gG,gG} = 4(1 - \pi)^2\pi^2$, $P_{gg,GG} = 2(1 - \pi)^2\pi^2$, etc. Then the marginal hazard is given by

$$\lambda(t,\underline{z}) = \lambda_0(t)\, e^{\underline{z\beta}}\, \Sigma_{\gamma 1 \gamma 2} P_{\gamma_1\gamma_2}(\pi)\, R_{\gamma_1} P(D_1 \mid \gamma_1, E_1) \\ \times P(D_2 \mid \gamma_2, E_2)$$

where $P(D|\gamma,E) = (E\gamma)^D e^{-E\gamma}$ and $E = \Lambda_0(t)e^{\underline{z'\beta}}$.

REFERENCES

1. Self, S. G., and Prentice, R. L. Incorporating random effects into multivariate relative risk regression models. In: Modern Statistical Methods in Chronic Disease Epidemiology (S. H. Moolgavkar and R. L. Prentice, Eds.), John Wiley and Sons, New York, 1986, pp. 167–177.
2. Clayton, D., and Cuzick, J. Multivariate generalizations of the proportional hazards model. J. R. Stat. Soc. Ser. A 148: 82–117 (1985).
3. Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B 39: 1–38 (1977).
4. Tanner, M. A., and Wong, W. H. The calculation of posterior distributions by data augmentation. J. Am. Stat. Soc. 82: 528–550 (1987).
5. Bonney, G. E. Regressive logistic models for familial disease and other binary traits. Biometrics 42: 611–625 (1986).
6. Peto, J. Genetic predisposition to cancer. In: Banbury Report 4: Cancer Incidence in Defined Populations (J. Cairns, J. Lyon, and M. Skolnick, Eds.), Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1980, pp. 203–213.
7. Khoury, M. J., Beaty, T. H, and Liang, K. Can familial aggregation of disease be explained by familial aggregation of environmental risk factors? Am. J. Epidemiol. 127: 674–683 (1988).
8. Breslow, N. E. Contribution to the discussion of paper by D. R. Cox. J. Roy. Stat. Soc. Ser. B 34: 216–217 (1972).
9. Mack, W. Frailty models for the study of familial aggregation of disease. Unpublished Ph.D. dissertation, University of Southern California, Los Angeles, CA, 1988.
10. Levitan, M., and Montagu, A. Textbook of Human Genetics. Oxford University Press, New York, 1971, pp. 419–422.
11. McKeown-Eyssen, G. E., and Thomas D. C. Sample size determination for case-control studies: the influence of the distribution of exposure. J. Chron. Dis. 38: 559–568 (1985).
12. Thomas, D. C., Langholz, B., Mack, W., and Floderus, B. Bivariate survival models for analysis of genetic and environmental effects in twins. Genet. Epidemiol. 7: in press.
13. Hrubec, Z., Floderus-Myrhed, B., de Faire, U., and Sarna, S. Familial factors in mortality with control of epidemiological covariables: Swedish twins born 1886–1925. Acta Genet. Med. Gemellol. 33: 403–412 (1984).
14. Wu, A. H., Yu, M. C., Thomas, D. C., Pike, M. C., and Henderson, B. E. Personal and family history of lung disease as risk factors for adenocarcinoma of the lung. Cancer Res. 48: 7279–7289 (1988).
15. Bonney, G. E., Lathrop, G. M., and Lalouel, J. M. Combined linkage and segregation analysis using regressive models. Am. J. Hum. Genet. 43: 29–37 (1988).