
Universiteit
Leiden
The Netherlands

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

Callegaro, A.

Citation
Callegaro, A. (2010, June 17). *Using survival data in gene mapping : using survival data in genetic linkage and family-based association analysis*. Retrieved from https://hdl.handle.net/1887/15696

Version:	Corrected Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/15696

Note: To cite this publication please use the final published version (if applicable).

Chapter 5

Weighted statistics for aggregation and linkage analysis of human longevity in selected families: The Leiden Longevity Study

Abstract

Typically long-lived sibling pairs have been collected for linkage analysis of human longevity and information on life span of first degree relatives is available to assess familial aggregation of life span. We propose a weighted statistic for aggregation analysis which tests for a relationship between a family history of excessive survival of the sibships of the long lived pairs and the survival of their parents and their offspring. For linkage analysis, we derive a weighted score statistic from a simple gamma frailty model which assigns more weight to excessive long lived pairs. We apply the methods to data from the Leiden Longevity Study which consists of sibling pairs of age 90 years or above and their first degree relatives. The pairs have been genotyped for microsatellite markers in a candidate region. Association was present between survival within the sibships and survival of the offspring, but not with the parental generation. For linkage analysis, weighting increased the value of the test statistic, but the result was not statistically significant. About the methods we conclude that the statistic for aggregation provides insight in clustering of life span and the statistic for linkage provides a tool to include demographic information into the analysis.

This chapter has been published as: J. J. Houwing-Duistermaat, A. Callegaro, M. Beekman, R.G. Westendorp, P.E. Slagboom and J.C. van Houwelingen (2009). Weighted statistics for aggregation and linkage analysis of human longevity in selected families: The Leiden Longevity Study. *Statistics in Medicine 28(1)*, 140–151.

5.1 Introduction

Understanding the underlying biology of life span is an important challenge in life science. To reveal its genetic basis, families displaying exceptional longevity have been studied. For example Perls et al. (1998) analyzed survival in siblings of centenarians and Schoenmaker et al. (2006) analyzed survival in relatives of nonagenarian sibling pairs. Puca et al. (2001) performed a linkage scan in siblings pairs both older than 90 years and one of them older than 98 years. The necessity of these selection criteria were confirmed by a simulation study of Tan et al. (2004). They showed that to map a rare dominant genetic variant that reduces hazard of death by half, a large sample with at least one extremely long lived sibling (above 98 years) in each pair is needed when affected sibling pair methods are used. In Dutch and European studies on longevity (Franceschi et al., 2007; Schoenmaker et al., 2006) however, the selection criteria for sibling pairs are less stringent. Sibling pairs of age above 90 years will be genotyped for linkage. To improve efficiency, one may want to use a weighted score statistic (Hsu et al., 2002; Kruglyak et al., 1996; Whittemore, 1996). Here the weight should depend on the amount of excessive survival of a family.

Li and Zhong (2002) proposed gamma frailty models for linkage analysis of survival data. The correlation due to a trait locus at a certain position at the genome is modelled by a random effect. These models have been used for linkage analysis of age at onset data, i.e. for data containing subjects who experienced a particular event. For genetic analysis of human longevity however, the subjects need to be alive and for all siblings the outcome of interest (age at death) is censored. To apply the survival models to longevity data, population based information on mortality has to be used to standardize the age at entry of the siblings. For most European countries, life tables are available. Li and Zhong (2002) used a likelihood ratio statistic for testing. We prefer a score statistic since this statistic is robust against model deviations.

Before linkage analysis is performed, aggregation of the trait within families may be assessed. Clustering of an outcome within families can be studied by testing for the presence of a relationship between the outcome of an individual and a family history score based on the outcomes of the relatives (Khoury et al., 1994). For binary data, Houwing-Duistermaat and van Houwelingen (1998) derived a family history measure which is equal to a weighted sum of the observed number of cases in the family minus its expectation. The weights are used to take into account the different relationships within the family. In this paper, this approach will be adapted for observations on human longevity.

