# Using survival data in gene mapping : using survival data in genetic linkage and family-based association analysis

Callegaro, A.

CHAPTER 2

# Score test for age at onset genetic linkage analysis in selected sibling-pairs

**Abstract**

*A new score statistic is derived which uses information from registries (age-specific incidences) and family studies (sib-sib marginal correlation) to weight affected sibling pairs according to their age at onset. Age at onset of sibling pairs is modelled by a gamma frailty model. From this model we derive a bivariate survival function which depends on the marginal survival and on the marginal correlation. The score statistic for linkage is a classical Non Parametric Linkage statistic where the identical by descent sharing is weighted by a particular function of the age at onset data. Since the statistic is based on survival models, it can also be applied to discordant and healthy sibling pairs. Simulation studies show that the proposed method is robust and more powerful than standard nonparametric linkage methods. As illustration we apply the new score statistic to data from a breast cancer study.*

## 2.1 Introduction

Although many traits are heritable, identification of responsible genes appears to be a challenge. Recently new loci have been discovered by genome wide association studies, but they explain only a part of the genetic variation and a lot remains to be recovered. Focussing on chromosomal areas with linkage signals is a way to find these genes. In practice, to perform linkage analysis families are selected based on their phenotypes. A commonly used design for linkage is

This chapter has been published as: A. Callegaro, H. C. van Houwelingen, and J. J. Houwing-Duistermaat (2009). Score test for age at onset genetic linkage analysis in selected sibling-pairs. *Statistics in Medicine 28*, 1913–1926.

the affected sibling pair design (ASP). When age at onset is available and varies among sibling pairs the question arises whether age at onset contributes to the linkage effect. In the ASP data it is not possible to estimate the marginal (population) survival function and/or the marginal correlation between siblings from the data. However, registries provide accurate estimation of the marginal hazard rate since it is based on thousands of individuals. From twin studies, the estimates of sib-sib correlations are also available for many diseases (Wienke et al., 2003; Yashin et al., 1995; Zahl, 1997; Zdravkovic et al., 2002). Using this information, we propose a new score test for age at onset in linkage analysis. If age at onset plays a role the power to detect linkage will be increased.

As a motivating example, we consider a Dutch ASP data set on breast cancer for genetic linkage analysis. The ages at onset of the ASP are known. Information on age specific incidence is available from the Dutch breast cancer registry (`http://www.rivm.nl/vtv/object_document/o1502n17276.html`). Frailty parameters of the Dutch population are unknown, however they have been estimated for the Swedish population ($\rho = 0.125$ and $\sigma^2 = 25$) (Wienke et al., 2003). The question is whether evidence for linkage depends on age at onset. To answer this question we derive a new score test and apply it to these data.

The new score test is derived from a model for dependent survival data. Hougaard (2000) discusses various multivariate survival models. The dependence between relatives can be modelled using random effects survival models (frailty models). The observed times of a pair of relatives are independent given the unobserved frailty. Clayton (1978) and Vaupel et al. (1979) proposed a frailty model where the dependence between observations is modelled by one shared random effect. In order to describe more complex dependency structures the shared model was extended by Yashin et al. (1995), who proposed a correlated frailty model to jointly model the correlation structure of monozygotic and dizygotic twins. Since monozygotic twins share all their genes and dizygotic twins share only half of their genes, times observed in monozygotic twins are more correlated than the outcomes of dizygotic twins when genetic effects play a role. To model these correlations Yashin et al. (1995) divided the frailty into a sum of independent gamma-distributed effects. These additive models are mathematically appealing and can be considered as variance components models for life times (Petersen, 1998; Petersen et al., 1996).

For genetic linkage analysis, methods based on frailty models have been proposed for survival data (age at onset) (Commenges, 1994; Jonker et al., 2009; Li and Zhong, 2002; Pankratz et al., 2005; Siegmund and Todorov, 2000; Sun and Li, 2004). Commenges (1994) proposed a frailty model where the linkage effect is decomposed into the sum of two random effects representing the two alleles at the locus linked to a disease susceptibility gene. This model considers only

one linkage effect, hence under the null hypothesis of no linkage, relatives are assumed to be independent. For complex genetic traits this assumption is not realistic. In order to take into account residual correlation Li and Zhong (2002) proposed an additive gamma-frailty model where the frailty is decomposed into the sum of the linkage effect and a shared residual effect. This model is a shared frailty model under the null hypothesis of no linkage (correlation equal to one). In the line of the work of Yashin et al. (1995), Jonker et al. (2009) extended the Li and Zhong model by adding an individual effect, in such a way that, even at an unlinked locus the model is a correlated frailty model. The main limitation of all these frailty methods for linkage is that they are not valid for selected samples.

In this paper, using a correlated gamma frailty model (Jonker et al., 2009) together with a retrospective likelihood (Kruglyak et al., 1996; Li and Zhong, 2002; Whittemore, 1996), we derive a score test valid for any ascertainment scheme. It is a weighted nonparametric linkage (NPL) statistic (Kruglyak et al., 1996) where the (centered) proportions of alleles shared identical by descent (IBD) are weighted by a known function of the ages at onset of the two siblings. The proposed method is an extension of the score test proposed by Houwing-Duistermaat et al. (2009), which is derived from a shared-frailty model. Both methods are similar in spirit to the linkage analysis of quantitative traits in selected samples (Lebrec et al., 2004; Sham and Purcell, 2001; Tang and Siegmund, 2001; Tritchler et al., 2003), which also uses known information on the distribution of the phenotype. In general, with respect to the likelihood-ratio approaches (Jonker et al., 2009; Li and Zhong, 2002) the score test is less model dependent and therefore more robust against misspecification of the model.

In section 2.1 we briefly describe the correlated frailty model for twins studies (Yashin et al., 1995). In section 2.2 we show that a similar model can be used for linkage analysis and we derive the conditional hazard ratio for relative pairs who share 0,1, or 2 IBD alleles. In section 2.3 we derive a score statistic from the retrospective likelihood of the marker data conditional on the phenotype in order to test for linkage. The power and the robustness of the proposed method is studied in section 3 by means of simulation. In section 4 we illustrate the score test by analyzing age at onset data on breast cancer (Oldenburg et al., 2008). Conclusions, some remarks and suggestions for future developments are given in section 5.

