Title

Ascertainment-Adjusted Maximum Likelihood Estimation for the Additive Genetic Gamma Frailty Model

Authors

Sun, Wanlong
Li, Hongzhe

Publication Date

2003-12-01

Peer reviewed

Ascertainment-Adjusted Maximum Likelihood Estimation for the Additive Genetic Gamma Frailty Model

WANLONG SUN
Rowe Program in Human Genetics, University of California Davis School of Medicine, Davis, CA
HONGZHE LI
Rowe Program in Human Genetics, University of California Davis School of Medicine, Davis, CA

Running title: Ascertainment correction for frailty model.

Address for correspondence:

Hongzhe Li, Ph.D.
Rowe Program in Human Genetics
School of Medicine
University of California
Davis, CA 95616, USA

Tel: (530) 754-9234; Fax: (530) 754-6015
E-mail:hli@ucdavis.edu

Abstract

The additive genetic gamma frailty model has been proposed for genetic linkage analysis for complex diseases to account for variable age of onset and possible covariates effects. To avoid ascertainment biases in parameter estimates, retrospective likelihood ratio tests are often used, which may result in loss of efficiency due to conditioning. This paper considers when the sibships are ascertained by having at least two affected sibs with the disease before a given age and provides two approaches for estimating the parameters in the additive gamma frailty model. One approach is based on the likelihood function conditioning on the ascertainment event, the other is based on maximizing a full ascertainment-adjusted likelihood. Explicit forms for these likelihood functions are derived. Simulation studies indicate that when the baseline hazard function can be correctly pre-specified, both approaches give accurate estimates of the model parameters. However, when the baseline hazard function has to be estimated simultaneously, only the ascertainment-adjusted likelihood method gives an unbiased estimate of the parameters. These results imply that the ascertainment-adjusted likelihood ratio test in the context of the additive genetic gamma frailty may be used for genetic linkage analysis.

1 Introduction

Correction for ascertainment has been a long-standing problem in statistical genetics (Cannings and Thompson, 1977). For genetic linkage analysis, families are often ascertained by collecting only families with multiple affected individuals, leading to non-random sampling of families from the study population. There has been new interest in studying this problem in the context of including the latent trait heterogeneity. For example, Burton *et al.* (2000) noted that classical approaches to ascertainment fail to acknowledge the unseen population structure and are inconsistent. Epstein *et al.* (2002) and Glidden and Liang (2002) further examined the scenario considered in Burton *et al.* (2000) by considering different formulations of the likelihood function, with both concluding that proper construction of the ascertainment-adjusted likelihood can yield population-based parameter estimates. However, both papers considered the ascertainment problem for binary affected/unaffected traits within the framework of the logistic variance-components models. Pfeiffer *et al.* (2001) and Neuhaus *et al.* (2002) studied inference for covariates that accounts for ascertainment and random genetic effects in family studies for binary traits.

In order to account for variable age of onset and potential environmental risk factors in genetic linkage analysis, Li (1999), Li and Zhong (2002) and Li (2002) developed the additive genetic gamma frailty model, which utilizes inheritance vectors (Lander and Green, 1987). Within this modelling framework, Li and Zhong (2002) (abbreviated as LZ in the rest of the paper) used a retrospective likelihood formulation for estimating the parameters and for performing linkage analysis. While this approach is free of potential ascertainment biases, it may result in loss of efficiency due to conditioning. If the ascertainment scheme is known and the ascertainment probability is taken into account, one would expect an increase in efficiency in parameter estimates. Li (2002) proposed a semiparametric prospective likelihood approach for genetic linkage analysis using the EM algorithm, but did not consider the issue of ascertainment correction.

In order to rigorously account for ascertainment in model estimation, one must clearly define the sampling scheme of the families (Thompson, 1993; Thompson and Cannings, 1979). In this paper, we consider the ascertainment scheme of collecting only sibships with at least two affected sibs with age of disease onset before a fixed age t_0. This sampling scheme is commonly used in genetic linkage studies of diseases with variable age of onset. Under this ascertainment scheme, we use the additive genetic gamma frailty model of LZ to consider the ascertainment-adjusted (AA) maximum likelihood estimation for censored age of onset traits. At present, this is the first attempt to study the ascertainment issue for

linkage analysis relative to survival data and random effects models. We demonstrate by simulation studies that the proposed methods can be applied to obtain unbiased estimates of both the population baseline hazard function and the parameters involved in the random frailty effects.

The rest of the paper is organized as follows. We first briefly review the additive genetic gamma frailty model defined in LZ. We then present two different likelihood formulations for including ascertainment when estimating the parameters in the model. One formulation is based on a conditional likelihood, the other based on an ascertainment-adjusted likelihood to explicitly account for ascertainment probabilities. We then present simulation results on evaluating these methods. Finally, we give a brief discussion on the implications for genetic linkage analysis.

2 The Additive Genetic Gamma Frailty Model and the Joint Survival and Density Functions

2.1 *The additive genetic Gamma frailty model*

Consider a sibship with K sibs. Let T_j be the random variable of age at disease onset for the jth sib. Let (t_j, δ_j) be the observed data, where t_j is the observed age at onset if $\delta_j = 1$, and age at censoring if $\delta_j = 0$. We assume that the hazard function of developing disease for the jth individual at age t_j is modelled by the proportional hazards model with random effect Z_j,

$$\lambda_j(t_j|Z_j) = \lambda_0(t)\exp(X_j^{'}\beta)Z_j, \text{ for } j = 1, 2, \cdots, K, \tag{1}$$

where $\lambda_0(t)$ is the unspecified baseline hazard function, X_j is a vector of observed covariates for the jth sib, and β is a vector of regression parameters associated with the covariates. Z_j is the unobservable random genetic frailty. LZ constructed the genetic frailties based on the inheritance vector at the putative disease locus d, denoted by V_d. In the following discussion, the paternal chromosomes containing the locus of interest are labelled by (1,2), the maternal chromosomes by (3,4). The inheritance vector (Kruglyak *et al.*, 1996; Lander and Green, 1987) of a sibship at the d locus is defined as