The goals of this paper are to derive and apply methods for aggregation of life span in families selected for excessive survival and to test for linkage

of longevity taking into account the amount of excessive survival of the siblings. Our study is motivated by the Leiden Longevity Study, in which sibling pairs with ages above 90 years were ascertained. Data were available on date at birth and if applicable on date at death of the parents, siblings and offspring of these nonagenarians. Schoenmaker et al. (2006) showed that these relatives of the long lived siblings live longer than their Dutch birth cohort and conclude therefore that they collected a set of families that show familial enrichment for excessive survival. Since around 10% of the birth cohort of the long lived siblings achieve the age of 90 years, the question can be raised whether inter family variation in life span within this study is present. We will test for a relationship between a family history based on survival within the selected sibships and the survival of the parents and offspring of the long lived sibling pairs. Also genetic data to perform linkage analysis is available in the Leiden Longevity study. Beekman et al. (2006) replicated the linkage analysis of Puca et al. (2001) in the Leiden Longevity study. A standard affected sibling pair method was used. From a simple gamma frailty model, we will derive a weighted score statistic which assigns more weight to sibling pairs who show excessive survival. For both methods (testing for aggregation and linkage), Dutch life tables will be used to standardize the current age or age at death.

5.2 Methods

Data description

The families studied in this paper are a subset of the Leiden Longevity Study. The design is described in detail by Schoenmaker et al. (2006). Briefly families participating in the Leiden Longevity Study have at least two siblings meeting four inclusion criteria: (1) men are aged 89 years or above and women are aged 91 years or above, (2) subjects have at least one living brother or one living sister who fulfils the first criterion and is willing to participate, (3) the nonagenarian sib ship has an identical mother and father, (4) the parents of the nonagenarian sib ship are Dutch and Caucasian. Note that for selection different age cut off points for males and females were used because of differences in life expectancies. For the present study, we had available 368 nonagenarian subjects belonging to 166 sib ships (age range from 89 to 103, mean age 94.1). These nonagenarian siblings are called participants in this paper. The number of participants per sibship varied from 2 to 4 participants. In addition we had information on the age at entry or ages of death of 1317 offspring, of 330 parents, and of 881 siblings of the participants. An example of a family of the Leiden Longevity study is given in figure 1.

Ascertainment of the families in the Leiden Longevity Study, depends on

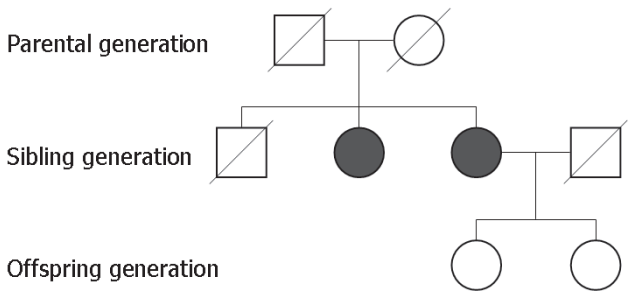

FIGURE 5.1: *Example of family from Leiden Longevity Study*

the sizes of the sibships and the distribution of life span within the sibship (Shute and Ewens, 1988). The sizes of the sibships (participants and siblings of participants) vary from 2 to 17 siblings with a mean of 7.5 siblings. Among the siblings of the two largest sibship of size 16 and 17 subjects, 10 respectively 9 subjects are deceased, 4 respectively 5 are alive but too young to be a participant, and 3 respectively 2 are participants. Modelling ascertainment of these sibships is complicated since for the Dutch population no information is available on joint mortality within sibships. Therefore we restrict ourselves to study excess survival in the parents and the offspring of the participants and the relationship between their survival distributions and the survival of the selected sibship. Because ascertainment of the parents depends on the number of their offsprings, we conditioned on the age of the parents at the birth of their last child. The mean age at birth of their last child was 38 (range of 24 to 48) and 41 (range of 26 to 59) for mothers and fathers respectively.

For the Dutch population, sex specific life tables are available with the percentages of death for each year of age in the range of 0 to 100 years and each birth year since 1860. We extended these tables to birth years from 1852 by using the numbers of 1860 for birth year 1852 to 1860 and to ages of 112 years by using the numbers of 100 years. From these tables, we computed for each individual the sex and birth year specific cumulative hazard.

