Biostatistics (2009), **10**, 1, *pp.* 187–200
doi:10.1093/biostatistics/kxn027
Advance Access publication on August 19, 2008

Gamma frailty model for linkage analysis with application to interval-censored migraine data

M. A. JONKER*, S. BHULAI

Department of Mathematics, Faculty of Sciences,
Vrije Universiteit, De Boelelaan 1081 a, 1081 HV Amsterdam, The Netherlands
majonker@few.vu.nl

D. I. BOOMSMA, R. S. L. LIGTHART, D. POSTHUMA

Department of Biological Psychology, Vrije Universiteit, Van der Boechorststraat 1,
1081 BT Amsterdam, The Netherlands

A. W. VAN DER VAART

Department of Mathematics, Faculty of Sciences, Vrije Universiteit, De Boelelaan 1081 a,
1081 HV Amsterdam, The Netherlands

Summary

For many diseases, it seems that the age at onset is genetically influenced. Therefore, the age-at-onset data are often collected in order to map the disease gene(s). The ages are often (right) censored or truncated, and therefore, many standard techniques for linkage analysis cannot be used. In this paper, we present a correlated frailty model for censored survival data of siblings. The model is used for testing heritability for the age at onset and linkage between the loci and the gene(s) that influence(s) the survival time. The model is applied to interval-censored migraine twin data. Heritability (obtained from the frailties rather than actual onset times) was estimated as 0.42; this value was highly significant. The highest lod score, a score of 1.9, was found at the end of chromosome 19.

Keywords: Additive gamma frailty model; Heritability; IBD sharing; Interval censoring; Likelihood ratio test; Linkage analysis; Migraine; Survival data.

1. Introduction

Studies on several diseases show strong correlation of ages at onset between family members, for example, breast cancer (Claus *and others*, 1991) and Alzheimer disease (Meyer *and others*, 1998). The results of these studies suggest that not only the occurrence but also the age at onset of the disease is genetically influenced. Therefore, information on age at onset is often collected to map the disease gene(s) and the gene(s) that influence(s) the age at onset of the disease. The exact ages are often censored. This makes techniques that are often used for mapping genes for complex quantitative traits (Haseman–Elston (1972)

[*] To whom correspondence should be addressed.

regression and variance decomposition models [e.g. Sham, 1998]) inapplicable for censored survival data. Statistical methods for gene search which combine techniques of survival analysis and methods of quantitative genetics are needed.

In survival analysis, the emphasis is on modeling the hazard function of independent individuals. The analysis of dependent survival data is more difficult because the dependence structure has to be modeled too. Frailty models for survival data typically include a single shared frailty to model simple patterns of dependence between survival data of related individuals (Vaupel *and others*, 1979). In these models, groups of individuals share the same frailty. In statistical genetics, a group usually consists of family members; individuals who share a part of their segregating genes. The shared frailty model is inappropriate to model this kind of data because different family relationships should correspond to different frailty correlations. Shared frailty models have been extended to correlated frailty models to deal with more complex dependence structures between individuals (Yashin *and others*, 1995; Andersen *and others*, 1992) and have been used to model age-of-onset data within families to test linkage between the loci and the gene(s) that influence(s) the survival time (see, e.g. Petersen, 1998; Korsgaard and Andersen, 1998; Yashin and Iachine, 1999; Li, 1999; Zhong and Li, 2002, 2004; Li and Zhong, 2002; Iachine, 2001) .

In this paper, our aim is to test the genetic contribution to the age at which people experience their first migraine attack and to find locations on the chromosomes that show linkage with the genes that influence this age at onset of migraine. The data we use are from a longitudinal study of Dutch twins and their family members. The ages at migraine onset are interval censored. Furthermore, identical-by-descent (IBD) information for 63–284 markers on the autosomes is available for 258 dizygotic twins. First, we want to test heritability of migraine onset, and second, for each of the markers we want to test whether they are linked to age at onset of migraine. In order to test for linkage, we model the migraine data of the siblings with an additive gamma frailty model. The frailty term of each sibling is decomposed as a linear combination of independent gamma-distributed random variables which represent the genetic contribution to the age at onset of migraine due to part of the genome at a marker, genetic contribution due to loci unlinked to the specific marker, and contributions due to shared familial effects and unshared environmental effects. Before testing linkage for all markers, we test for heritability, that is, a genetic contribution to the variability of age at onset of migraine. Usually, heritability is defined as the proportion of variance of the quantitative trait associated with genetic effects. Because age at onset and in particular its variance is difficult to handle, we define the trait in this context as the frailty and hence define heritability as the proportion of variance of the frailty associated with genetic effects. Basing heritability on a latent trait in this way is not unusual and is closely linked to our method to test for linkage at specific markers.

Mainly for mathematical convenience, we assume that the frailty variable follows a gamma distribution. In that case, an explicit expression of the bivariate survival function in terms of the marginal survival function exists. Estimation of the parameters and testing heritability and linkage are considered in the cases that the marginal survival function is completely unknown and that it belongs to a family of parametric distributions. Gamma frailty models for testing linkage have been proposed before (see, e.g. Li, 1999; Li and Zhong, 2002; Zhong and Li, 2002, 2004; Yashin *and others*, 1999). The model we propose can be used for testing both linkage and heritability.

The remainder of the paper is organized as follows. In Section 2, we introduce the additive gamma frailty model for survival data of siblings. Next, we explain how to test heritability and linkage. In Section 3, we apply the frailty model to interval-censored migraine data of Dutch twins. We describe the data, explain how to estimate model parameters, and show the results of our analysis. In Section 4, we give some concluding remarks. Appendices A, B, and C (in the supplementary material available at *Biostatistics* online [http://www.biostatistics.oxfordjournals.org]) contain the derivations of expressions of the bivariate and the trivariate survival functions of age at onset and the results of a simulation study.

2. An additive gamma frailty model for heritability and linkage

2.1 *The model for linkage analysis*

In this section, we describe a frailty model for sib pairs to test linkage between one locus and the gene that influences the age at onset time. We also explain how to extend the model to more siblings and to test linkage for 2 or more loci simultaneously.