$$V_d = (v_1, v_2, \cdots, v_{2j-1}, v_{2j}, \cdots, v_{2K-1}, v_{2K}),$$

where $v_{2j-1} = 1$ or 2, $v_{2j} = 3$ or 4 for $j = 1, 2, \cdots, K$. The inheritance vector indicates which parts of the genome at locus d are transmitted to the K children from the father

and the mother. LZ defined the frailties for a sibship of size K as

$$Z = HU, \tag{2}$$

where

$$Z = \{Z_1, Z_2, \cdots, Z_K\}',$$

$$H = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & 1 \\ & & \vdots & & \\ a_{K1} & a_{K2} & a_{K3} & a_{K4} & 1 \end{pmatrix},$$

$$U = \{U_{d1}, U_{d2}, U_{d3}, U_{d4}, U_p\}'.$$

Here H is the inheritance matrix with elements

$$\begin{cases} a_{j1} = 1, a_{j2} = 0 & \text{if } v_{2j-1} = 1 \\ a_{j1} = 0, a_{j2} = 1 & \text{if } v_{2j-1} = 2 \\ a_{j3} = 1, a_{j4} = 0 & \text{if } v_{2j} = 3 \\ a_{j3} = 0, a_{j4} = 1 & \text{if } v_{2j} = 4 \end{cases}$$

for $j = 1, \cdots, K$. The random effects U_{d1} and U_{d2} are used to represent genetic frailties from the father's two chromosomes, and similarly, U_{d3} and U_{d4} are used to represent the genetic frailties inherited from the mother. Here U_p is employed to account for possible genetic contributions to the disease due to loci unlinked to locus d. We further assume that the U_{d1}, U_{d2}, U_{d3} and U_{d4} are independently and identically distributed and follow $\Gamma(\nu_d/2, \eta)$, and U_p follows $\Gamma(\nu_p, \eta)$ across different sibships, where parameter η is the inverse scale parameter and ν_d and ν_p are the shape parameters. Then Z_j is distributed as $\Gamma(\nu_d + \nu_p, \eta)$, for $j = 1, 2, \cdots, K$. To make the baseline hazard function $\lambda_0(t)$ identifiable, let $\nu_d + \nu_p = \eta$, which sets $E(Z_j) = 1, j = 1, 2, \cdots, K$, and prevents arbitrary scaling in model (1). Under this restriction, there are two free parameters, ν_d and ν_p. We may also consider reparameterization in terms of the two frailty means, $\mu_d = 2E(U_{dj}) = \nu_d/\eta$, and $\mu_p = E(U_p) = \nu_p/\eta$ and the variance $\sigma^2 = Var(Z_j) = 1/\eta$. Note that $\mu_d + \mu_p = 1$, so there are only two free parameters, μ_d and σ^2. We use this parameterization in our simulation studies. We denote $\Theta = (\lambda_0(t), \beta, \mu_d, \sigma^2)$ as the parameters in model (1), where μ_d and σ^2 are the parameters related to the genetic frailties.

2.2 *The joint survival and density functions*

Assuming conditional independence and based on model (1), we observe that conditioning on the frailty vector Z and the inheritance vector, the joint survival function for a sibship can be written as

$$S(t_1, t_2, \cdots, t_K | Z_1, Z_2, \cdots, Z_K, V_d) \;=\; \exp[-\Lambda_1(t_1)Z_1 - \Lambda_2(t_2)Z_2 - \cdots - \Lambda_n(t_K)Z_K],$$

where

$$\begin{aligned} \Lambda_j(t_j) = &\quad \Lambda_0(t_j)\exp(X_j'\beta), \qquad j = 1, 2, \cdots, K, \\ \Lambda_0(t_j) = &\quad \int_0^{t_j} \lambda_0(u)du, \qquad\quad j = 1, 2, \cdots, K. \end{aligned}$$

It is simple to verify that the marginal joint survival function by integrating out $Z_1, Z_2, \cdots, Z_K$ is given by

$$\begin{aligned} & S(t_1, \cdots, t_K | V_d) = Pr(t_1, \delta_1 = 0, \cdots, t_K, \delta_K = 0 | V_d) \\ = & \left\{ \prod_{i=1}^{4} \frac{\eta^{\nu_d/2}}{[\sum_{j=1}^{K} \Lambda_j(t_j)a_{ji} + \eta]^{\nu_d/2}} \right\} \times \left\{ \frac{\eta^{\nu_p}}{[\sum_{j=1}^{K} \Lambda_j(t_j) + \eta]^{\nu_p}} \right\}, \end{aligned}$$

(see LZ for derivation). In practice, observations are often censored and therefore we need not only the joint survival function but also combined densities and survivor functions. For a sibship with a affected sibs (indexed by $j = 1, \cdots, a$) and $K - a$ unaffecteds, the joint survival and density function is

$$Pr(t_1, \delta_1 = 1, \cdots, t_a, \delta_a = 1, t_{a+1}, \delta_{a+1} = 0, \cdots, t_K, \delta_K = 0 | V_d) = (-1)^a \frac{\partial^a S(t_1, \cdots, t_K | V_d)}{\partial t_1, \cdots, \partial t_a}.$$

For sibship with all sibs affected, the joint density function is

$$Pr(t_1, \delta_1 = 1, \cdots, t_K, \delta_K = 1 | V_d) = (-1)^K \frac{\partial^K S(t_1, \cdots, t_K | V_d)}{\partial t_1, \cdots, \partial t_K}.$$

The closed forms of these expressions and the detailed derivations can be referenced in LZ.

2.3 *Retrospective likelihood, prospective likelihood and sampling scheme*

Based on the joint density and survival functions presented in the previous section, LZ showed that testing whether the putative disease locus d affects the disease risk can be formulated as testing $\nu_d = 0$. Since families collected for genetic linkage analysis are not typically a random sample from the study population, estimation of the parameters in model (1) based on the usual prospective likelihood function can result in potentially large

biases and lead to incorrect conclusions of linkage. LZ proposed using a retrospective likelihood ratio test (Whittemore, 1996) for testing this null hypothesis. This likelihood function is defined as the probability of the observed marker data conditioning on the sibship age of onset/age at censoring data. Although this retrospective likelihood ratio test is valid for any ascertainment scheme, it requires specification of the baseline hazard function and may suffer loss of efficiency due to conditioning on the phenotypes of the whole sibships.