For linkage analysis, subjects from 160 sibships were genotyped for the same six micro satellite markers at which initially linkage was detected by Puca

et al. (2001) namely D4S2964, D4S1534, D4S414, D4S1572, D4S406, D4S402 with an average inter-markers spacing of 7.22 cM. In total 215 pairs belonging to 154 sib ships were successfully genotyped (from 6 families only one subject was successfully genotyped). The heterozygosity of the markers varied from 74.6 to 89.6 in this study. Beekman et al. (2006) applied the likelihood ratio method of Kong and Cox (1997) for linkage analysis on this set. The obtained maximum of the lod score is -0.02 for the whole sample and 0.48 for the subset of sib ships with at least one sibling above 98 years of age (21 sib ships).

Testing for excess survival

To quantify the age at disease onset distribution within a family, Hille et al. (1997, 1999) proposed to use the family specific standard mortality ratio (SMR). For families selected for excessive survival, we compute three SMR's to describe life span distributions within a family, namely one for the sibling generation on which selection was based (SMR^s), one for the parental generation (SMR^p)and one for the offspring generation (SMR^o). Let D_{ij} be 1 if subject j of family i is deceased and 0 otherwise and let H_{ij} be the sex and birth cohort specific cumulative hazard for subject j of family i. Then for generation l, SMR$_i^l$ of family i is defined as follows

$$\text{SMR}_i^l = \frac{\sum_j^{n_i} D_{ij}}{\sum_j^{n_i} H_{ij}}, \tag{5.1}$$

with n_i the number of subjects in generation l. Note that due to selection on excessive survival of the sibling generation, SMR^s will be smaller than one.

For the parental and offspring generations, one may test for excessive survival (SMR<1). Ignoring the correlation between family members, the likelihood function for parents or offspring is equal to

$$L(\lambda|D_{ik}, H_{ik}) = \prod_i \prod_k (H_{ik}\exp(\lambda))^{D_{ik}} \exp(-H_{ik}\exp(\lambda)), \tag{5.2}$$

with (D_{ik}, H_{ik}) the data on the parents or on the offspring, respectively. The parameter λ models deviation of the survival from the corresponding Dutch birth cohort, i.e. $\lambda < 1$ represents excess survival. The score statistic U to test the null hypothesis $\lambda = 1$ versus the alternative $\lambda < 1$ is given by

$$U = \sum_i \sum_k^{n_i} (D_{ik} - H_{ik}), \tag{5.3}$$

with n_i the number of relatives (parents or offspring) in family i. The variance

of this statistic can be empirically estimated by

$$\mathrm{Var}(U) = \sum_i (\sum_k^{n_i} (D_{ik} - H_{ik}))^2.$$

To test if excess survival in the parents and offsprings depends on the excess survival in the sibship generation, the model may be extended by letting the parameter λ depend on a family history score describing survival in the sibship. A measure for excess survival of the sibship is the sum of the martingale residuals

$$\mathrm{sumMR}_i^s = \sum_j^{n_i} (D_{ij} - H_{ij}). \tag{5.4}$$

The sumMR_i^s measures the deviation in survival within the sibship generation of family i from the mean survival in the corresponding birth cohort. Since the genetic distance between children and their parents is smaller than the genetic distance between children and their aunts and uncles, for the offspring generations the survival of aunts and uncles should obtain less weight than the parents of the offspring (see Houwing-Duistermaat and van Houwelingen (1998)). A straightforward family history measure x_{ik} for relative k of family i is the sum of the kinship coefficient Γ_{jk}^i between relative k and the sibling j times the sibling's martingale residual $(D_{ij} - H_{ij})$:

$$x_{ik} = \sum_j^{m_i} \Gamma_{jk}^i (D_{ij} - H_{ij}),$$

with m_i the number of siblings. Now replace λ in likelihood (5.2) by $\lambda_{ik} = \theta x_{ik}$, then the statistic $U(x_{ik})$ corresponding to this parametrization is given by

$$U(x_{ik}) = \sum_i \sum_k^{n_i} x_{ik} (D_{ik} - H_{ik}). \tag{5.5}$$

The variance of this weighted statistic can be computed analogously to the variance of the unweighted statistic.

