Journal of Human Genetics (2013) 58, 694–699
© 2013 The Japan Society of Human Genetics All rights reserved 1434-5161/13
www.nature.com/jhg

ORIGINAL ARTICLE

Joint analysis of bivariate competing risks survival times and genetic markers data

Alexander Begun

Bivariate survival models with discretely distributed frailty based on the major gene concept and applied to the data on related individuals such as twins and sibs can be used to estimate the underlying hazard, the relative risk and the frequency of the longevity allele. To determine the position of the longevity gene, additional genetic markers data are needed. If the action of the longevity allele does not depend on its position in the genome, these two problems can be solved separately using a two-step procedure. We proposed an extension of this method allowing us to search the position of two longevity genes at a chromosome using the bivariate survival data with correlated competing risks combined with genetic markers data. We have studied the properties of the model with two longevity genes located on the same and on different chromosomes using simulated data sets.

Journal of Human Genetics (2013) **58,** 694–699; doi:10.1038/jhg.2013.80; published online 1 August 2013

Keywords: bivariate survival analysis; competing risks; genetic markers; linkage analysis

INTRODUCTION

Longevity studies show lifespan correlation between related individuals such as twins, sibs and family members. These correlations can be caused by common environmental and genetic factors. Some genes responsible for longevity have already been discovered, but they explain only a small part of genetic variation. Genes associated with longevity can be identified using the information on genotype frequencies for two or more age groups. A significant age trend of these frequencies can indicate a gene-longevity association. The 'gene frequency' method is based on this idea.[1,2] This basic method can be extended using demographic information about a studied population to allow the estimation of initial frequencies, relative risks and the age trajectories of mortality for candidate genes.[3,4]

Bi- and multivariate correlated frailty models allow us to search for the association between genes and disease or mortality in the absence of observed covariates. The simplest way to find this association is to compare correlations for monozygotic and dizygotic twins.[5] More complex models can involve the decomposition of the frailty into additive genetic factors, genetic dominance factors, shared environmental factors and non-shared environmental factors.[6,7]

In addition to the frailty models with continuously distributed frailty, the major genes models with discretely distributed frailty and mixed models with discretely and continuously distributed frailty can also be considered.[8] The models with discretely distributed frailty have an important advantage, as they can be easily adapted to the family and pedigree data.[9]

If, in addition to longevity or morbidity data, the genetic markers data are also available, we can locate the position of the longevity or disease genes at the chromosome. Li and Zhong[10] proposed a retrospective likelihood approach based on the allele-sharing test for the genetic linkage analysis using sibship data. Jonker *et al.*[11] used an extension of this method to test for linkage and heritability. A weighted nonparametric linkage statistic has been proposed by Callegaro *et al.*[12] to test for linkage for the selected samples. All these approaches use the correlated frailty model, where the frailty is broken down into the sum of the linkage effect and a shared residual effect. The components of the frailty must be calculated for each locus separately by maximizing the retrospective likelihood. This can lead to an enormous increase in calculating time if several genetic markers are involved in these calculations.

If the markers are in linkage disequilibrium, we can use them as covariates in the Cox-like regression and estimate respective regression coefficients. Significant deviation of some coefficients from zero can indicate that respective markers are involved in survival or disease. Even if the genetic markers are in linkage equilibrium, we can determine the position of longevity or disease genes at the chromosome using the linkage analysis.[13] Under the assumption that location of the longevity or frailty gene does not influence survival, this approach involves the two-step procedure. In the first step, we estimate the parameters of the underlying hazard functions and the parameters of frailty distribution for the model with the major gene by maximizing the observed survival data likelihood. The second step is focused on determining the position of the longevity gene between observed markers.

In this paper, we propose an extension of this two-step method—a bivariate correlated competing risks survival model with two major

Institute of Biometrics and Epidemiology, German Diabetes Center at the Heinrich-Heine-University, Düsseldorf, Germany
Correspondence: A Begun, Institute of Biometrics and Epidemiology, German Diabetes Center at the Heinrich-Heine-University, Auf'm Hennekamp 65, Düsseldorf D-40225, Germany.
E-mail: alexander.begun@ddz.uni-duesseldorf.de

Received 31 May 2013; revised 5 July 2013; accepted 7 July 2013; published online 1 August 2013

genes and genetic markers data. The properties of this model are illustrated using examples based on simulated data.

MATERIALS AND METHODS

Survival analysis

Suppose that K_i related individuals belong to cluster i, $i=1,...,n$, and individuals from different clusters are statistically independent. Subjects from the same cluster can belong to a sibship or to a family. For subject k from cluster i, we denote the time to first failure or the censoring time, and the vector of time-independent covariates by t_{ik} and $\boldsymbol{u}_{ik}$, respectively. Let $l_{ik} \in (1,..., L)$, be the type of the first observed failure and $l_{ik}=0$ stand for right censoring. The censoring is denoted by $\delta_{ik}=I(l_{ik}\neq 0)$, where $I(x)=1$ if $x=$'True' and $I(x)=0$, otherwise. We define the cause-specific hazard function for subject k from cluster i by using the formula:

$$\begin{aligned}\lambda_{ikj}(t \mid \boldsymbol{u}_{ik}, Z_{ikj}) &= \lim_{\Delta t \to 0+} P(t \leqslant T+\Delta t, l_{ik}=j \mid T \geqslant t, \boldsymbol{u}_{ik}, Z_{ikj})/\Delta t \\ &= \lambda_{0j}(t) Z_{ikj} \exp(\boldsymbol{\beta}_j' \boldsymbol{u}_{ik})\end{aligned}$$

where $j=1,...,L$, Z_{ikj} is an individual frailty for the failure type j, λ_{0j} is an underlying cause-specific hazard, $\boldsymbol{\beta}_j$ are cause-specific regression coefficients' vectors and the symbol '′' stands for transposition.[14]

