

Proceedings

Open Access

## New score tests for age-at-onset linkage analysis in general pedigrees

Andrea Callegaro\*, Hae-Won Uh, Quinta Helmer  
and Jeanine J Houwing-Duistermaat

Address: Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands Medical Center, PO Box 9600, 2300RC Leiden, The Netherlands

E-mail: Andrea Callegaro\* - a.callegaro@lumc.nl; Hae-Won Uh - h.uh@lumc.nl; Quinta Helmer - q.helmer@lumc.nl; Jeanine J Houwing-Duistermaat - j.j.houwing@lumc.nl

\*Corresponding author

from Genetic Analysis Workshop 16  
St Louis, MO, USA 17-20 September 2009

Published: 15 December 2009

BMC Proceedings 2009, 3(Suppl 7):S97 doi: 10.1186/1753-6561-3-S7-S97

This article is available from: <http://www.biomedcentral.com/1753-6561/3/S7/S97>

© 2009 Callegaro et al; licensee BioMed Central Ltd.

This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Abstract

Our aim is to develop methods for mapping genes related to age at onset in general pedigrees. We propose two score tests, one derived from a gamma frailty model with pairwise likelihood and one derived from a log-normal frailty model with approximated likelihood around the null random effect. The score statistics are weighted nonparametric linkage statistics, with weights depending on the age at onset. These tests are correct under the null hypothesis irrespective of the weight used. They are simple, robust, computationally fast, and can be applied to large, complex pedigrees. We apply these methods to simulated data and to the Genetic Analysis Workshop 16 Framingham Heart Study data set. We investigate the time to the first of three events: hard coronary heart disease, diabetes, or death from any cause. We use a two-step procedure. In the first step, we estimate the population parameters under the null hypothesis of no linkage. In the second step, we apply the score tests, using the population parameters estimated in the first step.

### Background

It is well known that heterogeneity results in loss of statistical power when studying genetic factors of complex genetic diseases. To deal with heterogeneity additional data such as covariates (e.g., age at onset, known genetic factors) are collected. In this paper we are interested in adjusting linkage for age at onset.

Frailty models have been proposed for age-at-onset linkage analysis [1-5]. Gamma frailty models are particularly attractive because the gamma-distributed random effect can be easily integrated out and it allows the use of observable marginal survival functions [1-4]. A drawback of these models is that their corresponding likelihood becomes very complex for large pedigrees.

To solve this problem, we propose a score test based on a composite likelihood [6].

A second model for multivariate survival data is the log-normal frailty model. Using this model, Pankratz et al. [5] proposed a likelihood-ratio approach for linkage. In the spirit of Lebec and Houwelingen [7], we derive a robust and simpler score test, using an approximation of the likelihood around the null random effect.

**Methods**

**Gamma frailty model: pairwise likelihood approach**

Let  $T_{ij}$  be the random variable of age at onset for relative  $j$  in family  $i$ ,  $i = 1, \dots, N$ . Let  $(t_{ij}, d_{ij})$  be the observed data where  $t_{ij}$  is the observed age at onset if  $d_{ij} = 1$  and age at censoring if  $d_{ij} = 0$ . The conditional hazard for individual  $j$  in family  $i$ , with covariates  $x_{ij}$  and random effect  $Z_{ij}$ , is given by  $\lambda(t_{ij} | x_{ij}, Z_{ij}) = \lambda_0(t_{ij} | x_{ij})Z_{ij}$ . Without loss of generality, we assume that  $E[Z] = 1$ . The baseline hazard  $\lambda_0(t)$  is the hazard for  $x = 0$  and  $Z = 1$ . The frailty  $Z$  is decomposed into the sum of independent gamma distributed effects, namely a linkage effect, a residual additive effect, and a non-shared environment effect. The scale parameter is common to all of the effects and is defined as the sum of the shape parameters. When the proportion of alleles shared identically by descent (IBD) for a relative pair  $(j, k)$  is known  $(\pi_{jk})$ , the marginal bivariate survival function can be derived from the additive gamma frailty model [4]. The bivariate survival function depends on the marginal survival functions, on the variance of the random effect  $(\sigma_G^2)$ , and on the pairwise correlation. The correlation  $\rho_{jk}(\pi_{jk}) = (\pi_{jk} - E\pi_{jk})\gamma + \rho_{jk}$  depends on the IBD through the linkage parameter  $\gamma$ . Under the null hypothesis  $(H_0: \gamma = \gamma_0 = 0)$ , the correlation is equal to the correlation in the population  $(\rho_{jk})$ . The marginal correlation between the  $i^{th}$  and the  $j^{th}$  individual is a function of their expected proportion of alleles shared IBD,  $\rho_{jk} = a^2 E\pi_{jk}$ , where  $a^2$  is the portion of the variance explained by the total additive effect.