Linkage analysis

In the former section on aggregation analysis, the whole sibling generation was analyzed. In this section, we only consider long lived sibling pairs. Let for sibship i, $\hat{\pi}_i$ be a vector with as elements the proportions alleles shared identical by descent (IBD) by the long lived sibling pairs at a certain position at the genome. This proportion is usually estimated from the markers located in the

surrounding region using a multipoint approach. Then for additive effects, the score statistic Z (Kruglyak et al., 1996) to test for linkage is given by:

$$\hat{Z} = \frac{\sum_i w_i'(\hat{\pi}_i - \frac{1}{2})}{\sqrt{\sum_i w_i' var_0(\hat{\pi}) w_i}}, \tag{5.6}$$

with w_i a vector of known weights of the same lengths as $\hat{\pi}_i$. When the proportions of alleles shared IBD are observed, the covariances between π_{il} and π_{ik} of sibling pair l and k of family i are zero and the variance of Z is equal to $1/8$. For incomplete marker informativeness, the variance $var_0(\hat{\pi})$ has to be computed using multipoint simulations (Kong and Cox, 1997). Information available in the ages of the siblings at entry of the study can be incorporated via the weights (Hsu et al., 2002; Kruglyak et al., 1996; Whittemore, 1996).

To derive appropriate weights, we propose to use a simple gamma frailty model for the current ages. Let Y be the shared frailty for siblings 1 and 2. The marginal survival functions S_1 and S_2 are given by $L^{-1}(H_1)$ and $L^{-1}(H_2)$ respectively, with L the Laplace transformation corresponding to the distribution of Y and H_1 and H_2 the marginal cumulative hazards. The marginal bivariate survival function is given by the Laplace transformation L of $(L^{-1}(S_1) + L^{-1}(S_2))$. When Y follows a gamma distribution $(\frac{1}{\delta}, \frac{1}{\delta})$ the marginal bivariate survival function is given by $S_{12} = (\exp(\delta H_1) + \exp(\delta H_2) - 1)^{-\delta^{-1}}$ (Hougaard, 2000). Hence the log likelihood function for this sibling pair is given by

$$\begin{aligned} l(\delta|H_1, H_2) &= -\delta^{-1} \ln(\exp(\delta H_1) + \exp(\delta H_2) - 1) \\ &\approx -(H_1 + H_2 - \delta H_1 H_2 + 0.5\delta^2(H_1^2 H_2 + H_1 H_2^2)), \end{aligned}$$

where the last step is obtained by a second order Taylor expansion around $\delta = 0$. Note that in this formula we used the fact that all siblings are alive. Now a first order Taylor approximation of the score function for the covariance parameter δ is given by

$$U(\delta) = H_1 H_2(1 - \delta(H_1 + H_2)). \tag{5.7}$$

Now when a locus for longevity is linked to the position at which the IBD status was estimated, the correlation tends to increase with the amount of alleles shared IBD. The following parametrization for the correlation given the IBD $(\delta(\pi))$ can be used $\delta(\pi) = \delta_{pop} + \gamma(\pi - \frac{1}{2})$ with δ_{pop} the correlation in the general population (Tang and Siegmund, 2001). Similar to the optimal Haseman-Elston statistic, the optimal test for this model is obtained by regression of $\hat{\pi}$ on $U(\delta_{pop})$ (Haseman and Elston, 1972; Lebrec et al., 2004; Tang and Siegmund, 2001). For the Dutch population δ is unknown. We use $\delta = 0$ in (5.7) and the

numerator of the score statistic is as follows

$$\hat{U} = \sum_i \sum_{lk} H_{il} H_{ik} (\hat{\pi}_{i,lk} - \frac{1}{2}) \tag{5.8}$$

5.3 Results

Aggregation analysis in the Leiden Longevity Study

The unweighted statistic (5.3) was applied to the data on the parents and the offspring. These results agree with the results of Schoenmaker et al. (2006), namely parents and offspring live significantly longer than their birth cohorts ($P < 0.0001$). To describe the age distribution in the families of the Leiden Longevity study, we computed family specific SMR^p, SMR^s, SMR^o (5.1) for the parents, the participants and their siblings and the offspring respectively. The median and range of these SMRs are given in table 1. As expected for all families the sibling specific SMR^s were smaller than 1. Twelve sibships had a standard mortality ratio of 0, i.e. all siblings were alive in these sibships. The median of the parental and offspring specific SMRs were also smaller than 1, indicating an excess survival in these generations. About 50% (n=84) of the families had an offspring specific SMR^o of 0.