## 2.2 Methods

**Correlated frailty model for twin data**

Let $T_j$ and $Z_j$ $(j = 1, 2)$ be the life span and the frailties of two related individuals. The observed data are given by $y_j = (t_j, \delta_i)$ where $t_j$ is the observed age at onset if $\delta_j = 1$ and age at censoring if $\delta_j = 0$. The individual hazards are independent given $Z_j$ and represented by the hazard model $\lambda_j(t) = \lambda_0(t; X_j)Z_j$, $j = 1, 2$. Note that we assume a general dependence between the baseline hazard ($\lambda_0$) and the vector of covariates ($X_j$).

To jointly model survival data observed in monozygotic and dizygotic twins, Yashin et al. (1995) proposed an additive gamma model. Outcomes observed in twins may be correlated due to shared environmental and genetic effects. Because dizygotic twins share half of their genes and monozygotic twins share all their genes, the outcomes of monozygotic twins are more correlated than the outcomes of dizygotic twins when genetic effects are present. To model this difference in correlation, Yashin et al. (1995) used three gamma distributed components, namely one component for correlation due to genetic effects, one component model for correlation due to shared environmental effects, and one component for individual variation. Let $(Z_1, Z_2)$ be the frailties for twin 1 and twin 2 respectively. Then these frailties can be written as follows

$$\begin{aligned} Z_1 &= Z_{1a} + Z_c + Z_{1e} \\ Z_2 &= Z_{2a} + Z_c + Z_{2e}. \end{aligned} \tag{2.1}$$

with $(Z_{1a}, Z_{2a})$ modelling additive genetic effects which are common for monozygotic twins and partly shared by dizygotic twins, $Z_c$ modelling shared environmental effects which are common for monozygotic as well as dizygotic twins and $(Z_{1e}, Z_{2e})$ modelling individual variation. These latter components are independent. The three components follow a gamma distribution with the following parameters $Z_{ja} \sim \Gamma(\nu_a, 1/\sigma^2)$, $Z_c \sim \Gamma(\nu_c, 1/\sigma^2)$ and $Z_{je} \sim \Gamma(\nu_e, 1/\sigma^2)$. It is common to use the following constraint $1/\sigma^2 = \nu_a + \nu_c + \nu_e$ in order to obtain $E(Z_j) = 1$, $j = 1, 2$. Let $a^2 = \nu_a\sigma^2$, $c^2 = \nu_c\sigma^2$ and $e^2 = \nu_e\sigma^2$, then the correlation between the frailties $(Z_1, Z_2)$ is given by

$$\rho_\Psi = 2\Psi a^2 + c^2, \tag{2.2}$$

with $\Psi$ the kinship coefficient ($\Psi = 1/2$ for monozygotic twins and $\Psi = 1/4$ for DZ twins).

Now the marginal bivariate survival function $S_{12}(t_1, t_2)$ can be written as function of univariate survival functions, $\rho_\Psi$ and $\sigma^2$:

$$S_{12}(t_1, t_2) = \frac{S_1(t_1)^{1-\rho_\Psi} S_2(t_2)^{1-\rho_\Psi}}{(S_1(t_1)^{-\sigma^2} + S_2(t_2)^{-\sigma^2} - 1)^{\rho_\Psi/\sigma^2}}, \tag{2.3}$$

where $S_j = \exp(-H_j)$ is the marginal survival, and $H_j$ is the marginal cumulative hazard for the $j$th twin.

**Correlated frailty model for linkage analysis of sibling pairs**

For linkage analysis of survival data observed in sibling pairs, a similar model can be used. Instead of modelling the correlation of the frailties as a function of the kinship coefficients, the correlation is modelled as a function of the proportions of alleles shared identical by descent ($\pi$) at a particular locus. When linkage at a locus is present, the correlation between frailties tends to increase with the number of alleles shared IBD. The decomposition of the frailty is given by

$$\begin{aligned} Z_1 &= Z_{1l} + Z_s + Z_{1e} \\ Z_2 &= Z_{2l} + Z_s + Z_{2e}, \end{aligned} \tag{2.4}$$

with $(Z_{1l}, Z_{2l})$ correlated components which model linkage at the locus, $Z_s$ a component which models correlation due to shared effects and $(Z_{1e}, Z_{2e})$ independent components which model individual variation. To model the correlation due to sharing alleles IBD at a locus, Commenges (1994) and Li and Zhong (2002) wrote the random component $Z_{jl}$ $(j = 1, 2)$ as the sum of two random effects representing the two alleles of the $j$th individual (see appendix 1). Based on this decomposition and using an additive genetic model for the locus, the correlation between $(Z_{1l}, Z_{2l})$ is equal to the proportion of alleles shared IBD ($\pi$). Note that if only data on siblings are available, the shared environmental and residual genetic effects cannot be disentangled, hence $Z_s$ also models correlation due to unlinked genes. When relative pairs with different kinship coefficients are available, a residual genetic effect and a shared environmental effect can be modelled. The three components are gamma distributed with the following parameters: $Z_{jl} \sim \Gamma(\nu_l, 1/\sigma^2)$, $Z_s \sim \Gamma(\nu_s, 1/\sigma^2)$ and $Z_{je} \sim \Gamma(\nu_e, 1/\sigma^2)$. The variance of the random effect is $(1/\sigma^2 = \nu_l + \nu_s + \nu_e)$.

Let $\gamma = \nu_l \sigma^2$ be the linkage parameter. The correlation between the frailties $(Z_1, Z_2)$ of two siblings given $\pi$ at the locus is given by

$$\rho_\pi = \pi\gamma + s^2, \tag{2.5}$$

where $s^2 = \nu_s \sigma^2$ and the bivariate survival function is now

$$S_{12}(t_1, t_2|\pi; \gamma) = \frac{S_1(t_1)^{1-\rho_\pi} S_2(t_2)^{1-\rho_\pi}}{(S_1(t_1)^{-\sigma^2} + S_2(t_2)^{-\sigma^2} - 1)^{\rho_\pi/\sigma^2}}. \tag{2.6}$$

The correlation between the two frailties given $\pi$ (2.5) can also be written in terms of the marginal sib-sib correlation $\rho$ (Tang and Siegmund, 2001)

$$\rho_\pi = (\pi - E\pi)\gamma + \rho = (\pi - 1/2)\gamma + \rho, \tag{2.7}$$

with the marginal sib-sib correlation equal to $\rho = 1/2\gamma + s^2$. Note that if a linkage effect is present at the locus ($\gamma > 0$), the correlation between frailties of siblings sharing two alleles IBD will be higher and the correlation between frailties of siblings sharing zero alleles IBD will be lower than the correlation in the population. The model used by Li and Zhong (2002) corresponds to model (2.4) without the term $Z_{je}$, hence assuming $\nu_e = 0$. Note that for $\nu_e = 0$ the marginal correlation $\rho$ equals 1 under the null hypothesis of $\gamma = 0$. It follows that for unlinked loci the Li and Zhong (2002) model corresponds to a shared frailty model.