If only one of the failure types can occur, the full hazard function for the same subject is defined by using the formula:

$$\begin{aligned}\lambda_{ik}(t \mid \boldsymbol{u}_{ik}, Z_{ik1}, ..., Z_{ikL}) &= \lim_{\Delta t \to 0+} P(t \leq T < t+\Delta t \mid T \\ &\geq t, \boldsymbol{u}_{ik}, Z_{ik1}, ..., Z_{ikL})/\Delta t \\ &= \sum_{j=1}^{L} \lambda_{0j}(t) Z_{ikj} \exp(\boldsymbol{\beta}_j' \boldsymbol{u}_{ik})\end{aligned}$$

suppose that individual frailties Z_{ikj} can correlate for subjects from the same cluster and different types of failure. Dependency between subjects from a cluster can be caused by correlated genotypes for relatives. Complex diseases can be influenced by many genes and environmental factors and the same genes and factors can be involved in different diseases (onsets). This leads to correlations between types of failure (types of onsets or causes of death).

Example (two competing risks of death influenced by two major genes)

Assume that two causes of death are influenced by two unobserved major genes, and the data on the age at death or the age at censoring for each of the twins in the sample consisting of n twin pairs are available. Let $\lambda_{01}(t)$ and $\lambda_{02}(t)$ be the underlying hazards for the first and the second cause of death, respectively. Assume that two longevity alleles with dominant action a and b have frequencies p_a and p_b, respectively, and that they are located in different loci at the same or different chromosome. The neutral alleles in these loci are denoted by A and B with frequencies $1–p_a$ and $1–p_a$, respectively. Suppose that the presence of at least one longevity allele a in genotype decreases the risk of the type 1 failure by factor r_1, $r_1<1$, and the risk of the type 2 failure by factor q_2, $q_2<1$. Similarly assume that the presence of at least one longevity allele b in the genotype decreases the risk of the type 2 failure by factor r_2, $r_2<1$, and the risk of the type 1 failure by factor q_1, $q<1$. Suppose that the absence of the longevity allele corresponds to frailties $Z_{ik1}=Z_{ik2}=1$. If both longevity genes are in Hardy–Weinberg and linkage equilibrium and the action of the longevity genes does not depend on the location at the chromosome, then the possible longevity genotypes have the frequencies and frailties given in Table 1.

In this table notation $(aa+aA)\times BB$; for example, means that a subject has one of the possible genotypes from the set $(aa\times BB, aA\times BB)$. We assume that in the first locus of the genotype $xX\times yY$, the allele x is inherited from the mother and allele X from the father. Similarly in the second locus of this genotype, the allele y is inherited from the mother and allele Y from the father. Suppose that the parents are genetically independent and that their offspring inherit their genotypes independently. The frequencies of an offspring's genotypes depending on the parental genotypes can be calculated under the assumption that an offspring receives with equal probability one of the two alleles from its mother's genotype and likewise one of the two alleles from its father's genotype (the law of segregation).[15]

Table 1 Frailties and genotype frequencies. Both longevity genes are dominant

Z_{ik1}	Z_{ik2}	*Genotype*	*Frequency*
r_1q_1	r_2q_2	$(aa+aA)\times(bb+bB)$	$(1–(1–p_a)^2)(1–(1–p_b)^2)$
r_1	q_2	$(aa+aA)\times BB$	$(1–(1–p_a)^2)(1–p_b)^2$
q_1	r_2	$AA\times(bb+bB)$	$(1–p_a)^2(1–(1–p_b)^2)$
1	1	$AA\times BB$	$(1–p_a)^2(1–p_b)^2$

We denote the observed data in cluster i, $i=1, ..., n$, by $(\boldsymbol{X}_i, \boldsymbol{U}_i, \boldsymbol{L}_i, \boldsymbol{\delta}_i)$, where $\boldsymbol{X}_i=(X_{ik})$, $k=1,...,K_i$, is the time to the first failure or time to censoring vector for subjects in cluster i, $\boldsymbol{U}_i=(\boldsymbol{u}_{ik})$ is the set of vectors of observed covariates, $\boldsymbol{L}_i=(l_{ik})$ is the vector of the types of failure and $\boldsymbol{\delta}_i=(\delta_{ik})$ is the vector of event indicators. Let us assume that given the observed covariates $\boldsymbol{U}$ and frailties Z_{ikj}, the censoring times are independent of the failure times and do not correlate with frailties, frailties are independent of covariates, and covariates' effect is subject specific.[14] If the underlying cause-specific hazard functions are known up to the vector parameter ω, we can write the likelihood function in the form of the following:

$$\begin{aligned}&L(\boldsymbol{X}, \boldsymbol{L} \mid \boldsymbol{U}, \boldsymbol{\delta}; \boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\zeta}) \\ &\qquad = \prod_{i=1}^{n} E_Z\left(\prod_{k=1}^{K_i} \exp\left(-\sum_{j=1}^{L} Z_{ikj} e^{\boldsymbol{\beta}_j' \boldsymbol{u}_{ik}} H_{0j}(X_{ik} \mid \boldsymbol{\omega})\right)\right. \\ &\qquad \left. \times \left(\lambda_{0L_{ik}}(X_{ik} \mid \boldsymbol{\omega}) Z_{ikL_{ik}} e^{\boldsymbol{\beta}_{L_{ik}}' \boldsymbol{u}_{ik}}\right)^{\delta_{ik}}\right)\end{aligned}$$

here $H_{0j}(x \mid \boldsymbol{\omega}) = \int_0^x \lambda_{0j}(t \mid \boldsymbol{\omega})dt$ are depending on $\boldsymbol{\omega}$, the cause-specific cumulative hazard functions, $j=1, ..., L$ and E_Z is the expectation with respect to frailty Z. Unknown vector parameter $\boldsymbol{\zeta}=(p_a, p_b, r_1, q_1, r_2, q_2)$ characterizes the frailty distribution. In the case when clusters are dizygotic twin pairs and longevity is regulated by two dominant genes, we can rewrite the last formula as follows:

$$\begin{aligned}&L_{DZ}(\boldsymbol{X}, \boldsymbol{L} \mid \boldsymbol{U}, \boldsymbol{\delta};\ \boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\zeta}) = \prod_{i=1}^{n_{DZ}} \sum_{G_m, G_f} P(G_m \mid \zeta) P(G_f \mid \zeta) \\ &\times \left(\sum_G P(G \mid G_m, G_f) \exp\left(-\sum_{j=1}^{2} Z_{i1j}(G, \boldsymbol{\zeta}) e^{\boldsymbol{\beta}_j' \boldsymbol{u}_{i1}} H_{0j}(X_{i1} \mid \boldsymbol{\omega})\right)\right. \\ &\left. \times \left(\lambda_{0L_{i1}}(X_{i1} \mid \boldsymbol{\omega}) Z_{i1L_{i1}}(G, \boldsymbol{\zeta}) e^{\boldsymbol{\beta}_{L_{i1}}' \boldsymbol{u}_{i1}}\right)^{\delta_{i1}}\right) \\ &\times \left(\sum_G P(G_m, G_f) \exp\left(-\sum_{j=1}^{2} Z_{i2j}(G, \boldsymbol{\zeta}) e^{\boldsymbol{\beta}_j' \boldsymbol{u}_{i2}} H_{0j}(X_{i2} \mid \boldsymbol{\omega})\right)\right. \\ &\left. \times \left(\lambda_{0L_{i2}}(X_{i2} \mid \boldsymbol{\omega}) Z_{i2L_{i2}}(G, \boldsymbol{\zeta}) e^{\boldsymbol{\beta}_{L_{i2}}' \boldsymbol{u}_{i2}}\right)^{\delta_{i2}}\right)\end{aligned}$$

Here, G_m and G_f are the maternal and paternal genotypes, respectively. The frequencies of these genotypes $P(G_m|\boldsymbol{\zeta})$ and $P(G_f|\boldsymbol{\zeta})$ and the twin frailties $Z_{i1j}(G)$ and $Z_{i2j}(G)$, $j=1,...,2$ are given in Table 1. The segregation ratios of the mating types $P(G|G_m,G_f)$ (the proportions of the different genotypes in the offspring of all mating types) can be calculated using the law of segregation. In this formula, the second with third lines and the fourth with fifth lines stand for expected survivals and probability densities given parental genotypes for the first and the second twin, respectively. We assumed here that both the twins inherited their genes from their parents independently. In the first line, we take the average for all the possible parental genotypes.

The monozygotic twins have identical genotypes and the likelihood function has a form

$$\begin{aligned} L_{MZ}(\boldsymbol{X},\boldsymbol{L}\mid \boldsymbol{U},\boldsymbol{\delta};\boldsymbol{\beta},\boldsymbol{\omega},\boldsymbol{\zeta}) &= \prod_{i=1}^{n_{MZ}}\sum_{G_m,G_f}P(G_m\mid\zeta)P(G_f\mid\zeta)\\ &\times\sum_G P(G\mid G_m,G_f)\exp\left(-\sum_{j=1}^{2}Z_{i1j}(G,\zeta)e^{\boldsymbol{\beta}_j'\boldsymbol{u}_{i1}}H_{0j}(X_{i1}\mid\boldsymbol{\omega})\right)\\ &\times\left(\lambda_{0L_{i1}}(X_{i1}\mid\boldsymbol{\omega})Z_{i1L_{i1}}(G,\zeta)e^{\boldsymbol{\beta}_{L_{i1}}'\boldsymbol{u}_{i1}}\right)^{\delta_{i1}}\\ &\times\exp\left(-\sum_{j=1}^{2}Z_{i1j}(G,\zeta)e^{\boldsymbol{\beta}_j'\boldsymbol{u}_{i2}}H_{0j}(X_{i2}\mid\boldsymbol{\omega})\right)\\ &\times\left(\lambda_{0L_{i2}}(X_{i2}\mid\boldsymbol{\omega})Z_{i1L_{i2}}(G,\zeta)e^{\boldsymbol{\beta}_{L_{i2}}'\boldsymbol{u}_{i2}}\right)^{\delta_{i2}} \end{aligned}$$

The full likelihood is as follows:

$$\begin{aligned} L(\boldsymbol{X},\boldsymbol{L}\mid\boldsymbol{U},\boldsymbol{\delta};\boldsymbol{\beta},\boldsymbol{\omega},\boldsymbol{\zeta}) =& L_{MZ}(\boldsymbol{X},\boldsymbol{L}\mid\boldsymbol{U},\boldsymbol{\delta};\boldsymbol{\beta},\boldsymbol{\omega},\boldsymbol{\zeta})\\ &\times L_{DZ}(\boldsymbol{X},\boldsymbol{L}\mid\boldsymbol{U},\boldsymbol{\delta};\boldsymbol{\beta},\boldsymbol{\omega},\boldsymbol{\zeta}) \end{aligned}$$

We can directly find the estimate $(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\omega}},\hat{\boldsymbol{\zeta}})$ of the unknown vector parameters ($\boldsymbol{\beta}$, $\boldsymbol{\omega}$, $\boldsymbol{\zeta}$) maximizing the likelihood function $L(\boldsymbol{X}, \boldsymbol{L}|\boldsymbol{U}, \boldsymbol{\delta}; \boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\zeta})$ with respect to ($\boldsymbol{\beta}$, $\boldsymbol{\omega}$, $\boldsymbol{\zeta}$). Using the likelihood ratio, we can test different hypothesis about the parameters of the model.