We use a retrospective likelihood [4] and, in order to deal with general pedigrees, we consider a pairwise likelihood approach [6]. For  $N$  families, the corresponding score statistic is a weighted nonparametric linkage (NPL) statistic

$$NPL = \frac{\sum_{i=1}^N \text{vec}(W_i)' \text{vec}(\pi_i - E\pi_i)}{\sqrt{\sum_{i=1}^N \text{vec}(W_i)' \text{var}_0(\pi_i) \text{vec}(W_i)}} \quad (1)$$

Here, elements of the weight matrix  $W$  are given by  $W_{jk} = \partial \log L_{jk}^\pi(\gamma_0) / \partial \rho_{jk}$ , where  $L_{jk}^\pi(\gamma) = P(\delta_j, t_j, \delta_k, t_k | \pi_{jk}, \gamma)$

is the prospective bivariate likelihood. The operator  $\text{vec}(A)$  places the  $n$  columns of the  $m \times n$  matrix  $A$  into a vector of  $mn \times 1$ . In the case of uncertain IBD status, the variance of the proportion of allele shared IBD  $(\text{var}_0(\pi_i))$  can be estimated by simulations. Note that the classical mean IBD test is a weighted NPL statistic [Eq. (1)] with weight equal to  $W_{jk} = d_j \times d_k$ .

**Log-normal frailty model**

Let  $d$ ,  $\Lambda_0$ , and  $V = \log Z$  be the  $n$ -dimensional vectors of the disease status, the baseline cumulative hazards at the observed age, and the normally distributed random effects of the  $n$  members of a particular pedigree, respectively. The random effect  $V$  follows a multivariate normal distribution with mean zero, and variance-covariance matrix  $\Sigma$  with elements  $\Sigma_{jk} = \sigma_N^2 \rho_{jk}(\pi_{jk})$ . The log-likelihood can be approximated by using a second-order Taylor approximation around  $V = 0$ . For small random effects and known baseline cumulative hazard, the vector of standardized martingale residuals behaves as a normal distribution. Integrating over the distribution of the random effect gives  $M = (d - \Lambda_0) / \Lambda_0 \sim N(0, \Sigma_1)$ , where  $\Sigma_1 = \Sigma + \text{diag}(1/\Lambda_0)$ . The score statistic derived from the retrospective likelihood is a weighted NPL statistic [Eq. (1)] with weight matrix  $W = \Sigma_1^{-1} M (\Sigma_1^{-1} M)' - \Sigma_1^{-1}$  and  $\Sigma_1$  taken in  $\gamma = 0$ . In this paper we approximate the baseline cumulative hazard with the marginal cumulative hazard.

**Materials**

**Estimation of the population parameters**

Three phenotype files were provided: Original Cohort participants, Offspring participants, and Generation 3 participants. We combined the three files and used this dataset as a random sample from the population. The total number of individuals considered was 6879. The number of disease-free survival events was 644 (248 coronary heart diseases, 385 diabetes, and 98 deaths), with prevalence around 10%. We estimated the marginal survival functions stratified by sex using the Kaplan-Meier estimator. By age 60 years, 20% of males and 10% of females were affected. Using these estimated survival functions we fitted a marginal pairwise correlated gamma frailty model. The sib-sib marginal correlation was  $\rho = 0.46$  and the variance estimated by the gamma frailty models was  $\sigma_G^2 = 0.93$ . The sib-sib marginal correlation was  $\rho = 0.5$  and the variance estimated by a log-normal frailty model [5] was  $\sigma_N^2 = 0.43$ .

**Pedigree data preparation**

In the Genetic Analysis Workshop (GAW) 16 Framingham Heart Study (FHS) data 765 pedigrees with 2 to 301 genotyped subjects were available. To simplify the IBD computation, large pedigrees were split into  $n = 1599$

nuclear families. The number of nuclear families with at least one affected sibling was  $n = 488$ . Only 46 nuclear families were available with at least two affected siblings.