Model (2.6) with the parametrization (2.7) is particularly attractive in linkage analysis where sib-pairs are commonly selected on the phenotype, because it depends on information about the age at onset at the population level $(S(t), \rho, \sigma^2)$. This information is often available from registries and family (twin) studies. For example, when twin data have been analyzed using model (2.1) estimates of the variance $\sigma^2$ and the marginal sib-sib correlation $\rho$ are available $(\rho = 1/2\hat{a}^2 + \hat{c}^2)$.

From the proposed model, the conditional hazard ratio function for sib pairs (Clayton, 1978; Li and Zhong, 2002; Oaks, 1989) given $\pi$ is given by

$$\phi_\pi(t_1, t_2) = \frac{\lambda_\pi(t_1|T_2 = t_2)}{\lambda_\pi(t_1|T_2 > t_2)} = 1 + \frac{\rho_\pi \sigma^2}{\Delta_{1\pi}\Delta_{2\pi}}, \tag{2.8}$$

where $\Delta_{1\pi} = (1 + (1 - \rho_\pi)S(t_1)^{\sigma^2}(S(t_2)^{-\sigma^2} - 1))$ and $\Delta_{2\pi} = (1 + (1 - \rho_\pi)S(t_2)^{\sigma^2}(S(t_1)^{-\sigma^2} - 1))$. The conditional hazard ratio is a decreasing function from $1 + \rho_\pi\sigma^2$ to $1 + \rho_\pi\sigma^2/(2 - \rho^2)$ when $t_1 = t_2 = t$. When $e^2 = 0$ the conditional hazard ratios reduce to those derived in Li and Zhong (2002).

As an example we show the conditional hazard ratio for breast cancer using the Dutch marginal survival function (Figure 2.1(a)) and marginal frailty parameters for the Swedish population ($\rho = 0.125$, $\sigma^2 = 25$) (Wienke et al., 2003). Figures 2.1(b) to 1(d) show the conditional hazard ratio (2.8) for siblings sharing 0, 1 and 2 alleles shared IBD at the disease locus which explains 10% of the total variability ($\gamma = 0.1$). Clearly $\phi(t_1, t_2)$ depends on the ages of the two siblings: if $t_2$ is small then the cross ratio decreases with $t_1$ but when $t_2$ is large, the cross ratio is an increasing function of $t_1$. The ratio increases with respect to the number of alleles shared IBD. The cross-ratios observed under the null hypothesis of no linkage corresponds to the case of 1 allele shared IBD (Figure 2.1(c)).

**Score test for selected sib-pairs**

Denote the genotypes at the marker by MD. From these marker data the probability of the proportion of alleles shared IBD by a sib-pair $P_0(\pi|\text{MD})$ can be estimated at each location by standard software for linkage analysis (i.e. MERLIN (Abecasis et al., 2002), GENEHUNTER (Kruglyak et al., 1996)). For selected samples, the conditional distribution of the marker data given the phenotype $\mathbf{y}=(t_1, \delta_1, t_2, \delta_2)$ gives a natural framework for testing linkage (Kruglyak et al., 1996; Whittemore, 1996). The retrospective log-likelihood is given by

$$\log P(\text{MD}|\mathbf{y};\gamma) = \log[\sum_{\pi\in\{0,0.5,1\}} P(\mathbf{y}|\pi;\gamma)P_0(\pi|\text{MD})] - \log[\sum_{\pi\in\{0,0.5,1\}} P(\mathbf{y}|\pi;\gamma)P_0(\pi)] + \log P_0(\text{MD}).$$

$P(\mathbf{y}|\pi;\gamma)$ is the likelihood of two individuals who share $\pi$ proportion of allele shared IBD (Appendix 2) and is derived from the marginal bivariate survival function (2.6). $P_0(\pi)$ is the probability under the null hypothesis of no linkage.

To test the null hypothesis $H_0 : \gamma = \gamma_0 = 0$ versus $H_1 : \gamma > 0$, a score test based on the retrospective likelihood is proposed. The score statistic for $n$ independent sib-pairs is given by

$$\text{NPL}_\rho = \frac{\sum_{i=1}^n (\hat{\pi}_i - E\pi_i)\ell_{\rho,i}(\gamma_0)}{\sqrt{\sum_{i=1}^n \text{var}_0(\hat{\pi}_i)\ell_{\rho,i}(\gamma_0)^2}}, \tag{2.9}$$

where $\hat{\pi}_i$ is the expected value of the $\pi$ given the marker genotypes of the $i$th sib-pair, $E\pi_i = 1/2$ for sibling pairs and $\ell_{\rho,i}(\gamma_0) = \frac{\partial \log P(\mathbf{y}_i|\pi;\gamma_0)}{\partial\rho}$(Appendix 3). For the univariate survival functions the known population (marginal) survivals are used and for $(\rho, \sigma^2)$ the values are obtained from previous twin studies. The variance (the denominator) of the statistic is the robust empirical variance given by the square of the score function. In the case of incomplete data on IBD, $\text{var}_0(\hat{\pi}_i)$ can be estimated by multipoint simulations.

The statistic (2.9) is part of a class of statistics for genetic linkage called weighted NPL statistics (Kruglyak et al., 1996) where the excess IBD sharing $(\pi - E\pi)$ is weighted by particular weight functions. In our case the weight function, $\ell_\rho(\gamma_0)$, depends on the marginal survival distribution of the two siblings and on the population frailty parameters. For breast cancer we show how the weights depend on the ages of the two siblings. Figure 2.2(left) shows the contour plot of the weights for ASP. Figure 2.2(right) shows the contour plot of the weight for discordant sibling pairs, i.e. one sibling is affected and the other sibling is censored. Figure 2.2 shows that the largest weights are assigned to early-onset sibling pairs.