If the underlying cause-specific hazard functions are not parametrically defined, the nonparametric maximum likelihood estimators can be used to estimate the vector parameter $\boldsymbol{\beta}$ and the cause-specific cumulative hazard functions using the EM algorithm.[14,16]

Linkage analysis

After the unknown parameters ($\boldsymbol{\beta}$, $\boldsymbol{\omega}$, $\boldsymbol{\zeta}$) have been estimated, we can try to locate the positions of longevity genes in the genome. Both monozygotic and dizygotic twin pairs contribute to the likelihood $L(\boldsymbol{X}, \boldsymbol{L}|\boldsymbol{U}, \boldsymbol{\delta}; \boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\zeta})$. On the contrary, only the dizygotic twin pairs contribute to the likelihood function for determining the position of longevity genes. The search for linkage between two loci is based on calculating the Lod score, $\log_{10}[L(\theta)/L(\theta_0)]$, proposed by Barnard[17] and used later by Morton[18] in sequential test procedures of the null hypothesis $\theta_0 = 0.5$ versus an alternative value $\theta < 0.5$. Here θ is the probability of recombination between two loci. Lander and Green[19] proposed an algorithm based on the hidden Markov chain concept to calculate the multipoint pedigree likelihood. This algorithm was later modified and optimized by Kruglyak and Lander,[20] Kruglyak *et al.*[21] The hidden Markov chain algorithm is used to calculate the probability of the extended vector of markers (M_s^E), $s = 1, ..., S+N_g$ (this vector consists of S observed markers M and N_g non-observed major genes) taking into account the Markov property of a pair (M_s^E, $\boldsymbol{V}_s$), where ($\boldsymbol{V}_s$) are inheritance vectors. In experiments with twins, an inheritance vector $\boldsymbol{V}_s = (V_s^1, V_s^2, V_s^3, V_s^4)$' is a binary vector at each locus s, $s = 1, ..., S+N_g$, having four components. The first and the second components stand for alleles inherited in this locus by a twin, and the third and the fourth ones characterize alleles inherited by its co-twin. It is assumed that the first and the third components denote alleles inherited from the mother (0 if from the grandmother and 1 if from the grandfather). The second and the fourth components stand for alleles inherited from the father (same rules).

Given the location of the major genes at the chromosome, we calculate the Lod score for dizygotic twin pairs using the value

$$\begin{aligned} L_{DZ}(\boldsymbol{X},\boldsymbol{L},\boldsymbol{M}\mid\boldsymbol{U},\boldsymbol{\delta};\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\omega}},\hat{\zeta},\theta) &= \prod_{i=1}^{n_{DZ}}\sum_{G_m,G_f}P(G_m\mid\hat{\zeta})P(G_f\mid\hat{\zeta})\\ &\times\left(\sum_G P(G,\boldsymbol{M}\mid G_m,G_f;\theta)\exp\left(-\sum_{j=1}^{2}Z_{i1j}(G,\hat{\zeta})e^{\hat{\boldsymbol{\beta}}_j'\boldsymbol{u}_{i1}}H_{0j}(X_{i1}\mid\hat{\boldsymbol{\omega}})\right)\right.\\ &\times\left.\left(\lambda_{0L_{i1}}(X_{i1}\mid\hat{\boldsymbol{\omega}})Z_{i1L_{i1}}(G,\hat{\zeta})e^{\hat{\boldsymbol{\beta}}_{L_{i1}}'\boldsymbol{u}_{i1}}\right)^{\delta_{i1}}\right)\\ &\times\left(\sum_G P(G,\boldsymbol{M}\mid G_m,G_f;\theta)\exp\left(-\sum_{j=1}^{2}Z_{i2j}(G,\hat{\zeta})e^{\hat{\boldsymbol{\beta}}_j'\boldsymbol{u}_{i2}}H_{0j}(X_{i2}\mid\hat{\boldsymbol{\omega}})\right)\right.\\ &\times\left.\left(\lambda_{0L_{i2}}(X_{i2}\mid\hat{\boldsymbol{\omega}})Z_{i2L_{i2}}(G,\hat{\zeta})e^{\hat{\boldsymbol{\beta}}_{L_{i2}}'\boldsymbol{u}_{i2}}\right)^{\delta_{i2}}\right) \end{aligned}$$

Here (G,$\boldsymbol{M}$) is an extended genotype of a twin including genetic markers and longevity genes. A vector parameter characterizes the location of longevity genes in the genome (for example, the recombination distance from respective neighboring markers). The details about the calculation of the extended segregation ratios $P(G, \boldsymbol{M}|G_m, G_f;\theta)$ for twins can be found elsewhere.[13] Finally we calculate the Lod score in the form

$$LodScore = \log_{10}(L_{DZ}(\boldsymbol{X},\boldsymbol{L},\boldsymbol{M}\mid\boldsymbol{U},\boldsymbol{\delta};\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\omega}},\hat{\boldsymbol{\zeta}},\theta)/L_{DZ}(\boldsymbol{X},\boldsymbol{L},\boldsymbol{M}\mid\boldsymbol{U},\boldsymbol{\delta};\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\omega}},\hat{\boldsymbol{\zeta}},\theta_0))$$

where θ_0 stands for longevity genes located at the recombination distance of 0.5 from all markers. We position the longevity loci at different places of a chromosome between respective markers, and calculate in each case the Lod score values. This Lod score profile can be used for testing the linkage. In accordance with current standard criterion for declaring linkage, we construct a support interval containing all the points where the Lod score is higher than or equal to 3. To exclude the linkage, we regard the points where the Lod score is less than or equal to -2.