When the sampling scheme for families is clearly defined and followed, it becomes possible to make statistical inference based on a prospective likelihood formulation. We consider a particular population-based sampling scheme for collecting data for genetic linkage analysis. Under this sampling scheme, we sample sibships randomly from the population, until we obtain n sibships with at least two sibs who were affected with the disease before age t_0. These n sibships are then genotyped and used for genetic linkage analysis. Parents of these sibs can also be genotyped if their DNAs are available. However, phenotype information is not necessary for the parents, as it is not used in the proposed linkage analysis. Suppose that in order to obtain n such sibships, n_0 families with at least two sibs but with no sib being affected before t_0 and n_1 families with at least two sibs but with only one sib being affected before t_0 were contacted. This is a special case of the epidemiologic survey sampling of families. It is important to note that in this sampling scheme, n_0 and n_1 are random variables rather than parameters. In addition, by strictly following the sampling scheme, one can also obtain the empirical estimates of the distribution of families with different numbers of children in the study population. This distribution is used in the AA likelihood formulation presented in a later section. We propose in the subsequent two sections, two alternative approaches for estimating the parameters in model (1) based on the sibship data collected by this sampling scheme.

3 A Conditional Likelihood Approach for Ascertainment Correction

Under the ascertainment scheme where only the sibships with at least two sibs being affected before age t_0 are collected, we can assume the n sibships collected are a random sample from such sibships. Further, let K_i be the number of the sibs in the ith sibship, for $i = 1, 2, \cdots, n$. Define $AC = \{$At least two members of the family are affected before age $t_0\}$, then for a given sibship with known inheritance vector V_d, the conditional (on the random effects)

probability of the ascertainment event is

$$Pr(AC|Z,V_d) = 1 - \exp\left[-\sum_{j=1}^{K}\Lambda_0(t_0)\exp(\beta' X_j)Z_j\right]$$
$$- \sum_{i=1}^{K}\left\{\exp\left[-\sum_{j\neq i}\Lambda_0(t_0)\exp(\beta' X_j)Z_j\right] - \exp\left[-\sum_{j=1}^{K}\Lambda_0(t_0)\exp(\beta' X_j)Z_j\right]\right\}.$$

For simplicity, this paper only considers the case when no covariates are included in the model. We can then integrate out the vector of random effects Z and obtain

$$\begin{aligned} Pr(AC|V_d) &= 1 + (K-1)\prod_{l=1}^{4}\left[\frac{\eta}{\sum_{j=1}^{K}\Lambda_0(t_0)a_{jl}+\eta}\right]^{\frac{\nu_d}{2}} \times \left[\frac{\eta}{K\Lambda_0(t_0)+\eta}\right]^{\nu_p} \\ &\quad -\sum_{i=1}^{K}\prod_{l=1}^{4}\left[\frac{\eta}{\sum_{j\neq i}\Lambda_0(t_0)a_{jl}+\eta}\right]^{\frac{\nu_d}{2}} \times \left[\frac{\eta}{(K-1)\Lambda_0(t_0)+\eta}\right]^{\nu_p}. \end{aligned}$$

Then for the ith sibship, for a given inheritance vector V_{di}, the likelihood function, conditioning on the ascertainment event AC, is

$$L_i(\Theta|V_{di}) = I(AC) \times \frac{Pr(t_1,\delta_1=1,\cdots,t_{a_i},\delta_{a_i}=1,t_{a_i+1},\delta_{a_i+1}=0,\cdots,t_{K_i},\delta_{K_i}=0|V_{di})}{Pr(AC|V_{di})},$$

where $I(AC)$ is 1 if the values of $(t_1,\delta_1,\cdots,t_{a_i},\delta_{a_i},t_{a_i+1},\delta_{a_i+1},\cdots,t_{K_i},\delta_{K_i})$ meet the ascertainment criterion AC and 0 otherwise. Taking into account the uncertainty of the inheritance vector, the conditional likelihood for the ith sibship can be written as

$$L_i(\Theta) = \sum_{v_{di}} L_i(\Theta|V_{di}=v_{di})Pr(v_{di}|M_i),$$

where $Pr(V_{di}=v_{di}|M_i)$ is the probability distribution of the inheritance vector at the putative disease locus d given the marker data for the ith sibship. Such distribution can be calculated using multipoint methods (e.g., those of Lander and Green, 1987; Kruglyak *et al.*, 1996) using all the marker information for the ith sibship, M_i., including the parental genotypes if they are available.

If we assume that the baseline hazard function belongs to a parametric family, we can obtain an estimate of the parameters in the baseline hazard and the parameter associated with the frailties by maximizing the following conditional likelihood function:

$$L^{cond}(\Theta) = \prod_{i=1}^{n} L_i(\Theta). \tag{3}$$

4 An Ascertainment-Adjusted Maximum Likelihood Approach

4.1 *Ascertainment-adjusted likelihood functions*

As defined in the previous section, conditional likelihood function (3) relies on families meeting the ascertainment criteria. Such conditioning may result in loss of efficiency, especially when many of the sibships collected have only two affected sibs before age t_0. Considering the sampling scheme outlined in the previous section, suppose n families with at least two sibs who are affected with the disease before age t_0 are collected for genetic linkage analysis. Also presume that in order to obtain n such families, we contacted n_0 families with at least two sibs but with no sib being affected before t_0, and n_1 families with at least two sibs but with only one sib being affected before t_0. The corresponding AA likelihood function can be written as

$$L^{asc}(\Theta) = P_0(\Theta)^{n_0} P_1(\Theta)^{n_1} \prod_{i=1}^{n} P(i)(\Theta), \tag{4}$$

where $P_i(\Theta)$ is the probability that a randomly selected sibship of size greater or equal to 2 has i affected member before age t_0, for $i = 0, 1$, and

$$P(i)(\Theta) = \sum_{V_{di}} Pr(t_1, \delta_1 = 1, \cdots, t_{a_i}, \delta_{a_i} = 1, t_{a_i+1}, \delta_{a_i+1} = 0, \cdots, t_{K_i}, \delta_{K_i} = 0 | V_{di}) Pr(V_{di} | M_i), \tag{5}$$

which is the joint probability of the observed data for the ith sibship.