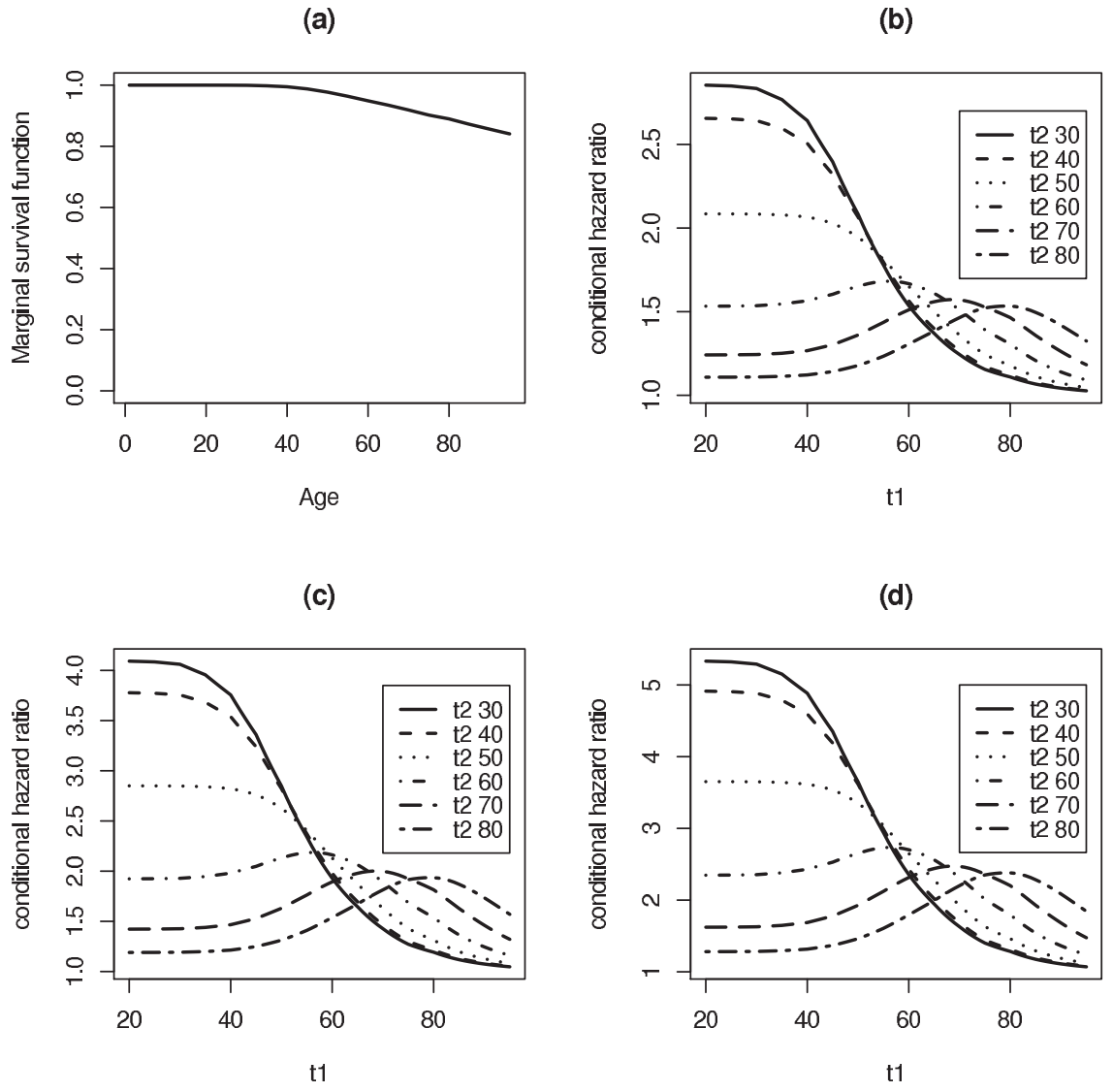


**FIGURE 2.1:** *Conditional hazard ratios for sib pair (1,2) with breast cancer for the three IBD categories. Fig 1(a) shows the Dutch breast cancer marginal survival function. Subfigures (b), (c) and (d) show the conditional hazard ratio for siblings sharing 0, 1 and 2 alleles IBD.*

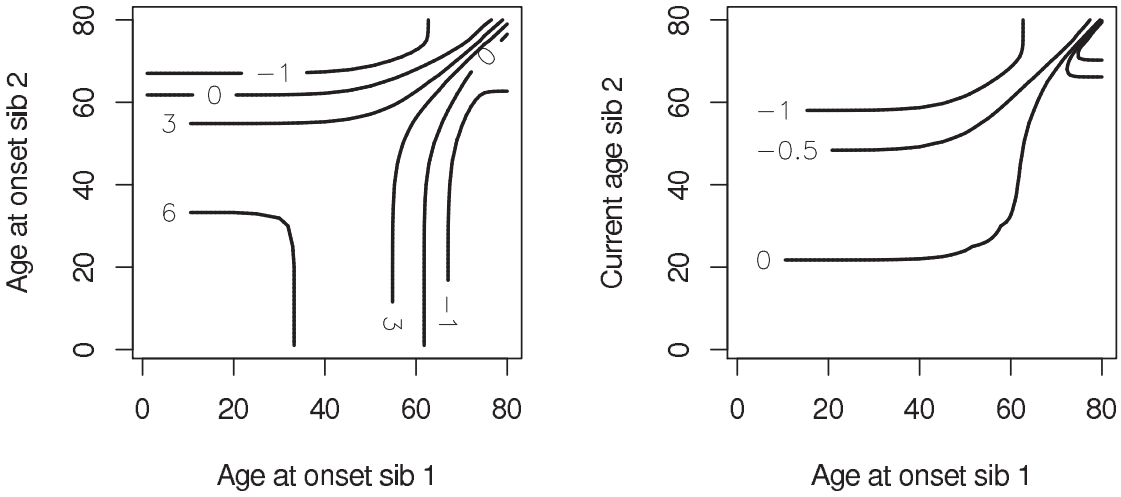


**FIGURE 2.2:** *Weight function for breast cancer. Figure Left: Contour plot of $\ell_\rho(\gamma_0)$ for ASP. Figure Right: Contour plot of $\ell_\rho(\gamma_0)$ for one affected sibling (sib 1) and one unaffected sibling (sib 2).*

In the next sections we will compare the proposed score statistic ($\text{NPL}_\rho$) with the unweighted mean test (NPL) (Blackwelder and Elston, 1985) and with two score tests derived from additive gamma frailty models namely $\text{NPL}_0$, and $\text{NPL}_{\rho(\rho=1)}$. $\text{NPL}_0$ is a weighted NPL statistic where the excess IBD is weighted by the product of the martingale residuals of the two relatives ($w = (\delta_1 - H_1)(\delta_2 - H_2)$) (Commenges, 1994; Houwing-Duistermaat et al., 2009). This statistic is derived from the retrospective likelihood corresponding to the model of Commenges (1994). The score test $\text{NPL}_{\rho(\rho=1)}$ is given by (2.9) with $\rho$ equal to one and the variance equal to the variance estimated at population level by a gamma shared frailty model. This test is the score test for the Li and Zhong (2002) model.