Simulation study

In order to investigate the properties of the model described above, we have generated three types of data sets. In the first data set type, both longevity genes were situated on the same chromosome with markers. In the second type of data set, the first longevity gene was situated on the same chromosome with markers and the second one out of this chromosome. In the third type of generated data sets, both longevity genes were situated out of the chromosome with markers. In all experiments, non-censored survival data for 1000 dizygotic twin pairs were controlled by two correlated competing mortality risks without observed covariates, and the earliest failure time of each twin has been chosen. For univariate cause-specific survival functions $S_j(x)$, $j = 1, 2$, we have used the gamma-Gompertz parameterization

$$\begin{aligned} S_j(x) = \left(1+s_j^2\tilde{H}_j(x)\right)^{-1/s_j^2} &= \sum_{G_m,G_f}P(G_m\mid\zeta)P(G_f\mid\zeta)\\ &\times\left(\sum_G P(G\mid G_m,G_f)\exp\left(-Z_j(G,\zeta)H_{0j}(x)\right)\right) \end{aligned}$$

Here $\tilde{\lambda}_j(x) = \mathrm{d}\tilde{H}_j(x)/\mathrm{d}x = b_je^{c_jx}$, $\tilde{H}_j(30) = 0$, $b_j>0$, $c_j>0$, $s_j^2>0$ are unknown parameters. That is, the cause-specific survivals are equal to the cause-specific survival in a population of individuals which has survived to age 30 years (left truncation at age 30) with underlying hazards $\tilde{\lambda}_j(x)$ and gamma-distributed frailty with mean 1 and variance $s_j^2>0$ at age 30 years. For survival data generation, we put $b_j = 2.5\times10^{-5}$, $c_j = 0.1$, $s_j^2 = 0.01$, $j = 1$, 2. Given the parameters ζ, b_j, c_j, s_j^2, $j = 1$, 2, the cause-specific underlying cumulative hazard functions $H_{0j}(x)$ can be found as follows: In the first step, we calculate $\tilde{H}_j(x)$ and $S_j(x)$ for given the age x using formulas $\tilde{H}_j(x) = (b_j/c_j)(e^{c_jx}-e^{30c_j})$ and $S_j(x) = (1+s_j^2\tilde{H}_j(x))^{-1/s_j^2}$. Then we calculate $P(G_m|\zeta)$, $P(G_f|\zeta)$ and $P(G|G_m, G_f)$ using assumptions of the model and the law of segregation. Finally, we find $H_{0j}(x)$ from the formula: $S_j(x) = \sum_{G_m,G_f}P(G_m\mid\zeta)P(G_f\mid\zeta)\left(\sum_G P(G\mid G_m,G_f)\exp\left(-Z_j(G,\zeta)H_{0j}(x)\right)\right)$ using a simple bisectional procedure. The functions $H_{0j}(x)$ are used to calculate the likelihood and the Lod score values. To generate the survival data, we have set $p_a = p_b = 0.5$ for longevity allele frequencies, $r_2 = q_1 = 0.1$ and either $r_1 = q_2 = 0.1$ or $r_1 = q_2 = 0.05$ for mortality risks. It is not difficult to show that the vector parameter $\zeta = (p_a, p_b, r_1, q_1, r_2, q_2) = (0.5, 0, 5, 0.1, 0.1, 0.1, 0.1)$ in the model with two dominant major genes and two causes of death with equal hazard functions corresponds to the vector of frailties $(Z_1, Z_2, Z_3) = (1.0, 0.55, 0.01)$ with probabilities $(P_1, P_2, P_3) = (0.25, 0.5, 0.25)$ in the model with one major gene and one cause of death. We can compare these values with results obtained in the longevity study of Danish twins using the major gene model [8]. In the case of autosomal locus with multiplicative action of one beneficial allele, it was found that $(\tilde{Z}_1,\tilde{Z}_2,\tilde{Z}_3) = (1.0, 0.37, 0.14)$ with probabilities $(\tilde{P}_1,\tilde{P}_2,\tilde{P}_3) = (0.21, 0.49, 0.30)$. The frailty $Z_3 = 0.01$ in our study is substantially smaller than the value of frailty $\tilde{Z}_3 = 0.14$, but this effect is caused by the assumption about multiplicative action of beneficial alleles.

For genetic data generation, we assumed that 10 genetic markers were uniformly distributed over a chromosome with a distance of 5 cM between neighboring markers. Each gene at the marker locus can be characterized by a pair from the set of 10 different alleles and each allele can be met in the population with a frequency equal to 0.1. In the first type of simulated data, the first longevity gene was situated in the middle between the third and fourth markers and the second longevity gene was situated in the middle between the seventh and eighth markers. In the second type of simulated data, the first longevity gene was situated in the middle between the fifth and sixth markers, whereas the second longevity gene was out of this chromosome. In addition, we have assumed that the observed markers and longevity genes are in linkage equilibrium and in Hardy–Weinberg equilibrium. The peaks for Lod scores were counted to calculate the statistical power and the type I error rate. We used a Lod score threshold of 3 as an indicator of linkage.

RESULTS

In Table 2, the results for parameter estimates b_j, c_j, s_j, $j=1$, 2 and p, p_1, q, q_1, r, r_1 based on 100 simulated data sets are given. Table 2 includes the empirical means and s.d.'s of the estimates. Taking into account s.d.'s, all parameter estimates are in agreement with true values. The Lod score profiles averaged over all simulations are shown in the Figures 1–6. A twofold decrease in the frailties r_1, q_2 can substantially increase the Lod score and the beneficial action of the first longevity gene, if this gene is situated on the same chromosome with markers (see Figure 1). On the contrary, this leads to a decrease of the Lod score if both longevity genes are situated out of the chromosome with markers (Figure 2). In this case, we observe background Lod score profiles without clear peaks. The coefficient of the correlation between life spans of siblings has not increased significantly (0.23 ± 0.03 versus 0.26 ± 0.03) by a twofold decrease in the frailties r_1, q_2. In summary, the decreased action of the longevity gene can improve the chances to reject or accept the hypothesis that the longevity gene is situated on the chromosome with markers. In experiments with two longevity genes situated on the same chromosome with markers, we observe two clear peaks situated symmetrically on the plots (see Figures 3 and 4). As expected, the heights of the peaks are similar, if the actions of both longevity genes are also similar (see Figure 4). In other words, we cannot distinguish between the first and the second longevity genes using the Lod score profile. The stronger the action of the longevity gene, the higher the respective peak in the Lod score profile. If we look at Figure 3, we can conclude that the first and the second longevity genes are most probably located on the first and on the second half of the chromosome with markers, respectively. The background Lod score profiles for both cases are shown in Figures 5 and 6. It seems that the values of the Lod score for these profiles depend only on the value of the distance between the possible positions of the first and the second longevity genes and on the distance to neighboring markers. In the neighborhood of the markers, the values of the Lod score are slightly smaller. We assessed the power of the likelihood ratio test when comparing true hypothesis H_1 (the data set was generated using the model with two dominant major genes and vector parameter $\zeta_1=(p_a, p_b, r_1, q_1, r_2, q_2)$, $p_a=p_b=0.5$, $r_1=q_1=r_2=q_2=0.1$) and false null hypothesis H_0 (the data set was generated using the model with one dominant major gene and vector parameter $\zeta_0=(p_a, r_1, q_2)$, $p_a=0.5$, $r_1=q_2=0.1$). The power in this experiment was equal to 0.92.