We assume that the observations of different sib pairs are independent. For ease of notation, we describe the model for one sib pair. Let (T_1, T_2) be the survival times of the sib pair and (Z_1, Z_2) a pair of latent variables ("frailties") such that T_1 and T_2 are conditionally independent given (Z_1, Z_2) with hazard functions $t \rightarrow Z_1\lambda(t)$ and $t \rightarrow Z_2\lambda(t)$, respectively, for a given "baseline hazard function" λ. Define N_{IBD} as the number of alleles IBD at the specific locus. The IBD number does not contain any information on the marginal distributions of the ages at onset T_1 and T_2 of the sibs: so for all $t \geqslant 0$ and $k = 0, 1, 2$, $P(T_1 > t|N_{\mathrm{IBD}} = k) = P(T_1 > t)$ and similarly for T_2. However, if the specific locus and the gene that influences the age at onset are in proximity on the chromosome, the association of the survival times of the sib pair should increase with the number of their alleles IBD at the specific locus; a sibling pair with 2 alleles IBD should have more similar survival times than a pair with zero alleles IBD at the locus. We model correlation between the frailties Z_1 and Z_2 dependent on N_{IBD}.

We model the frailties Z_1 and Z_2 as a linear combination of independent gamma-distributed random variables. The first 2 variables depend on N_{IBD} and represent the additive genetic contribution to the frailty by the alleles at the specific locus, the third term represents the common environment and sharing alleles unlinked to the specific locus, and the fourth term represents specific environment and non-sharing alleles. This decomposition is explained in more detail in the following.

We assume that the father and the mother of the sib pair are unrelated and that there is no assortative mating. Then, there are 4 unique alleles at the specific locus that are distinct by descent. We label the paternal chromosomes containing the specific locus by (1, 2) and the maternal chromosomes by (3, 4). The inheritance vectors for the 2 children are defined as

$$V_i = (V_{i,1}, V_{i,2}), \quad i = 1, 2,$$

where $V_{i,1}$ equals 1 or 2, $V_{i,2}$ equals 3 or 4, and i runs over the siblings of the sib pair (Li, 1999). The inheritance vectors indicate which alleles at the specific locus are transmitted from the father and the mother to their 2 children. So if $(v_{i,1}, v_{i,2}) = (1, 4)$, sib i inherited the information at the specific locus from chromosomes 1 and 4. The variables U_1, U_2, U_3, and U_4 represent the genetic frailties due to the alleles at the specific locus at the chromosomes of the father, (U_1, U_2), and the mother, (U_3, U_4). Since it is assumed that the father and the mother are unrelated and there is no assortative mating, U_1, U_2, U_3, and U_4 are independent.

We define the additive genetic frailties due to alleles at the specific locus for the father and mother as

$$Z_{\mathrm{F}} = U_1 + U_2,$$

$$Z_{\mathrm{M}} = U_3 + U_4.$$

For the 2 children, we define these additive genetic frailties as

$$\begin{pmatrix} Z_{1,\mathrm{g}} \\ Z_{2,\mathrm{g}} \end{pmatrix} = \begin{pmatrix} U_{V_{1,1}} + U_{V_{1,2}} \\ U_{V_{2,1}} + U_{V_{2,2}} \end{pmatrix}.$$

Taking into account the possible contributions to the disease not due to the specific locus, we add a term U_{C} which represents the frailty for the common environment and sharing alleles at loci that are unlinked to the specific locus and $U_{\mathrm{E},i}$ $(i = 1, 2)$ for the specific environment and non-sharing alleles for the 2 sibs. We assume that $U_1, \dots, U_4, U_{\mathrm{C}}, U_{\mathrm{E},1}$, and $U_{\mathrm{E},2}$ are independent and all have a gamma distribution with inverse scale parameter η and shape parameter ν for U_1, U_2, U_3, and U_4, ν_{c} for U_{C}, and ν_{e} for $U_{\mathrm{E},1}$ and $U_{\mathrm{E},2}$. Then, the pair of frailties equals

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} U_{V_{1,1}} + U_{V_{1,2}} + U_{\mathrm{C}} + U_{\mathrm{E},1} \\ U_{V_{2,1}} + U_{V_{2,2}} + U_{\mathrm{C}} + U_{\mathrm{E},2} \end{pmatrix}.$$

The pair (Z_1, Z_2) has a bivariate Gamma distribution, where Z_1 and Z_2 are both distributed as $\Gamma(2\nu + \nu_{\mathrm{c}} + \nu_{\mathrm{e}}, \eta)$ with mean $(2\nu + \nu_{\mathrm{c}} + \nu_{\mathrm{e}})/\eta$. We set $\eta = 2\nu + \nu_{\mathrm{c}} + \nu_{\mathrm{e}}$ so that the expectation of Z_1 and Z_2 equals 1. Then, the variance of Z_i and the correlation between Z_1 and Z_2 equal

$$\mathrm{Var}\ Z_i = \frac{2\nu + \nu_{\mathrm{c}} + \nu_{\mathrm{e}}}{\eta^2} = \frac{1}{\eta}, \qquad \text{for } i = 1, 2,$$
$$\rho = \mathrm{Cor}\,(Z_1, Z_2) = \frac{\nu}{\eta} + \frac{\nu_{\mathrm{c}}}{\eta}.$$

The variables $U_1, \dots, U_4, U_{\mathrm{C}}, U_{\mathrm{E},1}$, and $U_{\mathrm{E},2}$ are assumed to be independent of N_{IBD}, but the inheritance vectors V_1 and V_2 are dependent on N_{IBD} $\left(N_{\mathrm{IBD}} = 1_{\{V_{1,1}=V_{2,1}\}} + 1_{\{V_{1,2}=V_{2,2}\}}\right)$. Conditional on $N_{\mathrm{IBD}} = k$, the frailty pair, $(Z_1, Z_2)|(N_{\mathrm{IBD}} = k)$, still has a bivariate Gamma distribution with the marginal distributions as described before, but with correlation

$$\rho_k = \mathrm{Cor}\,(Z_1, Z_2 | N_{\mathrm{IBD}} = k) = \eta(k\mathrm{Var}\ U_1 + \mathrm{Var}\ U_{\mathrm{C}}) = \frac{k\nu}{\eta} + \frac{\nu_{\mathrm{c}}}{\eta}$$

(for a derivation of this expression, see Appendix A of the supplementary material available at *Biostatistics* online [http://www.biostatistics.oxfordjournals.org]). The first term of the correlation, $k\nu/\eta$, is positive if an increasing number of alleles IBD at the locus results in an increasing association between the survival times within the sib pairs. The parameter ν equals zero if there is no relation between the number of alleles IBD at the locus and the association between the survival times within the sib pairs. The second term of the correlation, ν_{c}/η, explains the association between the sibs due to common environment and sharing alleles that are unlinked to the locus of interest. In order to detect genes that affect the survival time, we try to find regions on the chromosome where genotypic similarity is highly correlated with similarity of survival times. Genotypic similarity is defined in terms of the IBD number. So for every locus, we test whether an increasing IBD number coincides with a more similar survival time. This means that for every locus, we test whether ν is positive.