If n_0 or n_1 are unknown, but $n_0 + n_1$ is known, we can also define the ascertainment-adjusted likelihood function as

$$L^{asc1}(\Theta) = (P_0(\Theta) + P_1(\Theta))^{n_0+n_1} \prod_{i=1}^{n} P(i)(\Theta), \tag{6}$$

which requires only the value of the random variable $n_0 + n_1$, that is, the total number of families with zero or one affected sib that were contacted during the sampling process of getting n sibships with at least two affected sibs with age of onset before t_0.

4.2 *Estimation of the probabilities*

We now give some details on how to calculate the probability P_0, and P_1 in the likelihood functions (4) and (6). For a given inheritance vector V_d, we introduce the following notation,

$$\begin{aligned} Q(K, j | V_d) &= Pr(T_1 \le t_0, \cdots, T_j \le t_0, T_{j+1} > t_0, \cdots, T_K > t_0 | V_d), \\ Q(K, j) &= Pr(T_1 \le t_0, \cdots, T_j \le t_0, T_{j+1} > t_0, \cdots, T_K > t_0), \end{aligned}$$

for $j = 0, 1, \cdots, K, K = 1, 2, \cdots$. It is clear that

$$Q(K, j+1|V_d) \;\; = \;\; Q(K-1, j|V_d) - Q(K, j|V_d),$$

and similarly,

$$Q(K, j+1) = Q(K-1, j) - Q(K, j), \tag{7}$$

for $j = 0, 1, \cdots, K-1$. Note that for a given inheritance vector V_d,

$$\begin{aligned} & Q(K,0|V_d) = Pr(T_1 > t_0, T_2 > T_0, \cdots, T_K > t_0|V_d) = S_K(t_0, t_0, \cdots, t_0|V_d) \\ = \;\; & \left\{ \prod_{i=1}^{4} \frac{\eta^{\nu_d/2}}{[\sum_{j=1}^{K} \Lambda_0(t_0) a_{ji} + \eta]^{\nu_d/2}} \right\} \times \left\{ \frac{\eta^{\nu_p}}{[\sum_{j=1}^{K} \Lambda_0(t_0) + \eta]^{\nu_p}} \right\}. \end{aligned}$$

In practice, the inheritance information for those families who were contacted during the sampling process but did not meet the sampling criteria is not available. Considering all possible inheritance patterns with uniform probabilities, we can obtain

$$\begin{aligned} & Q(K,0) = E[Q(K,0|V_d)] = E\left\{ \prod_{i=1}^{2} \frac{\eta^{\nu_d/2}}{[\sum_{j=1}^{K} \Lambda_0(t_0) a_{ji} + \eta]^{\nu_d/2}} \right\} \\ \times \;\; & E\left\{ \prod_{i=3}^{4} \frac{\eta^{\nu_d/2}}{[\sum_{j=1}^{K} \Lambda_0(t_0) a_{ji} + \eta]^{\nu_d/2}} \right\} \times \left\{ \frac{\eta^{\nu_p}}{[\sum_{j=1}^{K} \Lambda_0(t_0) + \eta]^{\nu_p}} \right\} \\ = \;\; & \left\{ \sum_{l=0}^{K} C_K^l \frac{1}{2^K} \frac{\eta^{\nu_d}}{[(\Lambda_0(t_0) l + \eta) \cdot (\Lambda_0(t_0)(K-l) + \eta)]^{\nu_d/2}} \right\}^2 \times \left\{ \frac{\eta^{\nu_p}}{[\sum_{j=1}^{K} \Lambda_0(t_0) + \eta]^{\nu_p}} \right\}. \end{aligned} \tag{8}$$

Using equations (7) and (8), we can calculate $Q(K,0)$, $Q(K,1)$, $\cdots$, $Q(K,K)$. For a given sibship of size K, let $P(j|K)$ be the probability that there are j members being affected before age t_0, then

$$P(j|K) = \left(\begin{array}{c} K \\ j \end{array} \right) Q(K,j), \quad j = 0, 1, \cdots, K.$$

Let q_K be the proportion of the families with K children among the families with at least two children, for $K \geq 2$. Averaging over K, we have

$$P_j(\Theta) = \sum_{K \geq 2} q_K P(j|K). \tag{9}$$

Here q_K can be estimated empirically by surveying the study population during the sampling process.

Finally, if we assume a particular parametric form for the baseline hazard function, we can obtain the parameter estimates by maximizing the likelihood function L^{asc} or L^{asc1} as defined in equation (4)or (6) over the model parameter Θ. We call the resulted parameter estimates the ascertainment-adjusted maximum likelihood estimates.

5 Simulation Studies and Results

In order to investigate the proposed conditional and AA likelihood approaches for estimating the parameters in the additive gamma frailty model, we performed extensive simulation studies for family samples ascertained by having at least two affected sibs before age t_0. We assumed a Weibull baseline hazard function,

$$\lambda_0(t) = \tau \frac{t^{\tau-1}}{b^\tau} \tag{10}$$

in model (1), with parameters $b = 80$ and $\tau = 5$. Age of onset data was simulated from the additive genetic gamma frailty model (1), while the current age was simulated as age at censoring from a uniform distribution $U(60, 80)$. We considered four different models by specifying four different combinations of model parameters, where the genetic variance is $\sigma^2 = 3.33$ for the first two models with $\mu_d = 1$ and μ_d=0.67 respectively, and $\sigma^2 = 1.25$ for model 3 and model 4 with $\mu_d = 1$ and μ_d=0.75 respectively. The corresponding (ν_d, η) is (0.3,0.3), (0.2,0.3), (0.8,0.8) and (0.6,0.8) respectively for the four models. Therefore, the degree of dependence is stronger for models 1 and 2 than for models 3 and 4. Figure 1 (a) shows the baseline survival curve and the corresponding population disease-free survival curves for parameters $\eta = 0.3$ and $\eta = 0.8$. Note that the model with $\eta = 0.8$ results in higher population disease risk than the model with $\eta = 0.3$. We considered three different sample sizes, $n = 100, 200$ and 500. For each model, we simulated families until the required sample size of families (n) with at least two affected sibs before age 40 is obtained. We then recorded n_0 and n_1, which are used in the estimation procedure. All simulations were based on 50 replications. For all the simulations, fully informative markers were simulated so that the inheritance vectors are known for all the sibships and sibships of size 4 were sampled.