For genome-wide analysis, the NPL statistic is computed at each marker locus. Morton (1955) obtained the threshold level of 3 (p=0.0001) for the LOD-score (defined as sign(NPL)$\text{NPL}^2/(2\ln 10)$) used to declare the presence of linkage in a genome-wide study.

## 2.3 Simulation studies

### Gamma frailty model

Age at onset was simulated from gamma frailty models, with Gompertz distributed baseline hazard ($\lambda_0(t) = be^{\tau t}, b = 1/10^5, \tau = 0.120$). Current age was simulated as age at censoring from a uniform distribution $U(60, 80)$. In order to obtain an estimate of the marginal survival we calculated Kaplan Meier survival using an independent simulated population of 5000 random pedigrees.

Respectively 10% and 40% of the population was affected by age 60 and 80. Two configurations of random effects were considered. In the first, data were simulated without environmental effects ($\gamma = 1, s^2 = 0, e^2 = 0$). In the second, data were simulated with shared and unshared environmental effects ($\gamma = 0.33, s^2 = 0.33, e^2 = 0.33$). In both cases, the variance of the random effect was equal to $\sigma^2 = 3.3$ and the marginal sib-sib correlation was equal to $\rho = 0.5$. Each scenario included 10000 replications. For each replication the first 200 affected sib pairs (ASPs) were used for analysis.
Figure 2.3 shows the power (proportion of LOD scores greater than 3) of the proposed score test ($NPL_\rho$) as a function of $\sigma^2$ for different values of the parameter $\rho$. Long-dashed lines represent the power of the $NPL_\rho$ method when the true population correlation is used in the weight function. The two horizontal lines represent the NPL (solid line) and the $NPL_0$ test (dashed line). These tests are independent of the frailty parameters. $NPL_\rho$ has maximal power around the true population parameters values. There is a considerable gain in power achieved by our method when the population frailty parameters are known. In the case of no environmental effect, the gain in power of the proposed $NPL_\rho$ compared to the $NPL_0$ and to the unweighted NPL method is about 5% and 20%, respectively (Figure 2.3 left). In the case of environmental effects, the gain in power of the proposed $NPL_\rho$ compared to the $NPL_0$ and to the unweighted NPL method is about 15% and 50%, respectively (Figure 2.3 right). The proposed method performs better than the other methods for a wide range of $(\rho, \sigma^2)$ values specified in the weight. For example, in the first configuration (Figure 2.3 left) the marginal values of the frailty parameters are $(\rho, \sigma^2)$=(0.5,3.3) but the score test $NPL_\rho$ is more powerful than the classical NPL test for all the values of $\rho > 0.1$ and $\sigma^2 < 4$. For $\sigma^2$ values smaller than the population value, $NPL_\rho$ and $NPL_0$ have similar performance. Note that also $NPL_0$ performs better than the classical NPL method. When the value of $\sigma^2$ used in the weight function is larger than twice the marginal value, $NPL_\rho$ loses power with respect to the NPL method.

For each configuration of data, we also simulated under the null hypothesis. All the tests have correct type I error rates (data not shown).

**Dominant major gene model**

Age at onset data were generated using a major gene model with a normally distributed residual effect, as described by Commenges and Abel (1996). The annual incidence was assumed to be constant and equal to 0.002 over the interval 0-60 years, then 11% of the population is affected by age 60 years (Figure 2.4(left)). A normally distributed underlying liability, $s = \mu + \epsilon$, was assumed with a major gene effect ($\mu$), and an individual environment ef-

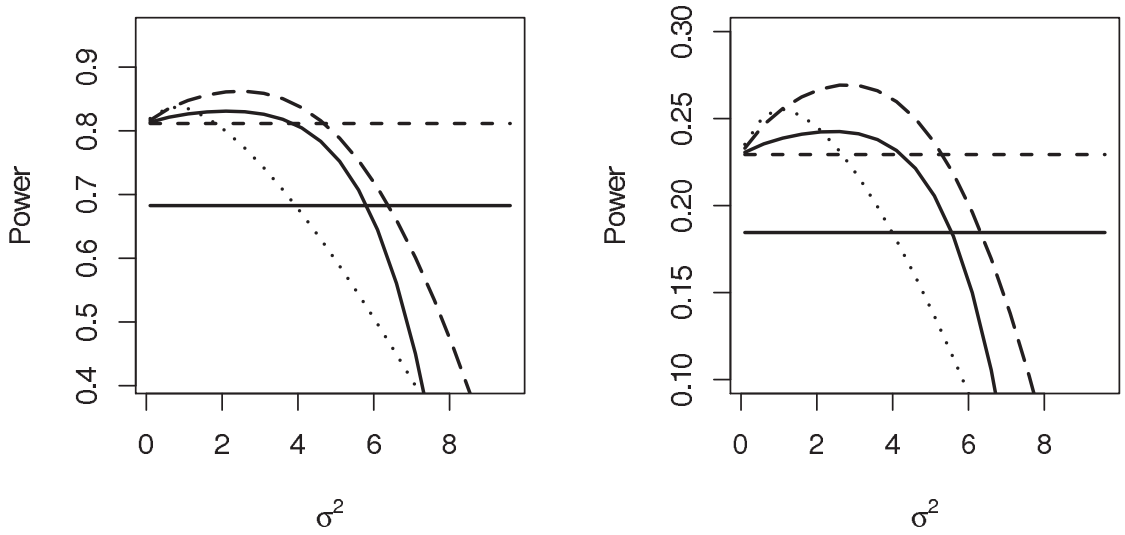


**FIGURE 2.3:** *Power of the test statistics based on Gamma frailty simulated data. Horizontal lines represent the unweighted NPL (solid line) and the NPL$_0$ method (dashed line). Dotted lines, long-dashed lines and solid lines represent the power of NPL$_\rho$ with correlation values ρ=0.1, 0.5 and 1, respectively. Figure Left: No environmental effect. Figure Right: environmental effects.*