Conditional on N_{IBD}, the inheritance vectors are not uniquely determined. For instance, if $N_{\mathrm{IBD}} = 0$, there are 4 combinations of the inheritance vectors possible ($v_1 = (1, 3)$ and $v_2 = (2, 4)$, $v_1 = (1, 4)$ and $v_2 = (2, 3)$, $v_1 = (2, 4)$ and $v_2 = (1, 3)$, or $v_1 = (2, 3)$ and $v_2 = (1, 4)$). However, all the different possibilities give the same marginal distributions of the frailties and the same correlation between Z_1 and Z_2.

The frailties are assumed to be gamma distributed. In that specific case, it is possible to write the bivariate survival function of (T_1, T_2) conditional on N_{IBD} in terms of the marginal survival function:

$$\begin{aligned} S_k(t_1, t_2) &= P(T_1 > t_1, T_2 > t_2 | N_{\mathrm{IBD}} = k) \\ &= \left(\frac{1}{S(t_1)^{-\sigma^2} + S(t_2)^{-\sigma^2} - 1}\right)^{\rho_k/\sigma^2} S(t_1)^{1-\rho_k} S(t_2)^{1-\rho_k}, \end{aligned} \tag{2.1}$$

with S the marginal survival function for the survival times T_1 and T_2 and $\sigma^2 = 1/\eta$ the variance of Z_1 and Z_2 (for a derivation of this expression, see Appendix A of the supplementary material available at

Biostatistics online [http://www.biostatistics.oxfordjournals.org]). Note that $S_k(0,t) = S_k(t,0) = S(t)$ is independent of the number of alleles IBD. If males and females have different marginal survival functions, S is replaced by gender-specific survival functions.

With this model, we try to transform the variance decomposition model for quantitative traits to a model for possibly censored (or truncated) survival data. In the variance decomposition model, the trait value is decomposed into a linear combination of independent normally distributed random variables. In the case of censored or truncated survival times, age at onset of the disease/event may not be observed. As an alternative, we decompose the frailties as described above. So the frailties are viewed as latent phenotypes.

In the decomposition of the frailty, the terms for specific environment and non-sharing alleles, $U_{\mathrm{E},1}$ and $U_{\mathrm{E},2}$, seem to be superfluous because $U_{\mathrm{E},1}$ and $U_{\mathrm{E},2}$ are always independent. However, if the terms are excluded from the model, the correlation between the frailties of sibs with $N_{\mathrm{IBD}} = 2$ is always equal to 1 and the model cannot be used to test heritability (see Section 2.2).

More sibs. The model can easily be generalized to more than 2 sibs. In Appendix B of the supplementary material available at *Biostatistics* online (http://www.biostatistics.oxfordjournals.org), we explain how to derive an expression for the survival function for 3 siblings in terms of the marginal survival function. Here, we give only some special cases.

Let $N_{\mathrm{IBD}} = (N_{\mathrm{IBD},12}, N_{\mathrm{IBD},13}, N_{\mathrm{IBD},23})$, the alleles IBD between the first and the second individual, the first and the third, and between the second and the third individual. If $N_{\mathrm{IBD},12} = 2$, $N_{\mathrm{IBD},13} = 0$, and $N_{\mathrm{IBD},23} = 0$,

$$\begin{aligned} S_{2,0,0}(t_1,t_2,t_3) &= P(T_1 > t_1, T_2 > t_2, T_3 > t_3 | N_{\mathrm{IBD}} = (2,0,0)) \\ &= \left(\frac{1}{S(t_1)^{-\sigma^2} + S(t_2)^{-\sigma^2} - 1}\right)^{2\nu} \\ &\quad \times \left(\frac{1}{S(t_1)^{-\sigma^2} + S(t_2)^{-\sigma^2} + S(t_3)^{-\sigma^2} - 2}\right)^{\nu_{\mathrm{c}}} S(t_1)^{\nu_{\mathrm{e}}\sigma^2} S(t_2)^{\nu_{\mathrm{e}}\sigma^2} S(t_3)^{(2\nu+\nu_{\mathrm{e}})\sigma^2}. \end{aligned} \tag{2.2}$$

If $N_{\mathrm{IBD},12} = 1$, $N_{\mathrm{IBD},13} = 2$, and $N_{\mathrm{IBD},23} = 1$,

$$\begin{aligned} S_{1,2,1}(t_1,t_2,t_3) &= P(T_1 > t_1, T_2 > t_2, T_3 > t_3 | N_{\mathrm{IBD}} = (1,2,1)) \\ &= \left(\frac{1}{S(t_1)^{-\sigma^2} + S(t_2)^{-\sigma^2} + S(t_3)^{-\sigma^2} - 2}\right)^{\nu} \left(\frac{1}{S(t_1)^{-\sigma^2} + S(t_3)^{-\sigma^2} - 1}\right)^{\nu} \\ &\quad \times \left(\frac{1}{S(t_1)^{-\sigma^2} + S(t_2)^{-\sigma^2} + S(t_3)^{-\sigma^2} - 2}\right)^{\nu_{\mathrm{c}}} S(t_1)^{\nu_{\mathrm{e}}\sigma^2} S(t_2)^{(\nu+\nu_{\mathrm{e}})\sigma^2} S(t_3)^{\nu_{\mathrm{e}}\sigma^2}. \end{aligned} \tag{2.3}$$

Multiple loci model. The frailty model can be extended so that linkage can be tested for 2 or more unlinked loci simultaneously. Define the variables U_1, U_2, $\tilde{U}_1$, and $\tilde{U}_2$ as the genetic frailties due to the 2 alleles on the first locus, (U_1, U_2), and the second locus, $(\tilde{U}_1, \tilde{U}_2)$, of the father. The variables for the mother, U_3, U_4, $\tilde{U}_3$, and $\tilde{U}_4$, are defined analogously. Since it is assumed that the father and the mother are unrelated, that there is no assortative mating and that the 2 loci are unlinked, all variables are independent.

