University of Pennsylvania
UPenn Biostatistics Working Papers

Year 2006 *Paper* 7

Survival Analysis Methods in Genetic Epidemiology

Hongzhe Li*

*University of Pennsylvania, hli@cceb.upenn.edu

Survival Analysis Methods in Genetic Epidemiology

Hongzhe Li

Abstract

Mapping genes for complex human diseases is a challenging problem due to the fact that many such diseases are due to both genetic and enviromental risk factors and many also exhibit phenotypic heterogeneity, such as variable age of onset. Information on variable age of disease onset is often a good indicator for disease heterogeneity and incorporation of such information together with enviromental risk factors into genetic analysis should lead to more powerful tests for genetic analysis. Due to the problem of censoring, survival analysis methods have proved to be very useful for genetic analysis. In this paper, I review some recent methodological developments on integrating modern survival analysis methods and human genetics in order to rigorously incorporate both age of onset and enviromental covariates data into aggregation analysis, segregation analysis, linkage analysis, association analysis and gene risk characterization. I also briefly discuss the issue of ascertainment correction and survival analysis methods for high-dimensional genomic data. Finally, I outline several areas that need further methodological developments.

Survival Analysis Methods in Genetic Epidemiology

Running title: Survival Analysis Methods

A Chapter in *"Current Topics in Human Genetics: Studies of Complex Diseases"*

(Eds. Hong-Wen Deng, Hui Shen, Yongjun Liu, Hai Hu)

Hongzhe Li

Professor of Biostatistics

Department of Biostatistics and Epidemiology

University of Pennsylvania School of Medicine

Philadelphia, PA 19104-6021, USA

Email: hli@cceb.upenn.edu

Abstract

Mapping genes for complex human diseases is a challenging problem due to the fact that many such diseases are due to both genetic and environmental risk factors and many also exhibit phenotypic heterogeneity, such as variable age of onset. Information on variable age of disease onset is often a good indicator for disease heterogeneity and incorporation of such information together with environmental risk factors into genetic analysis should lead to more powerful tests for genetic analysis. Due to the problem of censoring, survival analysis methods have proved to be very useful for genetic analysis. In this paper, I review some recent methodological developments on integrating modern survival analysis methods and human genetics in order to rigorously incorporate both age of onset and environmental covariates data into aggregation analysis, segregation analysis, linkage analysis, association analysis and gene risk characterization. I also briefly discuss the issue of ascertainment correction and survival analysis methods for high-dimensional genomic data. Finally, I outline several areas that need further methodological developments.

1 Introduction

The major burden of ill health in western society, and to a growing extent in developing societies, is due to complex chronic diseases such as coronary heart disease, stroke, breast cancer, prostate cancer, and diabetes. It is believed that both genetic and environmental factors contribute to both the risk of developing many of these common human diseases and also the responses to treatments. Because multiple genetic and environmental factors may play important roles in the susceptibility of individuals to develop these diseases and in the variation in treatment responses they are often referred to as complex traits. While the data necessary to study different complex traits are trait specific, the underlying principles and statistical methods of analysis of the genetic component are applicable to a variety of traits.

One important feature of many complex human diseases is disease heterogeneity due to genetic and other etiological factors. For example, many complex diseases exhibit variability in age of

onset, and early age of onset has been a hallmark for genetic predisposition in many diseases that aggregate in families. Therefore, age of onset outcomes such as age at diagnosis, are frequently gathered in genetic and epidemiological studies, including both genetic association and linkage studies. An important feature of age of onset data is the censorship resulting from being too young to develop the disease or death before developing the disease. This makes it possible for some of the unaffected siblings to share the disease gene with the affected siblings, who might be too young to exhibit the trait. In fact, affected relatives with different ages of onset may also be the result of different genetic etiologies. Age of onset data have been used to distinguish between two sub-forms of breast cancer (Claus *et al.*, 1990; Hall *et al.*, 1990) and prostate cancer (Carter *et al.*, 1992). For these adult onset cancers, carriers of high-risk alleles were estimated to have an earlier onset of cancer than noncarriers (sporadic cases). Taking into account age of onset information has been shown to be important in studying disease correlation and aggregation (Li *et al.*, 1998; Li and Thompson, 1997), in parametric linkage analysis (Morton and Kidd, 1980; Haynes *et al.*, 1986), in segregation analysis (Li and Thompson, 1997; Li *et al.*, 1998), and in allele-sharing based linkage analysis (Li and Zhong, 2002; Zhong and Li, 2004; Li *et al.*, 2002). A study by Li and Hsu (2000) also indicates that ignoring age of onset can reduce the power of both the allele-sharing-based linkage test and the transmission/disequilibrium test (TDT).