In figure 2 the martingale residuals for the participants and their siblings are given (range -7.54 to 0.97, mean of -1.27). Six subjects had a martingale residual smaller than -6. These subjects were alive, over 100 years of age and members of six different families. Among these subjects, there were five participants and one sibling of a participant; five males and one female. The oldest subject is female and 103 years of age. The histogram also shows that in the selected sibships most of the siblings who satisfy the selection criteria are participants i.e. siblings of participants either died relatively young or are too young to be a participant. For the parental generation 25% died before 67 years while for the offspring and sibling generations the first quartiles were 4 and 56 years respectively.

The mean and standard deviation of the sumMR^p, sumMR^s and sumMR^o (5.4) and their correlations with sumMR^s are also given in table 1. The correlation between the sums of the sibling and the parental generation appears to be small (cor=0.02). With the offspring generation some correlation exists (cor=0.25). Finally we computed for each offspring and parent the covariate x_{ij} and applied the weighted statistic. For the parents the weighted score statistics did not improve the significance level of the unweighted statistic (5.5). Also the separate analyzes of the data on fathers and mothers, showed no improvement in significance. For the offspring weighting reduced the standardized statistic from -4.70 to -5.12.

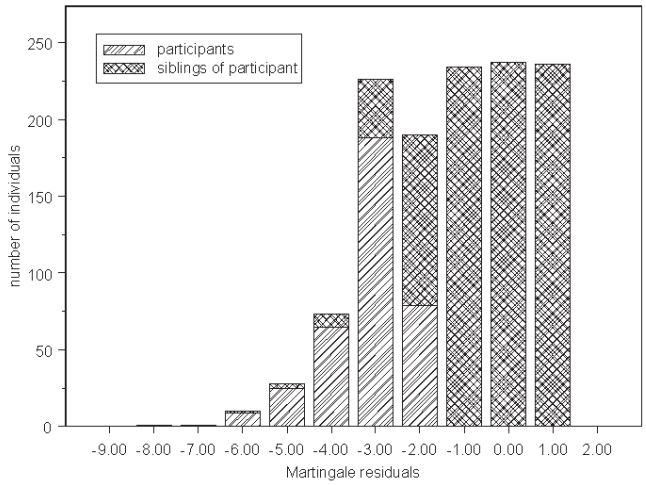

FIGURE 5.2: *Histogram of martingale residuals within in the sibling generation*

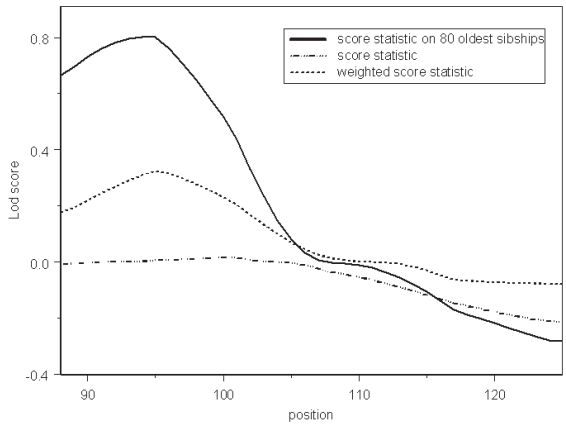

FIGURE 5.3: *Lod scores for various statistics*

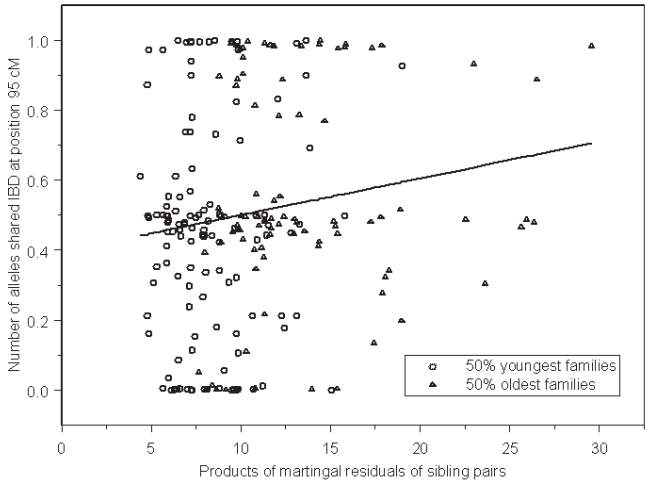

FIGURE 5.4: *For the sibling pairs, the relationship between the number of alleles shared IBD and the product of their martingale residuals.*

Linkage analysis

In figure 3, the lod-score curves are given for the unweighted statistic (5.6) applied to the whole data-set and applied to a subset of 50% oldest families (91 pairs). To obtain the oldest sibling pairs, we computed the minimum value of the cumulative hazard per family and took the 50% families with the highest minimum value. By doing so both siblings will be relatively old. The unweighted statistic applied to the whole data set confirmed the analysis of (Beekman et al., 2006) that there is no evidence for linkage in this set of families. Analysis of the oldest sibships gave a highest lod score of 0.8 at 95 cM.