When we model 2 unlinked loci, the frailties for the sib pair are decomposed as

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} U_{V_{1,1}} + U_{V_{1,2}} + \tilde{U}_{\tilde{V}_{1,1}} + \tilde{U}_{\tilde{V}_{1,2}} + U_{\mathrm{C}} + U_{\mathrm{E},1} \\ U_{V_{2,1}} + U_{V_{2,2}} + \tilde{U}_{\tilde{V}_{2,1}} + \tilde{U}_{\tilde{V}_{2,2}} + U_{\mathrm{C}} + U_{\mathrm{E},2}, \end{pmatrix}$$

with $(V_{i,1}, V_{i,2})$ and $(\tilde{V}_{i,1}, \tilde{V}_{i,2})$ the inheritance vectors for the ith sibling at locus 1 and 2, respectively. The variables U_C, $U_{E,1}$, and $U_{E,2}$ are defined as before. We assume that $U_1, \ldots, U_4$ and $\tilde{U}_1, \ldots, \tilde{U}_4$ all have a $\Gamma(\nu, \eta)$-distribution. Then, Z_1 and Z_2 have, marginally, a Gamma distribution $\Gamma(4\nu + \nu_c + \nu_e, \eta)$ with $EZ_1 = EZ_2 = (4\nu+\nu_c+\nu_e)/\eta$ and $\text{Var } Z_1 = \text{Var } Z_2 = (4\nu+\nu_c+\nu_e)/\eta^2$, with $4\nu/\eta^2 = 2\nu/\eta^2+2\nu/\eta^2$, the sum of the variances due to the 2 loci. We set $\eta = 4\nu + \nu_c + \nu_e$ so that Z_1 and Z_2 have expectation 1 again.

An expression of the bivariate survival function of (T_1, T_2) conditional on the number of alleles IBD at the 2 loci, $(N_{\text{IBD},1} = k, N_{\text{IBD},2} = l)$, can be derived in the same way. The conditional bivariate survival function, denoted by $S_{k,l}$, has the same form as S_k in (2.1) with ρ_k replaced by the correlation $\rho_{k,l} = \text{Cor}\,(Z_1, Z_2 | N_{\text{IBD},1} = k, N_{\text{IBD},2} = l) = (k\nu + l\nu + \nu_c)/\eta$. When we condition the frailty on the number of alleles IBD at a single locus, we recover the frailty for the single locus model. However, the frailty found from the multilocus model has a mixture of Gamma distributions rather than a bivariate Gamma distribution as we assumed for the single locus model.

2.2 *Heritability*

To investigate the genetic contribution to a quantitative trait, we reformulate the model. The model holds for siblings (also for monozygotic twins). For a sib pair, we decompose the frailty as $Z_i = A_i + C + E_i$, for $i = 1, 2$, where A_i represents the additive genetics, C the common environment, and E_i the non-shared, specific environmental effects for the ith sib in the pair (see also Yashin *and others*, 1999). Note that this decomposition is very similar to the decomposition for linkage analysis. We assume that (A_1, A_2), C, and E_1 and E_2 are independent and gamma distributed with inverse scale parameter η and shape parameters 2ν for A_1 and A_2, ν_c for C, and parameter ν_e for E_1 and E_2. Furthermore, we take A_1 and A_2 such that the correlation $\text{Cor}\,(A_1, A_2) = 1$ for a monozygotic twin and $\text{Cor}\,(A_1, A_2) = 1/2$ for dizygotic twins and siblings who are not twins (monozygotic twins share all alleles IBD and dizygotic twins and sibs who are no twins on average half of the alleles). Again, we set $\eta = 2\nu + \nu_c + \nu_e$ so that $EZ_1 = EZ_2 = 1$.

The lowercase letters h^2, c^2, and e^2 are defined as the proportions of variance of individual frailty associated with additive genetic effects, shared environmental factors, and non-shared environmental factors, hence $h^2 = 2\nu/\eta$, $c^2 = \nu_c/\eta$, and $e^2 = \nu_e/\eta$. Then, $h^2 + c^2 + e^2 = 1$. For these decompositions, the correlations between monozygotic twins and dizygotic twins and between sibs who are not twins are

$$\rho_{\text{MZ}} = \frac{2\nu}{\eta} + \frac{\nu_c}{\eta} = h^2 + c^2,$$

$$\rho_{\text{DZ}} = \frac{1}{2}\frac{2\nu}{\eta} + \frac{\nu_c}{\eta} = \frac{1}{2}h^2 + c^2. \tag{2.4}$$

We define heritability as h^2, estimate it by maximum likelihood, and test the hypothesis H_0: $h^2 = 0$ using a likelihood ratio test.

2.3 *Testing*

To test whether there is a genetic effect on survival time, we test the hypothesis H_0: $h^2 = 0$ versus H_1: $h^2 > 0$ or equivalently H_0: $\nu = 0$ versus H_0: $\nu \neq 0$. If the null hypothesis is rejected, one or more genes influence the survival times. To find the locations of these genes on the chromosomes, we test for all loci in the data set whether they are linked to one of these genes; we test H_0: $\nu = 0$ against H_1: $\nu > 0$.

The likelihood depends on the unknown marginal survival function S. If we assume that S belongs to a family of parametric survival functions, the limit distribution of the likelihood ratio statistic is a

50–50 mixture of a point mass at zero and a chi-square distribution with one degree of freedom (see, e.g. van der Vaart, 1998, Chapter 16). The family of parametric distributions can, for instance, be chosen by considering the non-parametric maximum likelihood estimator (NPMLE) based on only survival data.