5.1 *Evaluation of the conditional likelihood based parameter estimates*

First, consider the case when the baseline hazard function is known. Table 1 shows the mean of the parameter estimates and their empirical standard errors. The results show the excellent performance of the maximum conditional likelihood estimators based on the ascertained samples, indicating that when the baseline hazard function is known, the maximum conditional likelihood gives unbiased estimates of the parameters related to the random genetic effects. We also observed that the empirical standard error for the variance parameter is larger than that for the mean parameter of all the models considered.

We then examined how the conditional maximum likelihood estimation performs when the baseline hazard function is incorrectly specified. We performed simulation studies by assuming that model 1 or 2 are true models and a sample size of 500. We estimated the parameters by using four different slightly misspecified baseline hazard functions (see Figure 1 (b) for the disease-free survival curves corresponding to the true and the misspecified baseline hazard functions) in the conditional likelihood function L^{cond}. Table 2 presents the mean and the standard error of the estimates of the frailty parameters, indicating biased estimates even when the baseline is slightly misspecified.

When the baseline hazard function is unknown, we also considered the method of maximizing the conditional likelihood over both the parameters related to the baseline hazard function; by assuming a Weibull baseline function (10) and the parameters related to the random effects. We observed very large biases for the estimates of all the parameters (details not shown). In some cases, maximum values cannot be found. These results indicate that when the baseline hazard function has to be estimated simultaneously with other parameters in the model, the maximum conditional likelihood estimation cannot be used to obtain satisfactory estimates of the parameters.

5.2 *Evaluation of the ascertainment-adjusted likelihood based parameter estimates*

To evaluate the proposed AA likelihood estimation, we first considered the scenario when the baseline hazard function is known and not part of the estimated parameters. Table 3 shows the means of the parameter estimates and the empirical standard errors of the estimates. The results demonstrate the excellent performance of the maximum likelihood estimators from the ascertained samples. Compared to the results from the maximum conditional likelihood estimation, we note that the empirical standard errors are smaller based on the full likelihood, which is expected since the ascertainment-adjusted likelihood explicitly models how data are generated. As a comparison, since the conditional likelihood approach does not require the values of the random variables n_0 and n_1, it is practical to implement. However, as indicated in previous section, slight misspecification of the baseline hazard function can result in very biased estimates of the parameters.

We next considered the scenario when the baseline hazard function is unknown but can be specified as a Weibull baseline hazard function as defined in (10) with parameters b and τ. We maximize the likelihood function L^{asc} as in equation (4) over parameters in

the baseline hazard function and the parameters involved in the genetic frailties. Table 3 gives the means of the parameter estimates and the estimated standard errors for all the parameters in the model. This table clearly shows that when the values n_0 and n_1 are known, both parameters in the baseline hazard function and the parameters associated with the frailties can be estimated successfully by maximizing the ascertainment-adjusted full likelihood function L^{asc}. When the parameters in the baseline hazard function are estimated simultaneously with the frailty parameters, the estimated standard errors for the frailty parameters are higher than the estimates when the baseline hazard function is correctly specified.

We also investigated the alternative AA likelihood formulation as defined in equation (6). We noted that when $n_0 + n_1$ is observed, the estimates of the model parameters are unbiased, but are slightly less efficient than those obtained by maximizing the likelihood of equation (4) (details not shown). This is the expected outcome, as likelihood (4) utilizes more population information than likelihood (6). When $n_0 + n_1$ is unobserved and is incorrectly specified, estimate of the frailty variance could be biased, but estimates of the other parameters seem to be relatively robust. These simulation results indicate that the estimation procedure based on the likelihood formulation L^{asc} performs better than that based on the likelihood formulation L^{asc1}.

5.3 *Effects of misspecification of the population parameters on the ascertainment-adjusted maximum likelihood estimators*

The AA likelihood estimation procedure requires prediction of both n_0 and n_1 if they are not observed, as well as the distribution of the families with different numbers of children (the q_k parameters in equation (9)). It is therefore important to study how robust the proposed maximum AA likelihood estimators are to misspecification of n_0 or n_1 and the distribution of families with varying numbers of children.

We first investigated the biases resulting from the misspecification of values of n_0 and n_1. We considered cases when the values of n_0 and n_1 are under- or over- specified by 20%, 50% and 100% of their true observed values. Table 4 shows the parameter estimates by maximizing L^{asc}, but with incorrectly specified n_0 and n_1 for two different models and sample size of 500. Overall, we observed that the baseline parameters b and τ and the mean of the genetic frailties can still be estimated efficiently, however, large biases on the estimates of the frailty variance are observed. When n_0 and n_1 are over-predicted, the

frailty variance tends to be underestimated, and when n_0 and n_1 are under-predicted, the frailty variance tends to be overestimated. This follows with the expected result, as smaller frailty variance corresponds to higher prevalence of disease in the population (see Figure 1 (a)).

We then investigated how misspecification of the distribution of the family sizes affects the results. We considered model 2 and sample sizes of 100, 200 and 500 sibships. We generated the sibship data by assuming the true distribution of the family sizes to be $(q_2, q_3, q_4, q_5) = (0.10, 0.50, 0.35, 0.05)$ where q_j is the proportion of the families in the population with j sibs for $j = 2, 3, 4, 5$. We considered four different misspecifications in the estimation procedure (see Table 5 for the misspecified probabilities). The first two misspecifications overestimate the numbers of families with 4 and 5 children and the next two misspecifications overestimate the numbers of families with 2 and 3 children. Table 5 presents the simulation results. Clearly, all parameter estimates are very robust to misspecification of the distribution of the family sizes. In addition, the means and the estimated standard errors are all very similar across different specifications of the distribution of the family sizes.

6 Discussion

We have proposed and evaluated two different methods for estimating the parameters in the additive genetic gamma frailty model for sibships ascertained by having at least two affected sibs with disease before a fixed age t_0. This ascertainment scheme is often used in genetic linkage analysis for complex diseases with variable age of onset. The simulations have demonstrated that when the population disease rate data are known, the conditional likelihood maximization procedure provides unbiased estimates of the parameters related to genetic frailties. However, misspecification of the baseline hazard function can result in significant biases in the parameter estimates by maximizing the conditional likelihood. When the baseline hazard function is unknown but the distribution of different types of families in the population can be roughly estimated, our simulation results indicate that we can obtain unbiased estimates of the parameters in both the baseline hazard function and in the genetic gamma frailty by maximizing the AA likelihood function.