fect ($\epsilon \sim N(0, 0.648)$). The major gene was diallelic (A/B) with $p = 0.05$, $\mu_{BB} = -0.195$ and $\mu_{AA} = \mu_{AB} = -1.805$. Affection at age $k$ was defined by having a $s$ value above an age specific threshold ($thr_k$). An individual with a liability value less than the last threshold value ($thr_{60}$) was not affected; otherwise he became affected at age $k$ such that $thr_k < s < thr_{k-1}$ with $thr_0 = \infty$. The cumulative probability of being censored by age $k$ was equal to $Pc(k) = 1/[1 + \exp(10 - 0.2k)]$. We generated a marker with 21 alleles in order to have perfect IBD and 10000 replicates. For each replication the first 200 affected sib pairs (ASPs) were ascertained. The marginal survival and the population frailty parameters $(\hat{\rho}, \hat{\sigma}^2) = (0.3, 11)$ were estimated from an independent random data-set of 5000 sib-pairs.

Figure 2.4(right) shows the power of the NPL$_\rho$ statistic as function of $\sigma^2$ for various values of $\rho$. The long-dashed line represents the power of the NPL$_\rho$ test with the value of $\rho$ equal to the marginal correlation, $\rho = 0.3$. The two horizontal lines represent the power of NPL and NPL$_0$. When the population frailty parameters are known, the gain in power of NPL$_\rho$ compared to NPL$_0$ and to the unweighted NPL method is about 5% and 10%, respectively. The proposed score test has not maximal power at the marginal correlation value, but at $\rho = 0.1$. However, figure 2.4(right) shows that for a wide range of the $(\rho, \sigma^2)$ values, the score test NPL$_\rho$ is the most powerful test. On these data, the score test NPL$_0$ performs better than the shared residual effect score test (NPL$_{\rho(\rho=1)}$) because the true marginal correlation is small.

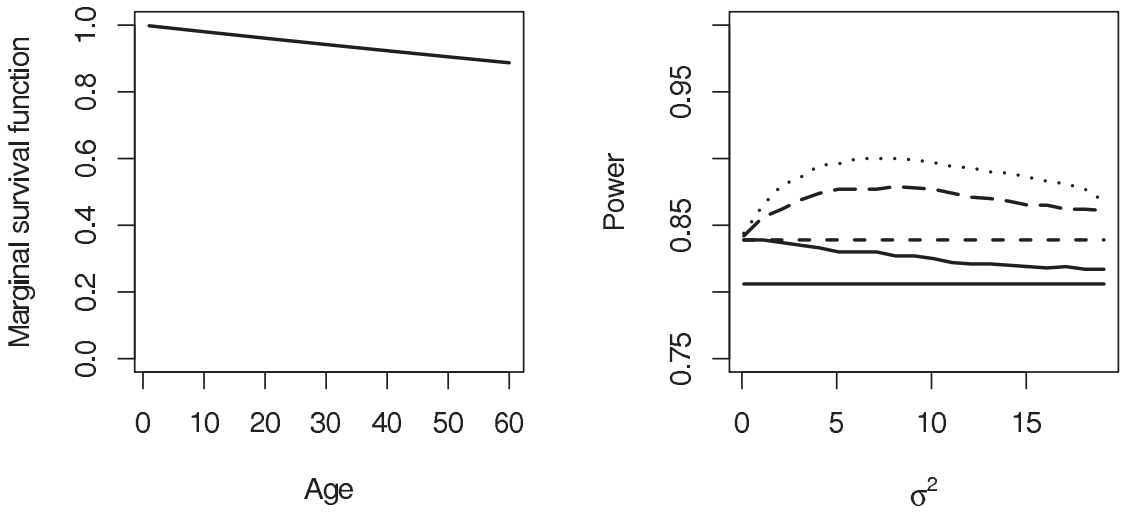


**FIGURE 2.4:** *Power of the test statistics based on dominant major gene simulated data. Figure Left: Marginal survival function. Figure Right: Power of the test statistics. Horizontal lines represent the unweighted NPL (solid line) and the $NPL_0$ method (dashed line). Dotted line, long-dashed line and solid line represents $NPL_\rho$ with correlation $\rho$=0.1, 0.3 and 1, respectively.*

## 2.4 Application to breast cancer data

We applied the new method to breast cancer ASP with known age at onset. For this analysis we used the ASP of the original set of 55 high-risk Dutch breast cancer families without any mutations in BRCA1 and BRCA2 described by Oldenburg et al. (2008). They found evidence for linkage at chromosome 9 around 82 cM. The question is whether taking into account age at onset will increase evidence for linkage. Data on 20 microsatellite markers at chromosome 9 were available. For the sibling pairs IBD status was estimated using MERLIN software (Abecasis et al., 2002) and the variances of the IBD status were estimated by simulations (Abecasis et al., 2002). Figure 2.2(left) shows the weight of the $NPL_\rho$ method. The proposed method weights the early-onset siblings (both affected before the age of 30) 6 times more than ASPs with "mean" age at onset. We applied the standard NPL method (NPL), $NPL_0$ and the proposed score test ($NPL_\rho$).

Taking into account the cumulative hazard ($NPL_0$) increases the maximum LOD score with respect to the NPL method from 2.9 to 3.0 (p=0.0001). Taking into account the dependence between siblings ($NPL_\rho$) further increases the maximum LOD score to 3.6 (p=0.000002) (Figure 2.5). Adjusting for age at onset increases evidence for linkage at chromosome 9 around 82 cM.