As an alternative, we use a common estimator for the nuisance parameter S for both the numerator and the denominator of the likelihood ratio statistic and maximize the likelihoods with respect to the remaining parameters. One way to estimate S is by maximizing the likelihood for the survival data of independent individuals. Generally, inserting an estimator for S into the likelihood ratio statistic will destroy the asymptotic mixture distribution of the statistic. If we estimate S separately, the data of individuals who have not been genotyped can also be used, and thus a considerably larger sample might be available to estimate S than is used to construct the test statistic. Then, the asymptotic distribution of the likelihood ratio statistic should not change much. This expectation was borne out in a simulation study (the data were simulated from the estimated model found in Section 3). In Appendix C of the supplementary material available at *Biostatistics* online (http://www.biostatistics.oxfordjournals.org), this simulation study is described. We also verified it by heuristic theoretical arguments (available on request), which show that estimating a nuisance parameter, based on additional independent observations, will always increase the variability of the likelihood ratio statistic. In our situation, it will asymptotically behave under the null hypothesis as a 50–50 mixture of a point mass at 0 and the distribution of $(1 + \mu\lambda)$ times a chi-square variable with one degree of freedom. Here, λ is the limiting value of the quotient n/m of numbers n and m of observations used to construct the test statistic and to estimate S, respectively, and μ measures the relative informativeness of the 2 types of observations to estimate the nuisance parameter. In our application (see Section 3), the factor $(1+\mu\lambda)$ is close to 1, since m is considerably larger than n, so that the asymptotic distribution is changed only little. Estimating S separately from the other parameters has the advantage that it reduces the computing time when testing linkage because S does not have to be estimated for all loci in the data set.

3. Application to interval-censored survival data

Interval censoring occurs, for example, when individuals are followed up only at fixed intervals, for example, by annual surveys. Let (B_1, C_1) and (B_2, C_2) be observation times for the 2 individuals of a sib pair with $B_j \leqslant C_j$ almost surely for $j = 1, 2$. We observe

$$(B_j, C_j), \quad \Delta_j := 1_{\{T_j \leqslant B_j\}}, \quad \Sigma_j := 1_{\{B_j < T_j \leqslant C_j\}}, \quad \mathrm{MD}_j,$$

for $j = 1, 2$ (j indicates the sib in the pair), where $1_{\{\cdot\}}$ denotes the indicator function and MD stands for “marker data,” which are used to determine IBD numbers. Furthermore, the indicator function $1_{\{T \leqslant B\}}$ equals 1 if $T \leqslant B$ and 0 otherwise.

We assume that the observations of different sib pairs are independent. In the description of the data, we again restrict ourselves to one sib pair. We assume that the survival times (T_1, T_2) of a sib pair are independent of the observation times (B_1, C_1, B_2, C_2) and that the vector (B_1, C_1, B_2, C_2) has an unknown density $f_{(B_1,C_1,B_2,C_2)}$.

In order to derive an expression for the likelihood, we assume that the data at a marker and the survival data are conditionally independent given the numbers of alleles IBD at the particular marker. Then,

$$\mathrm{P}(T_1 \leqslant t_1, T_2 \leqslant t_2|\mathrm{MD}) = \sum_{k=0}^{2} P(T_1 \leqslant t_1, T_2 \leqslant t_2|N_{\mathrm{IBD}} = k)P(N_{\mathrm{IBD}} = k|\mathrm{MD}).$$

The likelihood for one related pair and one specific marker is proportional to

$$
\begin{aligned}
&\mathrm{lik}((B_1, C_1, B_2, C_2, \Delta_1, \Delta_2, \Sigma_1, \Sigma_2, \mathrm{MD}); \nu, \nu_{\mathrm{c}}, \nu_{\mathrm{e}}, \sigma, S)\\
&\quad = \sum_{k=0}^{2} \{(1 - S_k(0, B_2) - S_k(B_1, 0) + S_k(B_1, B_2))^{\Delta_1 \Delta_2}\\
&\qquad + (S_k(0, B_2) - S_k(0, C_2) - S_k(B_1, B_2) + S_k(B_1, C_2))^{\Delta_1 \Sigma_2}\\
&\qquad + (S_k(0, C_2) - S_k(B_1, C_2))^{\Delta_1 (1-\Delta_2)(1-\Sigma_2)}\\
&\qquad + (S_k(B_1, 0) - S_k(B_1, B_2) - S_k(C_1, 0) + S_k(C_1, B_2))^{\Sigma_1 \Delta_2}\\
&\qquad + (S_k(B_1, B_2) - S_k(B_1, C_2) - S_k(C_1, B_2) + S_k(C_1, C_2))^{\Sigma_1 \Sigma_2}\\
&\qquad + (S_k(B_1, C_2) - S_k(C_1, C_2))^{\Sigma_1 (1-\Delta_2)(1-\Sigma_2)}\\
&\qquad + (S_k(C_1, 0) - S_k(C_1, B_2))^{(1-\Delta_1)(1-\Sigma_1)\Delta_2}\\
&\qquad + (S_k(C_1, B_2) - S_k(C_1, C_2))^{(1-\Delta_1)(1-\Sigma_1)\Sigma_2}\\
&\qquad + S_k(C_1, C_2)^{(1-\Delta_1)(1-\Sigma_1)(1-\Delta_2)(1-\Sigma_2)} \mathrm{Prob}(N_{\mathrm{IBD}} = k|\mathrm{MD})\}.
\end{aligned}
\tag{3.5}
$$

An expression in terms of the marginal survival function S is found after inserting the expression given in (2.1). Because the observations of different sib pairs are assumed to be independent, the likelihood for all pairs is simply the product of the term in the previous display for the n sib pairs.

In Section 2, we discussed the possibility of estimating the marginal survival function S beforehand. For interval-censored survival data of m independent individuals, the likelihood is given by

$$\prod_{i=1}^{m} (1 - S(B_i))^{\Delta_i} (S(B_i) - S(C_i))^{\Sigma_i} S(C_i)^{1-\Delta_i-\Sigma_i}, \tag{3.6}$$

with S the unknown marginal survival function and B_i, C_i, Σ_i, and Δ_i as defined before. If S is assumed to be completely unknown, we can estimate the survival function by the NPMLE (see, e.g. Groeneboom and Wellner, 1992). Otherwise, if S is a parametric distribution, the unknown parameter(s) of the survival function can be estimated by their maximum likelihood estimators.