The additive genetic gamma frailty model was developed for genetic linkage analysis by testing the null hypothesis of ν_d=0. Although not studied in this paper, we would expect that the likelihood ratio test based on the AA likelihood function should provide a valid

test for $\nu_d = 0$. We call the proposed likelihood ratio test the prospective ascertainment-adjusted likelihood ratio (PAA-LR) test. We expect that the PAA-LR test has better power then the retrospective likelihood ratio (R-LR) test proposed in LZ for the particular ascertainment considered in this paper. First, similar to model-based parametric log-odds (LOD) score tests for linkage, the PAA-LR test formulates the likelihood function as the probability of observed phenotypes given the marker data through inheritance vectors. The R-LR test in LZ is similar to the allele-sharing based methods as it examines the identity-by-descent sharing among the sibs conditioning on their phenotypes. It is well-known that the model-based LOD methods often have better power than the allele-sharing based methods when the parametric genetic models are correctly specified. We would therefore expect the PAA-LR test to possess greater power to detect linkage than the R-LR test. However, it is critical to emphasize that the PAA-LR test is substantially different from parametric LOD score methods since the PAA-LR test does not require specifying the penetrance functions or mode of inheritance. Second, the R-LR test in LZ can be applied to sibship data ascertained by any methods since the phenotypes of the whole sibship are conditioned on. However, for sibships which are ascertained in a particular sampling scheme, the likelihood ratio test constructed based on how the data are generated is expected to have better power than the R-LR test. Future research should focus on comparison between the power of these two tests to detect genetic linkage.

Although the R-LR test in LZ can be applied to sibship data collected by any ascertainment scheme, it requires a correct specification of the baseline hazard function or population disease rates, which may not be available. LZ demonstrated the robustness in power and type 1 error rates of the R-LR test to modest misspecification of the baseline hazard function. However, for some diseases, it is very difficult or impossible to come up with an estimate of the baseline hazard function. The baseline hazard function can in principle be estimated by maximizing the retrospective likelihood, however, such an estimate can be very unreliable due to the ascertainment problem. To efficiently incorporate age of onset information into genetic linkage analysis, knowledge of the baseline hazard function is essential (Li and Hsu, 2000). In order to estimate the baseline hazard function, certain sampling procedure must be followed in collecting family data. A random sample of families from the study population would be ideal, but is rarely used for genetic linkage studies. We considered in this paper one particular sampling scheme; sibships with at least two affected sibs before a given age. We demonstrated by simulation studies that under this sampling scheme, the population baseline hazard function does not need to be spec-

ified but can be estimated simultaneously with the frailty parameters, assuming we can obtain an approximate estimate of the distribution of the family sizes. When the sampling scheme is strictly followed, the distribution is readily estimable based upon empirical data. In addition, our simulation studies demonstrated that the parameter estimates based on maximizing the AA likelihood are fairly robust to the misspecification of these population parameters.

There are several issues which deserve further investigations. First, for all our simulations, we assume that the data are generated from the true model. It is important to further determine how misspecification of the distribution of the frailty affects the results. Second, for both conditional and AA likelihood approaches, we assume a parametric Weibull distribution for the baseline hazard function. Future directions will include studies of how our simulation results are affected by misspecification of the baseline hazard function. Also providing future direction in this research area is development of semiparametric methods leaving the baseline hazard function unspecified. Finally, we considered only the cases when covariates are not included in the model. If the risk of developing a certain disease depends on some environmental covariates, we need to have the distribution of the covariates in the population and to integrate out the unobserved covariates for those sibships with zero or one affected sib before age t_0 in order to calculate the probabilities P_0 and P_1 in the ascertainment-adjusted likelihood functions. When practically applying this method, however, it is difficult to know the distribution of the covariates in the study population.

In summary, we conclude that when the models are correctly specified and the ascertainment procedure is followed, the ascertainment-adjusted maximum likelihood method provides unbiased estimates of both the parameters in the baseline hazard function and the frailty parameters. The likelihood ratio test based on such ascertainment-adjusted maximum likelihood can be potentially applied for genetic linkage analysis of complex diseases with variable age of onset.

Acknowledgements

This research was supported in part by NIH grant ES09911 and by a NSF SCREMS grant (0079430). We thank the two referees for many helpful comments and Ms. Yolanda Figueroa for carefully reading the final version of the paper.

References

P.R. Burton, L.J. Palmer,K. Jacobs, K.J. Keen, J.M. Olson, and R.C. Elston, "Ascertainment adjustment: where does it take us?," *American Journal of Human Genetics*, 67, pp1505-1514, 2000.

C. Cannings and E.A. Thompson, "Ascertainment in the sequential sampling of pedigrees," *Clinical Genetics*, 12, pp208-212, 1977.

M.P. Epstein, X. Lin, and M. Boehnke, "Ascertainment-adjusted parameter estimates revisited," *American Journal of Human Genetics*, 70, pp886-895, 2002.

D.V. Glidden, K.Y. Liang, "Ascertainment adjustment in complex diseases," *Genetic Epidemiology*, 23, pp1-9, 2002.

L. Kruglyak, M.J. Daly,M.P. Reeve-Daly, and E.S. Lander, "Parametric and nonparametric linkage analysis: a unified multipoint approach," *American Journal of Human Genetics* 58, pp1347-1363, 1996.

E. Lander and P. Green, "Construction of multilocus genetic maps in humans," *Proceedings of National Academy of Sciences USA* 84, pp2363-2367, 1987.

H. Li, "The additive genetic gamma frailty model for linkage analysis of age-at-onset variation," *Annals of Human Genetics*, 63, pp455-468, 1999.

H. Li, "The additive genetic gamma frailty model for linkage analysis of diseases with variable age of onset using nuclear families," *Lifetime Data Analysis*, 8, pp315-334, 2002.