**Single-nucleotide polymorphism (SNP) data selection**

The GAW16 Framingham dataset included 550 k SNP genotype data. Using the nuclear families with at least one affected individual (2275 individuals), we selected 15 k SNPs informative for linkage. First, markers with known physical position were selected (497 k). Second, 10 markers per centimorgan with minor allele frequency larger than 0.15 were considered (37 k). Finally, SNPs were simulated on 250 sib-pairs in order to select 15 k SNPs with the highest information content. The information content of the final set of SNP was around 85%.

**Simulated data**

To assess power and type I error rates, we simulated data using a frailty model with parameter values estimated in the GAW16 FHS data. The random effect was gamma-distributed with a mean of one and variance of  $\sigma_G^2 = 0.93$ . The baseline hazard was derived from the marginal hazard. The random effect was decomposed into the sum of three components: one locus-additive genetic effect (explaining 60% of the variability), one shared environmental effect (explaining 20% of the variability), and one unshared environmental effect. We simulated pedigrees with 15 members (Figure 1). Marker data were simulated far from any disease locus (null hypothesis) and close to the disease locus, which explains all the additive genetic variance (alternative hypothesis).

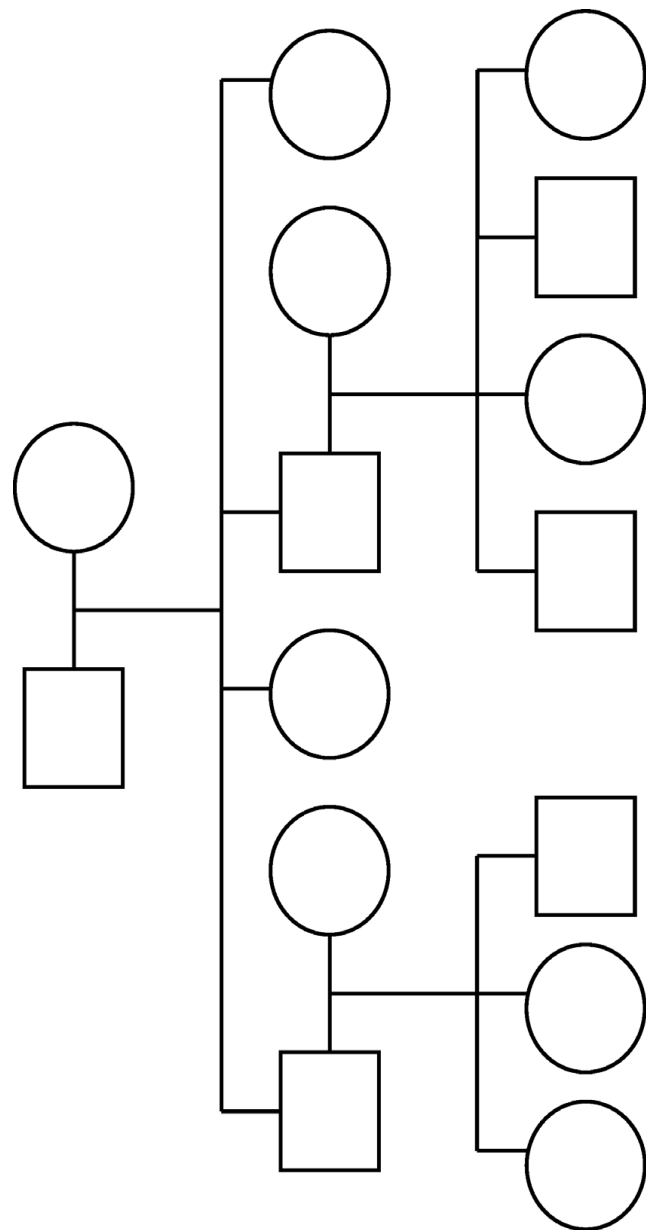
**Results**

**Simulated data results**

Table 1 shows the type I error rates based on 5000 replications and the power based on 1000 simulations, for sample size of 300 families with at least two affected siblings. On simulated data, the proposed methods have correct type I error rates. For our simulation settings, taking into account age at onset considerably increases the power to detect linkage. On a moderately sized pedigrees (15 members), the log-normal approach is more powerful than the pairwise gamma frailty approach.