3.1 *Application to interval-censored migraine data*

Migraine is a highly prevalent disorder characterized by recurrent attacks of headaches, which are typically unilateral and have a pulsating quality. The headache is accompanied by a variety of symptoms such as nausea or vomiting and an increased sensitivity to light and sound (photophobia and phonophobia). Due to a lack of biological markers, migraine diagnosis relies mainly on symptomatology; a certain number and combination of symptoms should be present in order to meet the commonly used diagnostic criteria for migraine (Headache Classification Subcommittee of the International Headache Society, 2004). Individuals who have several of the migraine symptoms, but not the right combinations or not enough symptoms, are assigned as no-migraine patient. This makes it difficult to measure phenotypic similarity.

The data. The analyses were performed on longitudinal migraine data collected in a large sample of Dutch twins and their families. The participants were volunteer members of the Netherlands Twin Registry, kept by the Department of Biological Psychology at the Vrije Universiteit in Amsterdam (Boomsma *and others*, 2002, 2006). The data were collected between 1991 and 2002, as part of an ongoing study of health, lifestyle, and personality. Surveys were mailed to the participants at 6 different time points. In each of these surveys, questions on headache and migraine were included. In surveys 2, 3, and 4 (1993, 1995, and 1997, respectively), participants were asked if they had ever been diagnosed with migraine by a physician. In surveys 1, 5, and 6 (1991, 2000, and 2002, respectively), the participants answered a series of more detailed questions concerning headache symptoms. Based on these questions, subjects were classified as affected or unaffected. Since a complete migraine diagnosis confirmed by a neurologist was not available, the phenotype will be referred to as "migrainous headache."

We have marker data at all 22 chromosome pairs (not the sex-chromosome). The number of markers per chromosome varies between 63 (chromosome 22) and 284 (chromosome 1). The conditional probabilities $P(N_{\text{IBD}} = k|\text{MD})$ for $k = 0, 1, 2$ were computed with the software package Merlin (Abecasis *and others*, 2002). We have IBD data of 258 dizygotic twin pairs. Heritability was estimated and tested based on 3975 monozygotic and dizygotic twin pairs (also data from not-genotyped individuals were used).

The analysis. The marginal survival function for the age at the first migrainous headache attack may differ between males and females. We therefore model gender-specific survival functions: S_{M} for males and S_{F} for females. The formulas in Section 2 remain valid, but S has to be replaced by S_{M} or S_{F}, depending on the gender of the individual. As mentioned before, we consider 2 models. For both models, we estimated heritability and tested whether this was significantly larger than 0. Next, we computed lod scores for all loci in the data set, that is, log to base 10 of the likelihood ratio.

In the first model, we assume that S_{M} and S_{F} follow shifted exponential distributions with unknown shift and intensity parameter. This choice was based on the form of the NPMLE of S_{M} and S_{F}. The unknown model parameters (including the shifts and intensities) were estimated by maximizing likelihood based on twin data.

In the second model, the survival functions S_{F} and S_{M} are completely unknown; in which case, we have a semiparametric frailty model. We estimated S_{M} and S_{F} by maximizing the likelihood in (3.6) based on interval-censored migraine data of all monozygotic and dizygotic twins and their siblings. We inserted these estimates in the likelihood function and assumed them known for further analysis. The asymptotic distribution of the likelihood ratio statistic is close to a 50–50 mixture of a point mass at 0 and a chi-square distribution with one degree of freedom, as argued in Section 2.

For both models, we used a grid search to maximize the likelihood. The expectation-maximization algorithm is difficult to use because the individual likelihoods are summed over the number of alleles IBD at the specific location at a chromosome.

Results for the parametric model. In the parametric model, we assume that the survival functions S_{M} and S_{F} are equal to shifted exponential distributions. The shift and intensity parameters for males and females are unknown. When estimating heritability or lod scores, the likelihood is maximized with respect to these parameters simultaneously with the other parameters in the model. Since we estimated the shift and intensity parameters for every locus separately, we found more than 1000 estimates (the number of loci). Fortunately, for most loci, the estimates were exactly the same. For the males, we found the estimates 12.4 and 0.0093 for the shift and intensity parameters, and for the females, these estimates were equal to 10.0 and 0.0220. The corresponding survival functions are shown in Figure 1.

For estimating heritability, data of monozygotic and dizygotic twins were used. In total, we used data of 3975 twin pairs. Heritability (based on the frailties) was estimated as 0.42 with a 95% confidence interval [0.374; 0.461], slightly wider than in the semiparametric case. We computed lod scores and found the most prominent peak at the end of chromosome 19, with a height of 1.86. A diagram of the lod scores found for chromosome 19 is given in Figure 2. The lod score of 1.86 corresponds to a p-value of 0.0017. Since many tests are performed, a multiple testing correction has to be made. In practice, the value 3

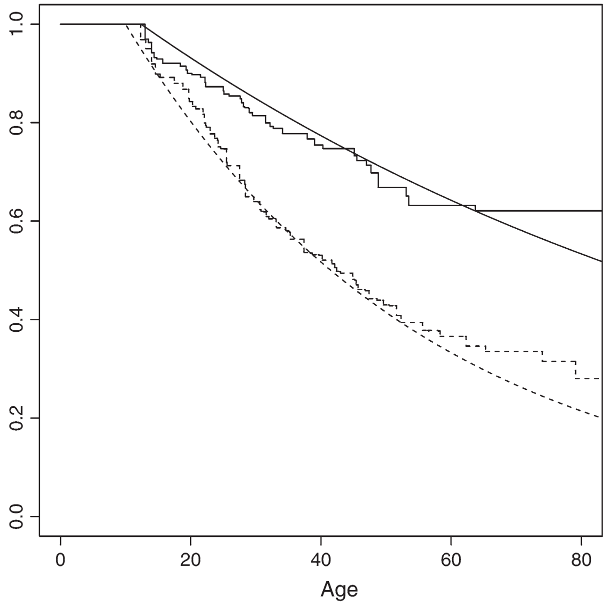

Fig. 1. The NPMLE of the survival functions S_{M} and S_{F} (step functions) and the estimated survival functions in the parametric model (smooth curves). Males, solid curves; females, dashed curves. The NPMLEs are based on interval-censored survival data of monozygotic and dizygotic twins and the sibs in our data set. The estimated survival functions in the parametric model re based on data of 258 dizygotic twins of whom estimated IBD numbers are available.

is often taken as a threshold for significant lod scores, in which case none of the lod scores would be significant.