H. Li and Hsu L., "Effects of Ages at Onset on the Power of the Affected Sib Pair and Transmission/Disequilibrium Tests," *Annals of Human Genetics,* 64, pp239-254, 2000.

H. Li and X. Zhong, "The additive genetic gamma frailty models induced by genetic frailties, with applications to linkage analysis," *Biostatistics*, 3(1), pp57-75, 2002.

J. Neuhaus, A.J. Scott, and C.J. Wild, "The analysis of retrospective family studies," *Biometrika*, 89,pp23-37. 2002.

R.M. Pfeiffer, M.H. Gail,and D. Pee, "Inference for covariates that accounts for ascertainment and random genetic effects in family studies," *Biometrika*, 88, pp933-948, 2001.

E.A. Thompson, "Sampling and Ascertainment in Genetic Epidemiology: A Tutorial Review,". Technical Report No. 243, Department of Statistics, University of Washington, 1993. http://www.stat.washington.edu/thompson/Genepi/Reports.shtml

E.A. Thompson and C. Cannings. "Sampling schemes and ascertainment," *Proceedings of the Workshop on the Genetics of Common Diseases: Applications to Predictive Factors in Coronary Disease.* Eds. C. Sing and M.H.Skolnick. Alan.R.Liss, Inc., NY. 363-382, 1979.

A.S Whittemore, "Genome scanning for linkage: An overview," *American Journal of Human Genetics*, 59, pp704-716, 1996.

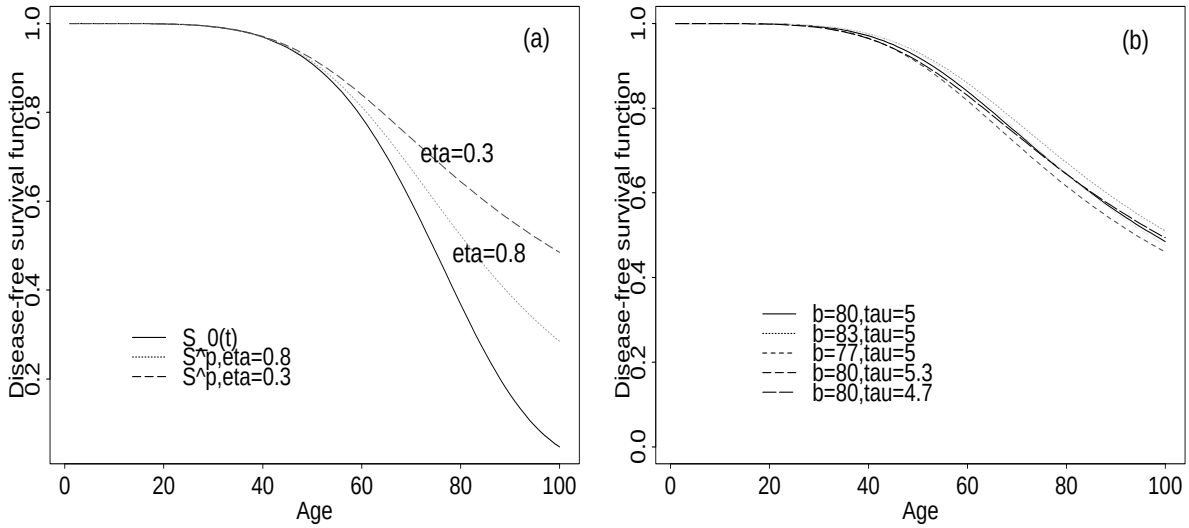

Figure 1: Survival functions used to generate data for simulations. (a) disease-free survival curves for the baseline and the population for $\eta = 0.3$ and 0.8 (σ^2=3.33 and 1.25); (b) population disease-free survival curves for different parameters in the Weibull baseline model for $\eta = 0.3$ ($\sigma^2 = 3.33$).

Table 1: Means and estimated standard errors (ESE) for parameter estimates based on maximizing the conditional likelihood L^{cond} over the frailty parameters for four different models and three different sample sizes when the baseline hazard function is assumed to be known. For each sample size, the first column is the mean (ESE) of the estimates of μ_d, the second column is the mean (ESE) of the estimates of σ^2.

Model	Sample Size		
Parameters	100	200	500
$\mu_d = 1$,$\sigma^2 = 3.33$	0.97(0.060),3.54(0.92)	0.98(0.034),3.32(0.61)	0.99(0.023),3.32 (0.33)
$\mu_d = 0.67$,$\sigma^2 = 3.33$	0.66(0.12),3.36(0.72)	0.66(0.099),3.35(0.57)	0.67(0.061),3.32(0.36)
$\mu_d = 1$,$\sigma^2 = 1.25$	0.95(0.14),1.29(0.45)	0.96(0.074),1.29(0.30)	0.96(0.077), 1.27(0.14)
$\mu_d = 0.75$,$\sigma^2 = 1.25$	0.79(0.19),1.26(0.36)	0.77(0.15),1.33(0.32)	0.76(0.087),1.32(0.21)

Table 2: Mean and estimated standard error (number in parenthesis) of the parameter estimates based on maximizing the conditional likelihood L^{cond} when the baseline hazard function is misspecified for two different models and sample size of 500. The true values for the Weibull baseline hazard function are $b = 80$ and $\tau = 5$.

Model	b or τ misspecified			
Parameters	$b = 83, \tau = 5$	$b = 77, \tau = 5$	$b = 80, \tau = 5.3$	$b = 80, \tau = 4.7$
			Model 1	
$\mu_d = 1$	0.96(0.052)	1.00(0.0073)	0.97(0.041)	1.00(0.011)
$\sigma^2 = 3.33$	3.12(0.28)	4.08(0.78)	3.02(0.56)	4.23(0.92)
			Model 2	
$\mu_d = 0.67$	0.58(0.14)	0.74(0.089)	0.61(0.11)	0.73(0.080)
$\sigma^2 = 3.33$	3.10(0.54)	4.03(0.74)	2.99(0.73)	4.15(0.86)

Table 3: Means and estimated standard errors (numbers in parenthesis) for parameter estimates based on maximizing the ascertainment-adjusted likelihood L^{asc} over the baseline hazard function and the frailty parameters for four different models and three different sample sizes. For each given sample size, the first column assumes that the baseline hazard and n_0 and n_1 are observed, the second column assumes that n_0 and n_1 are observed but the baseline hazard function is estimated from the data.