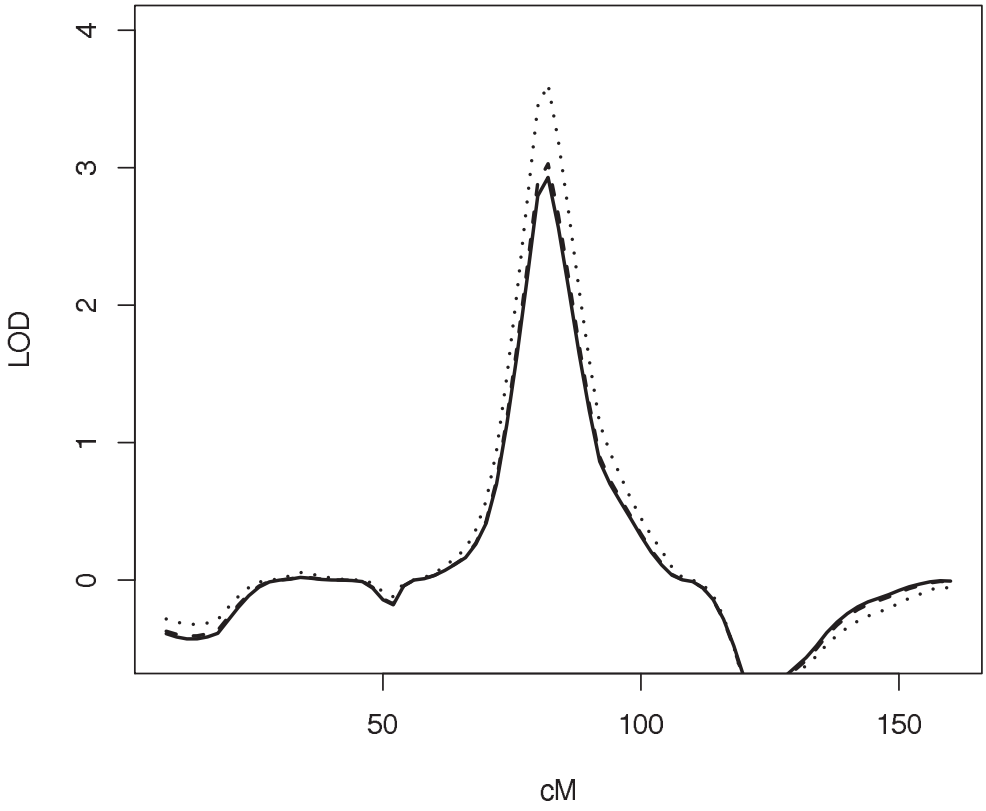


**FIGURE 2.5:** *Results of genetic linkage analysis of breast cancer data for chromosome 9. Solid line, dashed line and dotted line represent the unweighted NPL method, the $NPL_0$ method and the $NPL_\rho$ method, respectively.*

## 2.5 Discussion

We have proposed a new score statistic for age at onset linkage analysis for selected sibling-pairs. One of the main advantages of the method is that it can use marginal information known from previous family (twin) studies. Using this information, the proposed approach allows testing for age at onset linkage in affected-sibling pairs. Another advantage of this method with respect to likelihood-ratio approaches (Jonker et al., 2009; Li and Zhong, 2002) is that, under the null hypothesis of no linkage, the mean of the statistic is null. Hence it maintains the correct type I error for an arbitrary choice of population parameters. Thus if "known" population parameter values are not optimal the test loses power but it maintains the correct type I error.

Simulation results showed that our statistic is not very sensitive for the specified frailty parameters. We compared the proposed score statistic $\text{NPL}_\rho$ with the classical NPL test, with a test inspired by the Commenges tests and adapted for selected samples, $\text{NPL}_0$, and with a score test derived from the Li and Zhong (2002) model ($\text{NPL}_{\rho(\rho=1)}$). We simulated data with a gamma distributed random effect and with a dominant major gene effect. When population frailty parameters are approximately known the proposed methods outperforms the other methods. In fact, the expected gain in power of $\text{NPL}_\rho$ compared to $\text{NPL}_0$ and to the unweighted NPL method is about 5% and 10%, respectively. The $\text{NPL}_\rho$ and the $\text{NPL}_{\rho(\rho=1)}$ methods are equivalent when the true value of the marginal correlation is high. But, when the marginal correlation is small, $\text{NPL}_\rho$ performs better. When the frailty parameters are unknown, $\text{NPL}_\rho$ should not be used because when the difference between the used and the marginal parameters is large $\text{NPL}_\rho$ loses power with respect to NPL. In this case $\text{NPL}_0$ should be applied. In fact, the gain in power of $\text{NPL}_0$ compared to NPL method is about 5-10%. By means of simulations, we also showed the robustness of the proposed method with respect to the uncertainty on the marginal survival function. In fact the survival function has been estimated by Kaplan-Meier in random samples, and we never used the true value. In addition, data from registries are often based on large samples.

We applied the new score test to breast cancer data, where it increased evidence for linkage with respect to the standard NPL method and with respect to $\text{NPL}_0$. Since the frailty parameter values of the Dutch population are unknown, we used the frailty parameter values estimated in the Swedish population. Note that when the frailty parameters of the two populations are similar, the power of the test is maximized.

A number of extensions to the proposed model can be considered. For sake of notation we only considered sib-pairs but the extension to any kind

of relative-pairs is straightforward. In fact it can easily be proved that the score test for selected relative pairs is the same as derived in this paper, where the mean of the IBD ($E\pi$) and the marginal correlation depend on the genetic distance between relatives. For example, suppose the variance components have been estimated in a previous twin studies ($\hat{a}^2, \hat{c}^2, \hat{\sigma}^2$). In this case the score test for relative pairs is given by (2.9) where $E\pi = 2\Psi$, and the marginal correlation between relatives in the weight function is given by $\rho_\Psi = 2\Psi\hat{a}^2 + \hat{c}^2$ where $\Psi$ is their kinship coefficient. The proposed method can also deal with person-specific covariates for which known population incidences are stratified. For example, gender-specific incidences can be used. It is more complex to consider continuous covariates. We described the method using the common ASP design, but the method can include discordant and unaffected relative pairs. The extension of the model to general pedigrees can be done using a pairwise likelihood approach (Lindsay, 1998). We are currently exploring this approach. In this paper we assumed additive genetic effects. It is not straightforward to extend the additive gamma frailty model to model more complex genetic effects. An intuitive way to model dominance in sib-pairs is by using the marginal bivariate survival (2.6) with the parametrization of the correlation derived by Tang and Siegmund (2001) $\rho_\pi = (\pi - 1/2)\gamma - (I(\pi = 1/2) - 1/2)\varrho + \rho$ where $I()$ is the indicator function and $\varrho$ is the dominance effect. The linkage effect is modeled by two parameters $(\gamma, \varrho)$ with a joint null hypothesis $(\gamma, \varrho)=(0, 0)$. Detailed investigation on this case is one direction for future research. The model can also be extended to include frailties due to multiple disease loci (Jonker et al., 2009; Zhong and Li, 2002).