Another important feature of many complex traits is that many of these traits are known or suspected to be influenced by various environmental risk factors and interactions between genetic and environmental risk factors (G x E), e.g., breast cancer (Andrieu and Demenais, 1997) and rheumatoid arthritis (Brennan *et al.*, 1996). From a statistical standpoint, ignoring existing gene-environment interactions can result in underestimation of both genetic and environmental effects (Ottman, 1990), in incorrect conclusions with regard to the mode of inheritance and the magnitude of genetic effects in segregation analysis (Tiret *et al.*, 1993), and lower power in detecting genetic linkage (Towne *et al.*, 1997; Guo 2000a, 2000b).

Information on variable age of disease onset is often a good indicator for disease heterogeneity and incorporation of such information together with environmental risk factors into genetic analysis should lead to more powerful tests for genetic analysis. Due to the problem of cen-

soring, survival analysis methods, which are particularly developed for handling censoring, have proved to be very useful for genetic analysis. In this paper, I review some recent methodological developments in genetic epidemiology in order to rigorously take into account age of onset and environmental risk factors in aggregation analysis, segregation analysis, linkage and family-based association analysis and in gene risk characterization in the population. I also briefly discuss the issue of ascertainment correction and survival analysis methods for high-dimensional genomic data. Although I attempt a full and balanced treatment of most available literature, naturally the presentation leans in parts towards my own work. At the end of this review, I outline several areas that I think need further methodological developments, in particular, in the areas when high-throughput genomic data such as the genome-wide single nucleotide polymorphisms (SNPs) data are available.

2 Survival Analysis Methods for Aggregation Analysis

The purpose of aggregation analysis is to test whether disease aggregates within a family after some known environmental risk factors are taken into account. The ideal design is to collect a random sample of N families from the study population and to collect both age of disease onset/age at censoring data and the environmental risk factors of all the individuals within the families sampled. Then test of disease aggregation within family is equivalent to testing whether ages of onset of family members are correlated after adjusting for environmental risk factors.

I first define some notations that are used throughout this review. Suppose we have a collection of N families collected randomly or by some ascertainment criteria. Let the subscript ik indicate the ith individual in kth family, $i = 1, \cdots, m_k$, $k = 1, \cdots, N$. T_{ik} is the age at onset, C_{ik} the censoring age, t_{ik}=min(T_{ik}, C_{ik}), and $\delta_{ik} = I(t_{ik} = T_{ik})$, where $I(.)$ is the indicator function. The observed data are $(t_{ik}, \delta_{ik}, X_{ik})$, where X_{ik} is a p-dimensional vector of covariates that are independent of the genotype.

2.1 Shared frailty models based on random sample of families

The most commonly used model for assessing disease aggregation is the shared frailty model, which assumes the following conditional hazard function,