The product of the cumulative hazards of the sibling of each pair ranged from 4.4 to 29.6, with a mean of 10.4. The highest product corresponds to a sibling pair of 100 and 99 years of age. In figure 3, the lod score corresponding to the weighted score statistic (5.8) is given. A maximum lod score of 0.3 at position 95 cM was obtained. In figure 4, the relationship between the IBD sharing at position 95 cM and the product of the cumulative hazards is depicted. It appeared that a high product corresponds to a high proportion of alleles shared identical by descent.

5.4 Discussion

In this paper weighted statistics to test for aggregation and linkage of human longevity were derived and applied to the families of the Leiden Longevity

study. When selection criteria of older than 90 years are used, it makes sense to assign more weight to extremly long lived siblings as these statistics do. These statistics have a straightforward formula and are robust against model deviations. When the weights do not reflect the true underlying survival model, the test statistic is still valid.

Analysis of data from the Leiden Longevity Study showed excess survival of the parents and offspring of the long-lived siblings. Evidence for an association between survival of the sibling generation and the survival of the offspring generation exists (cor=0.25), yielding an increased significance of the weighted statistic compared to the unweighted statistic. No association was present between survival of the sibling generation and survival of the parents (cor=0.02) and weighting did not improve significance of the test statistic for aggregation. For linkage analysis, the Leiden Longevity Study did not show evidence for a trait locus in the region which was identified by Puca et al. (2001). The maximum lod score corresponding to the weighted statistic was higher than the lod score of the unweighted statistic, but still far from significant.

Since the families are selected on excessive survival, the computed correlations between survival measures of the various generations are meant for illustration and cannot be generalized to the general population. Further early mortality is under represented in the parental generation, because parents with large offspring sizes are more likely to be in the sample. In contrast early mortality is present in the siblings and the offsprings of the nonagenarians. The generation of the offspring cannot express the longevity trait yet and hence any correlation with the sibling generation must be based on mean survival at middle age which may rather reflect the absence of mortality related variants (late-acting deleterious alleles) than the presence of exceptional longevity promoting variants. The excess survival in the parental generation is probably due to longevity promoting variants. The lack of correlation between parental and sibling generation may be explained by the fact that in the parental generation only longevity promoting variants are present while in the sibling generation both variants are present.