It is interesting to relate our approach to other published frailty methods for age at onset. Li and Zhong (2002) used a retrospective likelihood which depends on information (the baseline hazard and the shared residual effect) that have to be estimated together with the linkage effect. This approach cannot yield unbiased population-based parameter estimates. In fact, this is possible only through proper construction of the ascertainment-adjusted likelihood (Epstein et al., 2002; Sun and Li, 2004), but the probability of the sampled families are often unknown and hard to model. Recently, Jonker et al. (2009) proposed a correlated gamma frailty model for linkage similar to the one described in this paper. We used a different parametrization (Tang and Siegmund, 2001) which depends on marginal frailty parameters, which are potentially known from previous twin studies. In contrast to these papers we derived a score test instead of a likelihood-ratio test. Although the likelihood-ratio can have greater power for strong effects, the score test has the advantage to be always correct under the null hypothesis. Furthermore, it is more robust and simpler because it is computed under the null hypothesis.

Another approach to take into account age at onset is the conditional-logistic model implemented in LODPAL (Olson, 1999). Here age at onset is included as a covariate. Our approach offers several advantages above this model. First, it can be applied to unaffected samples, which are considered censored at their current age. Secondly, no arbitrary pairwise "covariates" has to be defined, in fact there is no theory justifying the use of the sum of the age at onset. Thirdly, our method is a score test. Note that the likelihood-ratio approach (like LODPAL) can give extraordinary large LOD scores due to instability of the maximization of the pseudo-likelihood (Schaid et al., 2007). Finally, the proposed method is more powerful because no degrees of freedom are spent to adjust for age at onset. In contrast, for LODPAL the degrees of freedom increase when covariates (age at onset) are considered. In fact, our score test is distributed as a 50:50 mixture of $\chi^2_0$ and $\chi^2_1$, while the LODPAL approach, adjusting for age at onset as a covariate is distributed as a 50:50 mixture of $\chi^2_1$ and $\chi^2_2$. We applied LODPAL approach to the breast cancer data. Probably due to the increment of the degrees of freedom, adjusting for the sum of the age at onset did not increase the evidence of linkage (data not shown).

We used additive gamma frailty models. These models have some drawbacks, first the additivity of the effects is mathematically appealing but it is not intuitive because in the individual hazard they act in a multiplicative way. Secondly only positive correlation can be taken into account. In order to deal with these issues a multivariate survival model with log-normal distribution has been proposed (Pankratz et al., 2005). However, these methods cannot be computed directly and Laplace approximations are needed. The computational demands may be high and convergence is not always ensured. More importantly, the log-normal frailty model cannot be formulated in terms of marginal parameters. In fact the baseline hazard and the residual additive effect have to be estimated, which is not straightforward for selected samples.

In summary, we have derived a new score test for age at onset linkage analysis. The proposed method is simple and is valid for any kind of ascertained samples. It is robust and computationally fast. When the marginal frailty parameters are known from previous twin studies, the proposed $\text{NPL}_\rho$ approach is more powerful than standard nonparametric linkage approaches. When the frailty parameters are unknown, we recommend to use $\text{NPL}_0$ which is less powerful than $\text{NPL}_\rho$ but has still more power than NPL.

A collection of compiled C++ programs which implements the proposed score test is available from our web site (http://www.msbi.nl/Genetics). The software uses Merlin (Abecasis et al., 2002) to compute the probabilities of IBD and to estimate the variance of the IBD by simulations.

## Appendix

### 1. Definition of the linkage random effect

Following Li and Zhong (2002) we define the component of the frailty due to the linkage effect by the sum of two random effects one inherited from the mother and the other from the father. Specifically, the component due to the causal gene can be modeled by $Z_{jl} = \sum_{j=1}^{4} a_{jk} U_{lk}$ $j = 1, 2$ where

$$a_{jk} = \begin{cases} 1, & \text{if } j \text{ has allele } k; \\ 0, & \text{if } j \text{ does not have allele } k \end{cases}.$$

$U_{lk} \sim \Gamma(\nu_l/2, 1/\sigma^2)$ is the effect of the $k$th allele of the parents ($U_{l1}$ and $U_{l2}$ represent the two alleles of the mother and $U_{l3}$ and $U_{l4}$ represent the alleles of the father).

### 2. Bivariate likelihood conditioned on $\pi$

The likelihood of two individuals with $\pi$ proportion of alleles shared IBD is given by

$$P(\mathbf{y}|\pi;\gamma) = P(t_1, \delta_1, t_2, \delta_2|\pi;\gamma) = (-1)^{\delta_1+\delta_2} \frac{\partial^{\delta_1+\delta_2}}{\partial t_1^{\delta_1} \partial t_2^{\delta_2}} S_{12}(t_1, t_2|\pi;\gamma),$$

where $S_{12}(t_1, t_2|\pi;\gamma)$ is given by the equation (2.6).

### 3. Weight function of the score statistic

The weight function of the proposed score statistic is given by

$$\begin{aligned} \ell_\rho(\gamma_0) =& \frac{\partial \log P(\mathbf{y}|\pi;\gamma_0)}{\partial \rho} = H_1 + H_2 - \frac{\log(K)}{\sigma^2} + \\ & \delta_1(1-\delta_2)\frac{k_1 - 1}{\theta_1} + \delta_2(1-\delta_1)\frac{k_2 - 1}{\theta_2} + \\ & \delta_1\delta_2 \frac{(k_2-1)\theta_1 + (k_1-1)\theta_2 + \sigma^2 k_1 k_2}{\theta_1\theta_2 + \rho\sigma^2 k_1 k_2}, \end{aligned} \tag{2.10}$$

where $H_j = \int_0^{t_j} h(t)dt$ is the marginal cumulative hazard of the $j$th individual; $K = (\exp(H_1\sigma^2) + \exp(H_2\sigma^2) - 1)$, $k_j = \exp(H_j\sigma^2)/K$, and $\theta_j = 1 + \rho(k_j - 1)$.