$$\lambda_{ik}(t|Z_k) = \lambda_0(t)\exp(X_{ik}\beta)Z_k, \tag{1}$$

where $\lambda_0(t)$ is the baseline hazard function, X_{ik} is the individual-specific covariate vector, β is the corresponding risk ratio parameters, and Z_k is the family-specific random effect or shared frailty. If $Z_1, \cdots, Z_k, \cdots, Z_N$ are assumed to be $i.i.d$ from a gamma distribution $\Gamma(\nu, \eta)$, where ν is the shape parameter and η is the scale parameter, the model is also called the gamma frailty model, Clayton or Clayton-Oakes model (Clayton, 1978; Oakes, 1982). For identifiability of $\lambda_0(t)$, it is assumed that $\nu = \eta$ so that $E(Z) = 1$. Estimation of such a model has been a subject of active research since the mid-eighties (Clayton and Cuzick, 1985; Self and Prentice, 1986; Klein, 1991; McGilchrist, 1993; Murphy, 1994; Nielsen *et al.*, 1992; Glidden and Self, 1999). The most common nonparametric maximum likelihood estimates (NPMLE) of the parameters can be obtained by the EM algorithm. Other assumptions on the frailty distribution include positive stable distribution (Hougaard, 1995) and log-normal distribution. Different distributions induce difference dependency structures of the age of onset within family. Glidden (1999) provides a model checking procedure for the gamma frailty model.

Under the shared frailty model (1), the null hypothesis of no disease aggregation can be formulated as testing

$$H_0 : var(Z) = \nu = 0.$$

For randomly sampled families, the standard inference procedure for the shared frailty model is based on the EM algorithm and likelihood ration test for H_0 (Klein, 1991; Nielsen *et al.*, 1992). The theoretical development was given by Murphy (1994) for the shared gamma frailty models.

2.2 Multivariate frailty models

The limitation of the shared frailty model for investigating disease aggregation within a family is that such a model assumes the same degree of dependency for any pairs of individuals within a

family, which is likely to be violated when disease aggregation is due to genetic segregation within the families. One way of extending the shared frailty model is to assume an individual specific additive frailty. For example, Peterson (1998) defined the following additive frailty model,

$$\lambda_{ik}(t|\eta_k) = \lambda_0(t)\exp(X_{ik}\beta)Z_{ik}, \quad (2)$$

$$Z_{ik} = Z_{k0} + Z_i, \quad (3)$$

where Z_{ik} is the individual-specific frailty, the frailty Z_{k0} is the shared frailty by family members in the kth family and is assumed to follow a $\Gamma(\nu_0, \eta)$, Z_i are the individual specific frailties and assumed to follow $\Gamma(\nu_1, \eta)$. For the purpose of identifiability of $\lambda_0(t)$, it is often assumed that $\nu_1 + \nu_0 = \eta$ so that $E(Z) = 1$. Under this additive gamma frailty model (2), the null hypothesis of interest is

$$H_0 : \nu_0 = 0.$$

Peterson (1998) presented an EM algorithm to obtain the NPMLE for the parameters and Parner (1998) developed the asymptotic theorem for the estimators and likelihood ratio test for H_0.

If the main goal is to test for disease aggregation due to genetic segregation, one should explicitly model the family dependence. Extensions of the frailty models to account for kinship relationship have been developed in recent years. One approach by Korsgaard and Andersen (1998) is in the framework of additive gamma frailty models, where the additive individual-specific frailties are explicitly defined based on gene segregation. Another approach is to assume that the log of the family-specific frailty vector $\log(Z_k) = \{\log(Z_{1k}), \cdots, \log(Z_{n_k k})\}$ follows a multivariate normal distribution $MVN(0, \Sigma)$, where the variance-covariance matrix is defined by the kinship coefficient matrix. Estimation of such multivariate normal frailty models includes the Monte Carlo or approximate EM algorithm (Palmgren and Ripatti, 2002) or the penalized partial likelihood approach (Ripatti and Palmgren, 2000).

2.3 Familiar aggregation based on case-control family design

As an alternative to family cohort design for assessing disease aggregation, the case-control sampling design is often used for rare diseases because a sufficient number of cases can be ascertained.

In family studies, investigators use the case-control design to enroll relatives for more detailed information, obtain medical records to validate reported disease, and obtain biospecimens for studies of genetic markers associated with the disease. Case-control family studies allow a direct examination of the disease outcomes in relatives and collection of both risk factor and exposure data on each individual, and with measured genetic markers, permits a more complete assessment of genetic and environmental factors through segregation and linkage analysis. Such a case-control study identifies a sample of diseased cases, and for each case, an independent sample of age-matched disease-free controls. For each identified individual (the proband), his environmental covariates, his family structure, and the disease status, age of onset (or age at censoring) and environmental covariates of his relatives are obtained.