The analysis of the subset of 80 oldest sibship was significant (lod score of 0.8), but the data-set is too small to prove linkage. Although analyzing various subsets is appealing, it has also some drawbacks, namely the sample size of the subset is reduced and for continues outcomes the choice of the cut-off is arbitrary. When multiple cut-off values are used adjustments for multiple testing have to be applied. As alternative for using subsets based on the current ages of the siblings, we derived a weighted statistic from a frailty model for the current ages of the siblings. Correlation in survival between the offspring and sibling generation is present in these data, therefore including also the information on

survival of siblings and offspring of the participants, may further improve the efficiency of linkage analysis to identify deleterious variants while efficiency for identification of longevity variants is unlikely to be improved. Here more research is needed. For European countries the weighted statistic provides a tool to deal with the heterogeneity in life expectancies between various countries.

In this paper we derived models for longevity and therefore assumed no events. The method can easily be extended to the situation that some of the subjects experienced the event of interest. For δ equal to zero, the general score function becomes the product of the martingale residuals $(d_1 - H_1)(d_2 - H_2)$. Note that for $\delta = 0$, the derived weighted statistic corresponds to the Haseman-Elston method (Haseman and Elston, 1972; Sham and Purcell, 2001) by considering the martingale residuals as quantitative outcomes. This approach was proposed by Yoo et al. (2001). The numerator of our weighted statistic is also equivalent to the numerator of the statistic of Commenges (1994), but the variances are different. Commenges (1994) used the conditional likelihood of the outcomes given the IBD probabilities and did not use regression of the IBD probabilities on $U(\delta)$. Therefore this statistic may not be appropriate for selected data while our statistic is valid for selected data.

Another weighted method that has been proposed is the score statistic of Hsu et al. (2002) for age at onset data. Hsu et al. (2002) propose the following weight $w_{jk} = (X_j - X_0)(X_k - X_0)$ where X is the age at onset and X_0 is the age where the mean IBD between ASP is 0.5. For binary outcomes with additional endophenotypes or covariates, Whittemore and Halpern (2006) proposed a (Kong and Cox (1997)) model with person specific weights. For our data the current age can be considered as additional information yielding a product of the cumulative hazards as weight. The main limitation of the latter two methods is that the weighting functions were not based on a bivariate survival model.

Software to apply the new weighted score statistic for linkage will soon be available. Alternatively to a score statistic, one may want to apply Kong and Cox likelihood ratio method weighted the product of the cumulative hazards (see also Whittemore and Halpern (2006)). To apply this method, any software which allows for weights can be used.

To asses aggregation of life span within the Leiden Longevity Study, we tested for a relationship between the sum of martingale residuals of the sibship generation and survival of the parental and offspring generations. Alternative measures for family history may be considered. Murad et al. (2007) compared various family history measures and concluded that most quantitative scores perform well and better than dichotomous scores.

To conclude we identified aggregation of life span within the families of the

Leiden Longevity study. Therefore the efficiency of future linkage and association studies in the Leiden Longevity Study will be improved if the information available on the age distributions within the families is used. For linkage analysis of human life span in long-lived siblings pairs of age 90 years and above, we recommend to apply our weighted score statistic. The weighted linkage statistic also provides a tool for meta analysis of linkage studies for Longevity of various European countries with different life expectancies.

Table 5.1: *Descriptives of life span in families of Leiden Longevity Study and results of testing for excessive survival.*

type of relative	number	number deceased	age at death*	family specific SMR*	family specific sumMR**	cor#	Unweighted statistic%	Weighted statistics&
siblings	1249	686	76 (0-101)	0.29 (0.00,0.67)	-9.5 (3.9)	1	-	-
parents	330	330	77 (27-104)	0.77 (0.22,5.20)	-0.8 (1.9)	0.02	-5.63	-5.43
mothers	164	164	77 (27-104)	1.04 (0.15,42.8)	-0.2 (1.3)	0.02	-3.34	-3.32
fathers	166	166	78 (34-103)	0.89 (0.16,34.2)	-0.5 (1.4)	0.05	-4.60	-4.31
offspring	1317	138	44 (0-72)	0.00 (0.00,22.5)	-0.4 (1.0)	0.25	-4.70	-5.12

*median (range), see formula (1)
**mean (sd), see formula (4)
correlation between sumMR and $sumMR^s$
%standardized scores, see formula (3)
&standardized weighted scores, see formula (5)