If the data set was generated under the assumption of a single major gene and the Lod score profile was calculated using the present method, we will observe a peak situated near the diagonal $\theta_1=\theta_2$. Finally, we have calculated the Lod score profile (averaged over 100 simulations) using the model with a single dominant major gene applied to the data set generated using the model with two dominant major genes (both situated on the same chromosome—in the middle between the third and fourth markers and in the middle between the

Table 2 Summary of simulation results for unknown parameters

	True value	*Mean*	*s.d.*	*True value*	*Mean*	*s.d.*
10^{-5} b_1	2.50	2.30	0.48	2.50	2.28	0.58
10^{-1} c_1	1.00	1.02	0.03	1.00	1.02	0.04
$10^{-1}\cdot s_1$	1.00	1.55	1.46	1.00	1.37	1.43
$10^{-5}\cdot b_2$	2.50	2.28	0.59	2.50	2.29	0.55
$10^{-1}\cdot c_2$	1.00	1.02	0.04	1.00	1.02	0.04
$10^{-1}\cdot s_2$	1.00	1.17	1.55	1.00	1.43	1.59
$10^{-1}\cdot p_a$	5.00	5.11	1.01	5.00	5.04	0.76
$10^{-1}\cdot r_1$	1.00	1.26	0.73	0.50	0.44	0.21
$10^{-1}\cdot p_b$	5.00	4.86	1.03	5.00	4.97	0.68
$10^{-1}\cdot r_2$	1.00	1.06	0.38	1.00	0.95	0.35
$10^{-1}\cdot q_1$	1.00	0.85	0.43	1.00	0.96	0.29
$10^{-1}\cdot q_2$	1.00	0.81	0.41	0.50	0.47	0.29

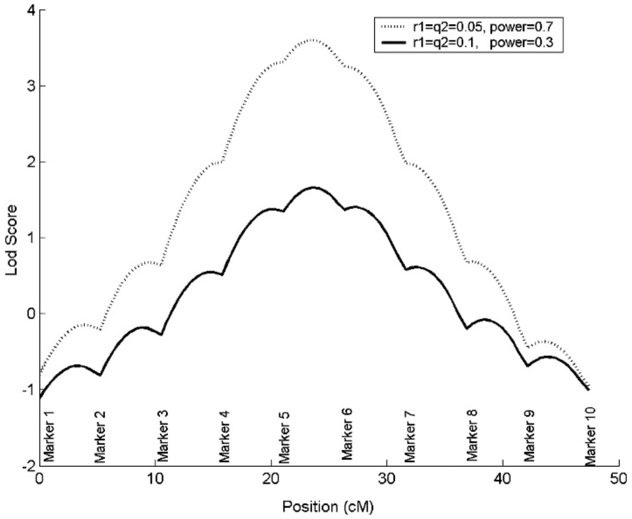

Figure 1 Lod score profile. The first longevity gene is situated between the fifth and sixth chromosomes. The second longevity gene is out of the chromosome. $p_a=p_b=0.5$; $r_2=q_1=0.1$.

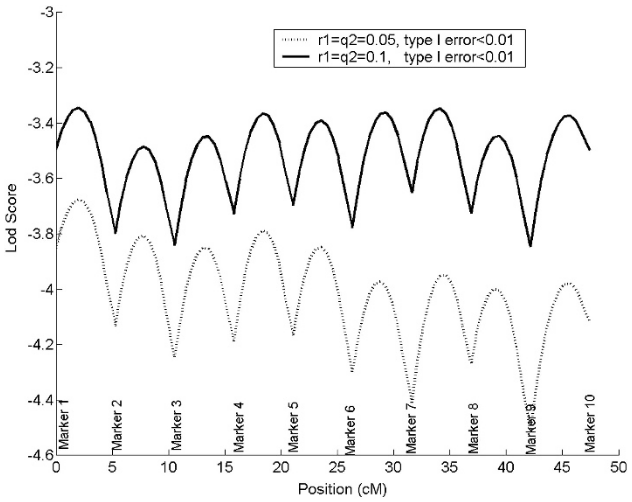

Figure 2 Lod score profile. Both longevity genes are out of the chromosome. $p_a=p_b=0.5$; $r_2=q_1=0.1$.

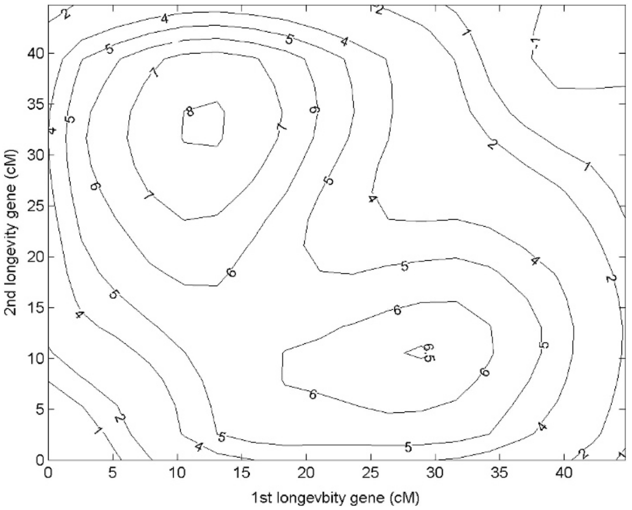

Figure 3 Contour map of the Lod score profile (smoothed). The first longevity gene was situated in the middle between the third and fourth markers and the second longevity gene was situated in the middle between the seventh and eighth markers. Power>0.99. $p_a = p_b = 0.5$; $r_2 = q_1 = 0.1$; $r_1 = q_2 = 0.05$.