**Application to the FHS dataset**

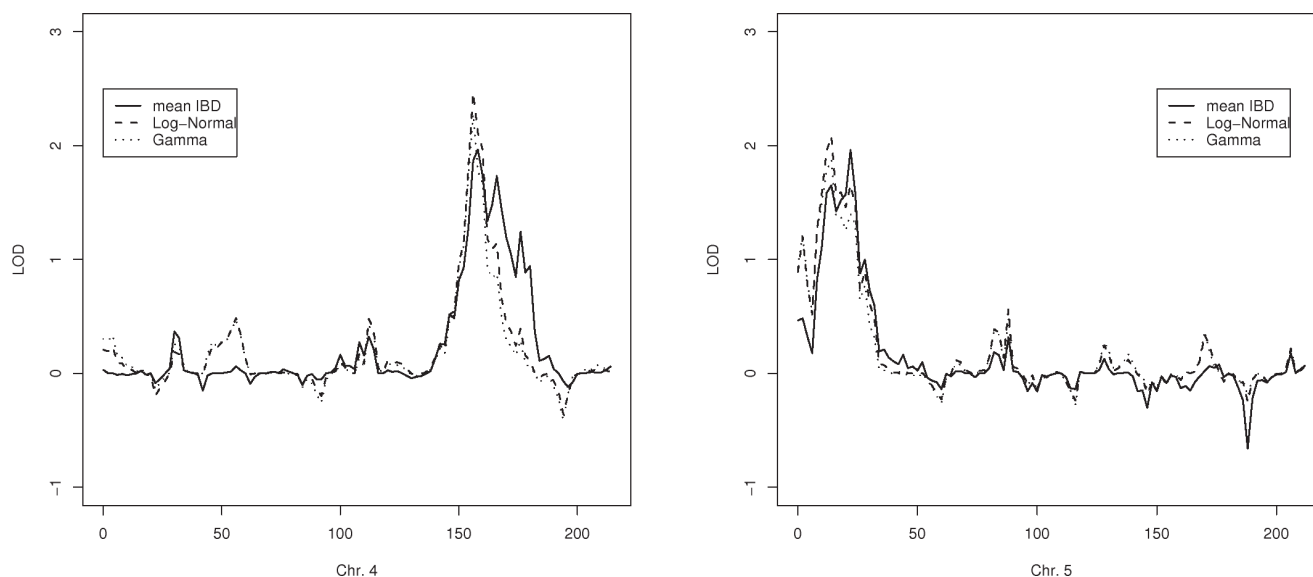
We performed a genome-wide linkage analysis using the unweighted NPL test (mean IBD test) with variance of the allele shared IBD estimated by simulations [8]. Figure 2 shows the two highest LOD scores (close to  $LOD = 2$ ), which are located on chromosomes 4 and 5, respectively.



**Figure 1**  
Pedigree structure of 15 individuals used for simulating data.

**Table 1: Estimates of type I error rates and power**

Method	Null hypothesis		Alternative hypothesis	
	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
Mean IBD	0.05	0.01	0.34	0.14
Gamma	0.05	0.01	0.94	0.80
Log-normal	0.05	0.01	0.98	0.85



**Figure 2**  
**Age at onset genetic linkage analysis of GAW16 FHS dataset.** LOD scores on chromosomes 4 (left) and on chromosome 5 (right).

We applied the proposed methods to the data of these two chromosomes. The linkage analysis was performed on all the nuclear families ( $n = 1599$ ), on the families with at least one affected siblings ( $n = 448$ ) and on the subset of families with at least two affected siblings ( $n = 45$ ). The maximum LOD-scores were obtained considering only families with at least two affected siblings. Figure 2 shows the results on this subset of families. On chromosome 4, adjusting for age at onset increases the maximum LOD score from 2 to 2.5. On chromosome 5, with the proposed methods the maximum LOD score is in a slightly different location (10 cM) with respect to the unweighted mean IBD test (25 cM). Results on chromosome 5 are replicated on the larger set of families with at least one affected sibling (data not shown).