Model	Sample Size					
	100		200		500	
			Model 1			
$b = 80$	-	80.59(3.42)	-	80.23(2.65)	-	80.25(1.35)
$\tau = 5$	-	4.97(0.29)	-	4.99(0.23)	-	4.99(0.11)
$\mu_d = 1$	0.97(0.003)	0.97(0.053)	0.98(0.002)	0.98(0.041)	0.99(0.001)	0.98(0.025)
$\sigma^2 = 3.33$	3.33(0.46)	3.32(0.45)	3.23(0.34)	3.22(0.33)	3.33(0.20)	3.23(0.21)
			Model 2			
$b = 80$	-	80.61(2.82)	-	80.53(2.39)	-	80.05(1.47)
$\tau = 5$	-	4.97(0.25)	-	4.97(0.21)	-	4.99(0.13)
$\mu_d = 0.67$	0.66(0.11)	0.66 (0.11)	0.66(0.097)	0.66(0.097)	0.67(0.059)	0.67(0.060)
$\sigma^2 = 3.33$	3.38(0.44)	3.37(0.47)	3.44(0.40)	3.43(0.39)	3.34(0.22)	3.35(0.21)
			Model 3			
$b = 80$	-	80.74(3.25)	-	80.93(2.47)	-	80.35(1.56)
$\tau = 5$	-	4.95(0.27)	-	4.93(0.22)	-	4.97(0.13)
$\mu_d = 1$	0.96(0.12)	0.95(0.12)	0.97(0.068)	0.96(0.077)	0.96(0.064)	0.96(0.066)
$\sigma^2 = 1.25$	1.32(0.29)	1.32(0.29)	1.25(0.17)	1.25(0.18)	1.23(0.12)	1.22(0.12)
			Model 4			
$b = 80$	-	80.69(4.04)	-	80.00(2.18)	-	80.07(1.41)
$\tau = 5$	-	4.96(0.35)	-	5.01(0.19)	-	5.00(0.13)
$\mu_d = 0.75$	0.76(0.18)	0.77(0.18)	0.76(0.14)	0.76(0.15)	0.76(0.0082)	0.76(0.085)
$\sigma^2 = 1.25$	1.32(0.33)	1.33(0.34)	1.33(0.22)	1.33(0.23)	1.30(0.11)	1.30(0.12)

Table 4: Means and estimated standard errors (numbers in parenthesis) for parameter estimates based maximizing the ascertainment-adjusted likelihood L^{asc} over both baseline hazard function and the frailty parameters for two different models and sample sizes of 500 when the prediction of n_0 and n_1 are subject to errors, where the percentage of misspecification is defined as $[n_j(\text{predicted}) - n_j(\text{true observed})]/n_j(\text{true observed})$, for $j = 0, 1$.

	% of misspecification of n_0 and n_1			
Parameters	20 %	-20 %	100 %	-50 %
		Model 1		
$b = 80$	80.11(1.27)	80.18 (1.31)	79.42(1.40)	79.36(1.43)
$\tau = 5$	5.05 (0.12)	4.91(0.14)	5.23(0.26)	4.78(0.25)
$\mu_d = 1$	0.99 (0.014)	0.97(0.039)	1.00(0.00)	0.94(0.070)
$\sigma^2 = 3.33$	2.81(0.54)	3.93(0.61)	1.60(1.73)	5.26(1.94)
		Model 2		
$b = 80$	79.80(1.50)	80.18(1.48)	78.41(2.19)	79.55(1.48)
$\tau = 5$	5.08(0.16)	4.90(0.16)	5.35(0.39)	4.74(0.29)
$\mu_d = 0.67$	0.69(0.061)	0.64 0.066)	0.76(0.11)	0.59(0.10)
$\sigma^2 = 3.33$	2.91(0.45)	3.85(0.56)	1.84(1.50)	4.89(1.58)

Table 5: Means and estimated standard errors (numbers in parenthesis) for parameter estimates based on maximizing the ascertainment-adjusted likelihood L^{asc} over both baseline hazard function and frailty parameters for model 2 and sample sizes of 100, 200 and 500 when the distribution of the family sizes is misspecified. The true distribution of families with 2,3,4 and 5 children is $(q_2, q_3, q_4, q_5) = (0.10, 0.50, 0.35, 0.05)$. The misspecified distributions are: miss-1=(0.05,0.35,0.50,0.10); miss-2=(0.05, 0.45,0.40,0.10); miss-3=(0.13,0.55,0.30.0.02); miss-4=(0.13,0.65,0.20,0.02).

Sample size	misspecified distribution			
Parameters	miss-1	miss-2	miss-3	miss-4
	Sample size=100			
$b = 80$	80.69(3.39)	80.35(3.30)	79.08(3.27)	78.70(3.35)
$\tau = 5$	5.08(0.31)	5.07 0.30)	5.04(0.29)	5.03(0.29)
$\mu_d = 0.67$	0.66(0.16)	0.66(0.16)	0.66(0.17)	0.67(0.17)
$\sigma^2 = 3.33$	3.53 (0.56)	3.50 (0.55)	3.41(0.53)	3.39(0.53)
	Sample size=200			
$b = 80$	81.11(2.85)	80.76(2.71)	79.50(2.52)	79.11(2.61)
$\tau = 5$	5.03(0.24)	5.03(0.23)	5.00(0.23)	4.99(0.23)
$\mu_d = 0.67$	0.64(0.095)	0.64(0.096)	0.65(0.10)	0.65(0.10)
$\sigma^2 = 3.33$	3.41 (0.34)	3.38(0.34)	3.29(0.33)	3.27(0.33)
	Sample size=500			
$b = 80$	80.65(1.62)	79.40(1.55)	81.00(1.81)	79.01(1.72)
$\tau = 5$	5.03(0.13)	5.00(0.13)	5.04(0.14)	4.99(0.13)
$\mu_d = 0.67$	0.66(0.052)	0.70(0.054)	0.66(0.052)	0.67(0.055)
$\sigma^2 = 3.33$	3.37(0.27)	3.28(0.27)	3.40(0.27)	3.27(0.27)