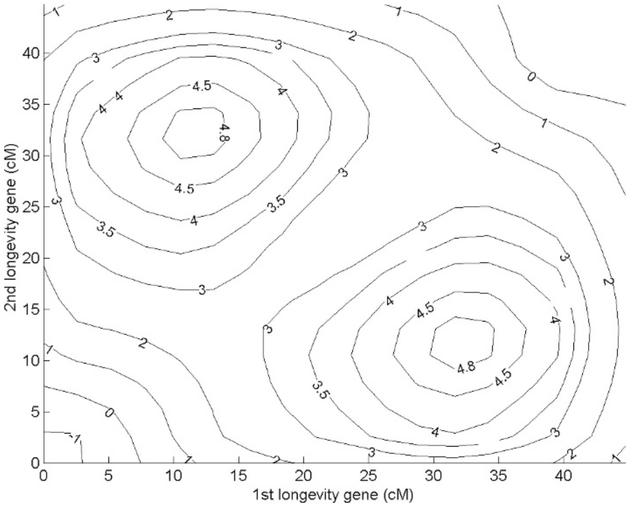

Figure 4 Contour map of the Lod score profile (smoothed). The first longevity gene was situated in the middle between the third and fourth markers and the second longevity gene was situated in the middle between the seventh and eighth markers. Power=0.89. $p_a = p_b = 0.5$; $r_2 = q_1 = 0.1$; $r_1 = q_2 = 0.1$.

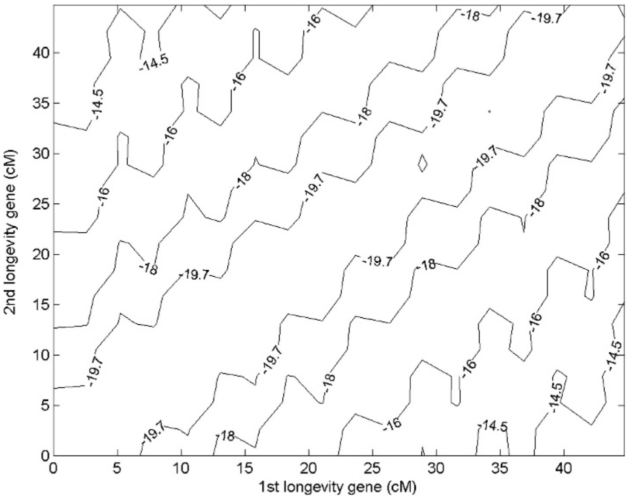

Figure 5 Contour map of the Lod score profile. Both longevity genes are out of the chromosome. Type I error<0.01. $p_a = p_b = 0.5$; $r_2 = q_1 = 0.1$; $r_1 = q_2 = 0.05$.

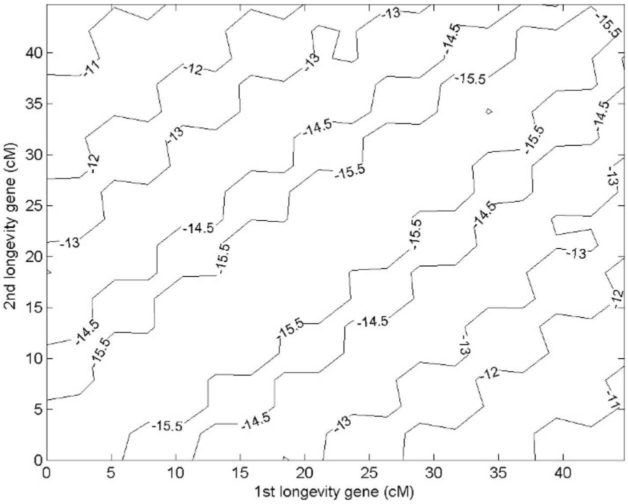

Figure 6 Contour map of the Lod score profile. Both longevity genes are out of the chromosome. Type I error<0.01. $p_a = p_b = 0.5$; $r_2 = q_1 = 0.1$; $r_1 = q_2 = 0.1$.

seventh and eighth markers) (Figure 7). In this case, we do not observed clear peaks, but the plateau. The area, where the Lod score is greater than 3, extends ∼32 cM.

DISCUSSION

In a previous paper[13], a two-step procedure was used to estimate the parameters of univariate fit, frailty distribution and the location of the longevity gene. In this paper, we extended this method for the case of two longevity genes and two correlated competing risks of mortality. The presence of longevity genes in the genome can be tested in the first step. In the second step, we locate the position of these genes in the genome. There is no problem to extend the case of twin data with two related individuals to the case of family data with more than two related individuals if we use the model based on the major gene concept. The major gene model makes it possible to take into account not only the correlation between individuals but also the correlations between competing risks of mortality using parameters q_1 and q_2. The covariates such as age, gender, disease status and so on, which influence the risk of mortality can be easily taken into account in the form of the Cox-type regression. The unknown regression coefficients can be estimated in the first step together with parameters b_j, c_j, s_j, $j = 1$, 2 and p_a, p_b, q_1, q_2, r_1, r_2. From our experiments with simulated data sets, we see that parameters p_a, p_b, q_1, q_2, r_1, r_2 influence the values of Lod scores and heights of possible peaks. The smaller the values of q_1, q_2, r_1, r_2, the higher the peaks of the Lod scores and the greater the possibility of longevity genes detection and localization. In principle, we can extend this model to one with mixed frailty by including an additional continuously distributed component of