### Discussion

In this paper we proposed two approaches for age-at-onset linkage analysis in general pedigrees. We applied the proposed methods to the GAW16 FHS data in two suggestive regions identified by the standard NPL method. The maximum LOD-scores were obtained analyzing only the set of families with at least two affected siblings. This can be due to the fact that affected individuals carry most of the information for linkage. On the densest pedigrees, adjusting for age at onset slightly increased the evidence for linkage. However, it is difficult to interpret the results because of the small number of events.

Because GAW16 FHS families were randomly selected, it was possible to estimate the marginal information

directly from the data. When marginal information is known from previous twin (family) studies, the proposed methods can be applied to ascertained families.

For the two identified regions, association analysis in the presence of linkage may be the next step. The proposed models can be easily extended to study association in the presence of linkage by including the genotype of the siblings as a covariate.

In this paper we computed IBD probabilities using MERLIN and we estimated the variance of the allele shared IBD using simulations [8]. Because this software can deal only with small to moderately large families, we split large families into nuclear families. An alternative approach is to estimate IBD probabilities using Markov-chain Monte Carlo methods, which now provide this information for general pedigrees. Sampled inheritance vectors can also be used to estimate the variance of the allele shared IBD in the denominator of the score statistic.

Software to apply the proposed methods is freely available [9].

### Conclusion

We proposed two new score tests for age of onset linkage analysis. Both methods are simple and can be applied to general pedigrees. Simulations showed that the proposed methods outperform the traditional affected-only NPL method. On the application to the GAW16 FHS data,

adjusting for age at onset slightly increased the interesting linkage peaks.

### List of abbreviations used

FHS: Framingham Heart Study; GAW: Genetic Analysis Workshop; IBD: Identical by descent; NPL: Nonparametric linkage; SNP: Single-nucleotide polymorphism.

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

AC participated in method development, carried out data analysis, and drafted the manuscript. QH carried out the SNP data selection. JJH-D participated in method development. H-WU acquired the data. All authors read, critiqued, and approved the final manuscript.

### Acknowledgements

The Genetic Analysis Workshops are supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences. JJH-D, AC, and QH are supported by a grant from the Netherlands Organization for Scientific Research (NWO 917.66.344). H-WU is supported by grants from IOP Genomics/SenterNovem (IGE05007).

This article has been published as part of *BMC Proceedings* Volume 3 Supplement 7, 2009: Genetic Analysis Workshop 16. The full contents of the supplement are available online at <http://www.biomedcentral.com/1753-6561/3?issue=S7>.

### References

1. Commenges D: **Robust genetic linkage analysis based on a score test of homogeneity: the weighted pairwise correlation statistic.** *Genet Epidemiol* 1994, **11**:189–200.
2. Jonker MA, Bhulai S, Boomsma DI, Ligthart RSL, Posthuma D and Vaart Van Der AW: **Gamma frailty model for linkage analysis with application to interval-censored migraine data.** *Biostatistics* 2008, **10**:187–200.
3. Houwing-Duistermaat JJ, Callegaro A, Beekman M, Westendorp RG, Slagboom PE and Van Houwelingen JC: **Weighted statistics for aggregation and linkage analysis of human longevity in selected families: The Leiden Longevity Study.** *Stat Med* 2009, **28**:140–151.
4. Callegaro A, Van Houwelingen JC and Houwing-Duistermaat JJ: **Score test for age at onset genetic linkage analysis in selected sibling-pairs.** *Stat Med* 2009, **28**:1913–1926.
5. Pankratz VS, de Andrade M and Therneau TM: **Random-effects Cox proportional hazards model: general variance components methods for time-to-event data.** *Genet Epidemiol* 2005, **28**:97–109.
6. Lindsay B: **Composite likelihood methods.** *Contemp Math* 1998, **80**:221–239.
7. Lebrech J and van Houwelingen HC: **Score test for linkage in generalized linear models.** *Hum Hered* 2007, **64**:5–15.
8. Abecasis G, Cherny S, Cookson W, Cardon L and Blangero J: **Merlin—rapid analysis of dense genetic maps using sparse gene flow trees.** *Nat Genet* 2002, **30**:97–101.
9. **Score Test for Age at Onset Linkage Analysis.** <http://www.msbi.nl/Genetics/Software>.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:  
[http://www.biomedcentral.com/info/publishing\\_adv.asp](http://www.biomedcentral.com/info/publishing_adv.asp)