Results for the semiparametric model. Based on interval-censored migrainous headache data (phenotypes only) of all monozygotic and dizygotic twins and all sibs, we estimated the NPMLE of age at first migrainous headache attack for males and females separately (based on 4791 males and 6796 females). The NPMLEs are shown in Figure 1. Note that the estimates in the parametric and the semiparametric model are quite close (although the data used are different). It is estimated that 72% of the Dutch females will eventually have migrainous headaches at least once. For the Dutch males, this percentage is 38%. We estimated heritability (based on the frailties), h^2, as 0.37 with 95% confidence interval [0.323; 0.415]. For all loci in the data set, we computed the lod scores, and the most prominent peak was again found at the end of chromosome 19 with a height of 1.36 (see Figure 2) and a p-value of 0.0062. None of the lod scores were greater than 3.

The curves of the lod scores computed in the parametric and in the semiparametric model are very similar for most chromosomes.

4. DISCUSSION AND SUMMARY

In this paper, we presented an additive gamma frailty model for analyzing interval-censored migraine data for siblings. The aim was to test whether the age at which people experience their first migraine attack is heritable and, if so, to find locations at the chromosomes that are linked to the migraine genes. Heritability was estimated and tested based on almost 4000 twin pairs. It was estimated as 0.42 in the

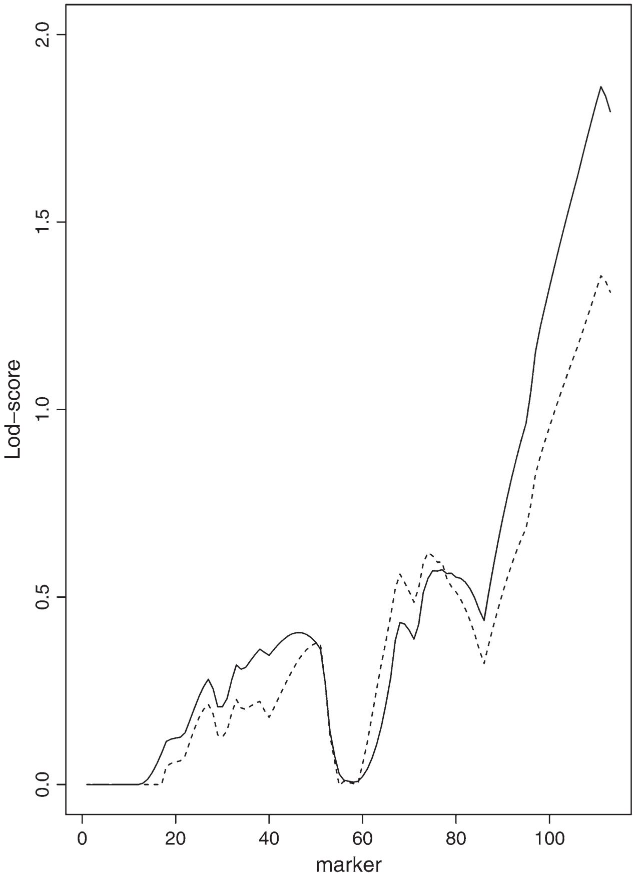

Fig. 2. Lod scores for testing linkage for the markers at chromosome 19 for the parametric model (solid curve) and the semiparametric model (dashed curve).

parametric model and 0.37 in the semiparametric model, in both cases with associated p-values less than 0.001. We defined heritability as the proportion of variance of the frailty associated with genetic effects since age-at-onset data is not available. Estimates of heritability based on actual onset times may differ from our estimates. Linkage analysis was only based on the genotyped dizygotic twin pairs; 258 pairs in total. The highest lod score we found was 1.86 at the end of chromosome 19, less than the threshold of 3. More twins will be genotyped in the near future to increase the sensitivity of the analysis.

With frailty models, dependent data are modeled such that the dependence structure between the relatives (in our case, siblings) does not effect the marginal survival function of the age at onset. Although the model was presented in the context of sibship data, it can be easily extended to medium-sized families. The genetic component of the frailty will remain the same, but chromosomes of more than 4 different individuals play a role (see, e.g. Korsgaard and Andersen, 1998). The term for the common environment should be adapted to the kinship between the 2 individuals; “the common environment” between sibs may be different from the common environment between, for instance, father and son. The last term, the specific environment and non-sharing alleles, is specific for all individuals and can be left unchanged. However, for larger sibships, even for a sibship with 3 siblings, the expression of the survival function and therefore also of the likelihood is complicated. The application of this model is therefore limited to small pedigrees or even to siblings. Variance component models can accommodate any type of pedigree, but the assumption of normal survival times that is often made is dubious. Recently, Diao and Lin (2006) generalized the variance component models. They presented a model in which the distribution function of the survival times is assumed to be unknown and is also applicable in practice for extended pedigrees. The unknown parameters in the model are estimated by the maximum likelihood estimators, and linkage is tested with the likelihood ratio test. Although the model can handle any kind of censoring in theory, the expression of the likelihood might be too complicated for current status and interval-censored data as used in this paper. Diao and Lin (2006) implemented their method in a computer program for right- and left-censored data.

Our aim is to find locations linked to genes that affect the age at the first migraine attack. The analyses were performed on data from a Dutch twin cohort. In practice for linkage analysis, families are often ascertained through their phenotypes; for instance by having children with the disease of interest. In that case, the likelihood should be adapted in order to get estimators and a test statistic that are free of ascertainment bias. In the literature, 3 different methods are proposed. The first method is to consider the retrospective likelihood; that is, the probability of the observed marker data conditional on the phenotypes (see, e.g. Li and Zhong, 2002). The likelihood ratio statistic based on the retrospective likelihood is valid for any ascertainment scheme. The second method is to use the conditional likelihood function conditional on the ascertainment event; for instance the event "at least 2 siblings are affected before a certain age t." In this case, we have to assume that the n sibships collected are a random sample of all sibships that satisfy the ascertainment condition (see Sun and Li, 2004). The third method is to use an ascertainment-adjusted maximum likelihood approach as described in Sun and Li (2004). This is only possible if the sampling scheme of families is clearly defined and followed. In case only families with at least 2 sibs being affected are of interest, information of families with at least 2 siblings (affected or not) are collected. Marker data of only the families with at least 2 affected siblings is gathered. The ascertainment-adjusted maximum likelihood is defined as the probability of the phenotypes of the siblings of all families (so also with zero or one affected sibling) and the observed marker data of the families with at least 2 affected siblings.