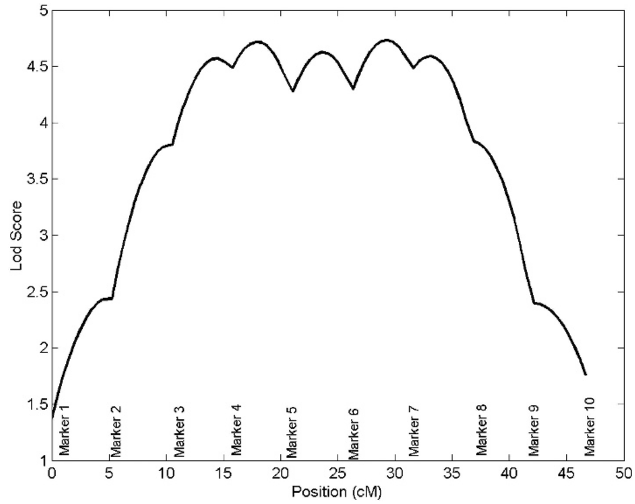

Figure 7 Lod score profile. The first longevity gene was situated in the middle between the third and fourth markers and the second longevity gene was situated in the middle between the seventh and eighth markers ($p_a = p_b = 0.5$; $r_2 = q_1 = 0.1$; $r_1 = q_2 = 0.1$). The Lod score profile was calculated under the assumption of a single dominant major gene model.

frailty. This continuous component will measure the averaged influence on mortality of a large number of genes and environment. However, it is not reasonable for sample sizes used in this study, as the bivariate probability density functions for the model with discretely and continuously distributed frailties are very similar.[8] In some cases, we can detect the presence of model misspecification analyzing the Lod score profiles. For example, large plateau in experiments with a single major gene can indicate the presence of two major genes situated on the same chromosome. On the contrary, if a single major gene influences the lifespan and the present method is used, we will observe a peak of Lod score situated near the diagonal.

1 De Benedictis, G., Carotenuto, L., Carrieri, G., De Luca, M., Falcone, E., Rose, G. *et al.* Gene/longevity association studies at four autosomal loci (REN, THO,PARP, SOD2). *Eur. J. Hum. Genet.* **6,** 534–541 (1998).

2 Garasto, S., Rose, G., Derango, F., Berardelli, M., Corsonello, A., Feraco, E. *et al.* The study of APOA1, APOC3, and APOA4 variability in healthy ageing people reveals another paradox in the oldest old subjects. *Ann. Hum. Genet.* **67,** 54–62 (2003).

3 Yashin, A., De Benedictis, G., Vaupel, J. W., Tan, Q., Andreev, K. F., Iachine, I. A. *et al.* Genes, demography, and life span: The contribution of demographic data in genetic studies on aging and longevity. *Am. J. Hum. Genet.* **65,** 1178–1193 (1999).

4 Begun, A. Detecting genes contributing to longevity using twin data. *Hum. Genomics* **4,** 73–78 (2009).

5 Yashin, A. I. & Iachine, I. A. Genetic analysis of durations: correlated frailty model applied to survival of Danish twins. *Genet. Epidemiol.* **12,** 529–538 (1995).

6 Neale, M. C. & Cardon, L. R. *Methodology for Genetic Studies of Twins and Families* (Kluwer, Dodrecht, Netherlands, 1992).

7 Wienke, A., Christensen, K., Skytthe, A. & Yashin, A. Genetic analysis of cause of death in a mixture model of bivariate lifetime data. *Stat. Modelling* **2,** 89–102 (2002).

8 Begun, A., Iachine, I. A. & Yashin, A. Genetic nature of individual frailty: comparison of two approaches. *Twin Res.* **3,** 51–57 (2000).

9 Begun, A., Desjardins, B., Iachine, I. & Yashin, A. in *Medical Infobahn for Europe (Hasman, A. et al. eds) 412–416 (IOS Press, 2000).*

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

11 Jonker, M. A., Bhulai, D. I., Boomsma, D. I., Ligthart, R. S. L., Posthuma, D. & Van der Vaart, A. W. Gamma frailty model for linkage analysis with application to interval censored migraine data. *Biostat.* **10,** 187–200 (2009).

12 Callegaro, A., Van Houwelingen, H. C. & Houwing-Duistermaat, J. J. Score test for age at onset genetic linkage analysis in selected sibling pairs. *Stat. Med.* **28,** 1913–1926 (2009).

13 Begun, A. & Yashin, A. Genetic markers data in survival studies of twins: the results of a Simulation Study. *Twin Res. Hum. Genet.* **8,** 34–38 (2005).

14 Gorfine, M. & Hsu, L. Frailty-based competing risks model for multivariate survival data. *Biometrics* **67,** 415–426 (2011).

15 Sham, P. *Statistics in Human Genetics* (Wiley, New York, NY, USA, 1998).

16 Zeng, D. & Lin, D. Y. Maximum likelihood estimation in semiparametric regression models with censored data. *J. R. Stat. Soc.* **B69,** 507–564 (2007).

17 Barnard, G. A. Statistical inference. *J. R. Stat. Soc.* **B11,** 115–139 (1949).

18 Morton, N. E. Sequential tests for the detection of linkage. *Am. J. Hum. Genet.* **7,** 277–318 (1955).

19 Lander, E. S. & Green, P. Construction of multilocus genetic maps in humans. *Proc. Natl Acad. Sci. USA* **84,** 2363–2367 (1987).

20 Kruglyak, L. & Lander, E. S. Complete multipoint sib pair analysis of qualitative and quantitative traits. *Am. J. Hum. Genet.* **57,** 439–454 (1995).

21 Kruglyak, L., Daly, M. J., Reeve-Dali, M. P. & Lander, E. S. Parametric and non-parametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.* **58,** 1347–1363 (1996).