For mathematical convenience, we assumed that the frailty is gamma distributed. Only if the frailty has a gamma distribution is it possible to give an explicit expression of the bivariate survival function in terms of the marginals. The frailty is decomposed as a sum of 4 independent gamma-distributed frailty components with a common scale parameter so that the frailty itself is again gamma distributed.

We have considered gender-specific survival functions in the analysis of interval-censored migraine data. In general, this approach works well for a discrete covariate without many categories. However, it is not applicable in case of continuous covariates or discrete covariates with many categories. Then, the likelihood function cannot be written in terms of the marginal survival function but should be written in terms of the hazard function. Linkage can still be tested with the likelihood ratio statistic, but now the likelihood function has to be maximized with respect to the hazard function or an estimator for the hazard function must be inserted into the likelihood ratio statistic (see, e.g. Zhong and Li, 2002; Li and Zhong, 2002).

The score test provides an alternative to the likelihood ratio test which is less model dependent and therefore more robust against misspecification of the model. Moreover, the score test has the advantage that the score statistic has to be computed under the null hypothesis only. However, defining the score statistic requires some work; especially in the case of interval censoring (like the migrainous headache case) and if the IBD numbers are not known exactly. The latter case requires a summation within the logarithm in the log-likelihood (see (3.5)). Zhong and Li (2004) propose a joint proportional hazards model for linkage and association due to linkage disequilibrium. The hazard function of developing disease is defined as a product of a baseline hazard, a frailty term, and a term for the genetic association. The frailty term was defined in a similar way as the frailty term in our model, except that they left out the term for specific environment. For this model, not only (censored) survival data and IBD numbers but also data on genotypes at the marker locus must be available.

Acknowledgments

We would like to thank Piet Groeneboom for using his computer program for estimating the NPMLE for the survival curve based on interval-censored survival data. We thank the referees for some insightful comments and suggestions. We also like to thank all Dutch twins and their families for participation in the study. *Conflict of Interest:* None declared.

Funding

Nonlinear Dynamics of Natural Systems and Netherlands Organization for Scientific Research.

References

Abecasis, G., Cherny, S., Cookson, W. and Cordon, L. (2002). Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* **30**, 97–101.

Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1992). *Statistical Models Based on Counting Processes*. New York: Springer.

Boomsma, D. I., Geus, E. J. C., de Vink, J. M., Stubbe, J. H. and Hottenga, M. A. D. J. J. (2006). Netherlands twin register: from twins to twin families. *Twin Research and Human Genetics* **9**, 849–857.

Boomsma, D. I., Vink, J. M., van Beijsterveldt, T. C. E. M., de Geus, E. J. C., Beem, A. L., Mulder, E. J. C. M., Derks, E. M., Riese, H., Willemsen, G. A. H. M., Bartels, M. *and others* (2002). Netherlands twin register: a focus on longitudinal research. *Twin Research* **5**, 401–406.

Claus, E. B., Risch, N. J. and Thompson, W. D. (1991). Genetic analysis of breast cancer in the cancer and steroid hormone study. *American Journal of Human Genetics* **48**, 232–242.

Diao, G. and Lin, D. Y. (2006). Semiparametric variance-component models for linkage and association analyses of censored trait data. *Genetic Epidemiology* **30**, 570–581.

Groeneboom, P. and Wellner, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Basel, Switzerland: Birkhäuser.

Haseman, J. K. and Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics* **2**, 3–19.

Headache Classification Subcommittee of the International Headache Society (2004). The international classification of headache disorders (2nd edition). *Cephalalgia* **24** (Suppl 1), 9–160.

Iachine, I. (2001). The use of twin and family survival data in the population studies of aging: statistical methods based on multivariate survival models, [PhD. Thesis]. Volume 8. Odense, Denmark: Department of Statistics and Demography, University of Southern Denmark.

Korsgaard, I. R. and Andersen, A. H. (1998). The additive genetic gamma frailty model. *Scandinavian Journal of Statistics* **25**, 255–269.

Li, H. (1999). The additive genetic gamma frailty model for linkage analysis of age-of-onset variation. *Annals of Human Genetics* **63**, 455–468.

Li, H. and Zhong, X. (2002). Multivariate survival models induced by genetic frailties, with application to linkage analysis. *Biostatistics* **3**, 57–75.

Meyer, M. R., Tschanz, J. T., Norton, M. C. and Welsh-Bohmer, K. A. (1998). Apoe genotype predicts when–not whether–one is predisposed to develop Alzheimer disease. *Nature Genetics* **19**, 321–322.

Petersen, J. H. (1998). An additive frailty model for correlated life times. *Biometrics* **54**, 646–661.

Sham, P. C. (1998). *Statistics in Human Genetics*. London: Arnold Publishers.

SUN, W. AND LI, H. (2004). Ascertainment-adjusted maximum likelihood estimation for the additive genetic gamma frailty model. *Lifetime Data Analysis* **10**, 229–245.

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

VAUPEL, J. W., MANTON, K. G. AND STALLARD, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16**, 439–454.

YASHIN, A. I., BEGUN, A. Z. AND IACHINE, I. A. (1999). Genetic factors in susceptibility to death: a comparative analysis of bivariate survival models. *Journal of Epidemiology and Biostatistics* **4**, 53–60.

YASHIN, A. I. AND IACHINE, I. A. (1999). Dependent hazards in multivariate survival problems. *Journal of Multivariate Analysis* **71**, 241–261.

YASHIN, A. I., VAUPEL, J. W. AND IACHINE, I. A. (1995). Correlated individual frailty: an advantageous approach to survival analysis of bivariate data. *Mathematical Population Studies* **5**, 145–159.

ZHONG, X. AND LI, H. (2002). An additive genetic gamma frailty model for two-locus linkage analysis using sibship age of onset data. *Statistical Applications in Genetics and Molecular Biology* **11**, Article 2.

ZHONG, X. AND LI, H. (2004). Score test of genetic association in the presence of linkage based on the additive genetic gamma frailty model. *Biostatistics* **5**, 307–327.

[*Received April 20, 2007; first revision January 11, 2008; second revision April 10, 2008; third revision July 21, 2008; accepted for publication July 22, 2008*]