When analyzing such case-control family data, one has to account for both the sampling issue and also the dependency of age of onset within the family. Estimating the marginal hazard function from the correlated failure time data arising from casecontrol family studies is complicated by non-cohort study design and risk heterogeneity due to unmeasured, shared risk factors among the family members. By assuming a Clayton multivariate survival model, Li *et al.* (1999) developed a procedure based on Prentice and Breslow's (1976) retrospective likelihood formulation assuming a parametric baseline hazard function. The method provides a way to combine the information relating disease incidence to risk factors in relatives with the information contained in the case-control contrasts in order to obtain more precise estimates of the effects of the putative risk factors. Shih and Chatterjee (2001) developed a similar estimation procedure but leave the baseline hazard function unspecified. Hsu *et al.* (2004) proposed a two-stage estimation procedure. At the first stage, we estimate the dependence parameter in the distribution for the risk heterogeneity without obtaining the marginal distribution first or simultaneously. Assuming that the dependence parameter is known, at the second stage we estimate the marginal hazard function by iterating between estimation of the risk heterogeneity (frailty) for each family and maximization of the partial likelihood function with an offset to account for the risk heterogeneity.

3 Survival Analysis Methods for Segregation Analysis

The goal of genetic segregation analysis is to develop a genetic model that best describes the disease aggregation within a family. Often it is assumed that a major gene with/without polygenes is involved in disease segregation. Segregation analysis based on parametric distributional assumptions on age of onset distribution is simple; instead, I review two semiparametric models developed for segregation analysis.

3.1 The Cox-Gene model for gene segregation

Assuming that a single major Mendelian diallelic locus governs the age-specific disease rate and the corresponding alleles are a and A, where A is the dominant disease allele with allele frequency $P(A) = q$, let g_{ik} be the genotype of ikth individual, taking one of three values aa, Aa, or AA. Li and Thompson (1998) developed the following Cox-Gene model. They assume that conditional on the unobserved major genotypes g_{ik}, ages of onset are assumed to be independent with a hazard function for the ikth individual:

$$\lambda_{ik}(t|X_{ik}, g_{ik}) = \lambda_0(t)\exp(\beta' X_{ik} + \mu_{ik}) \tag{4}$$

where

$$\mu_{ik} = \begin{cases} 0 & \text{if } g_{ik} = aa \\ \mu & \text{if } g_{ik} = Aa \text{ or } AA \end{cases}$$

under the assumption of a dominant mode of inheritance. The vector parameter β specifies the log of the risk ratios associated with the covariates, and $\lambda_0(t)$ is an unspecified baseline hazard function. Let $\Lambda_0(t) = \int_0^t \lambda(s)ds$ be the cumulative hazard function. Since A is the disease allele, we assume that $\mu \geq 0$. Li and Thompson (1998) further developed a Monte Carlo EM algorithm for estimating the parameters, especially for large pedigrees when the exact computation of the EM algorithm is not feasible.

Similar models were also developed and studied by Gauderman and Thomas (1994) and Siegmund and McKnight (1999). Chang *et al.* (2005) established the asymptotic properties

of the NPMLE from the EM algorithm. Chang *et al.* (2006) developed a faster algorithm for computing the NPMLE than the EM algorithm.

3.2 Cox model with major gene and random environmental effects for age of onset

The Cox-Gene model (4) assumes that the disease aggregation is due to segregation of one major gene, which accounts for all the correlation among the family members. To account for possible shared environmental effects, Li *et al.* (1997) defined a model to allow for both major gene effects and shared environmental effects by introducing a family-specific gamma random effect. Specifically, conditional on individual-specific major genotype g_{ik} and family-specific random environment ϵ_k, ages of onset are independent with the hazard function for the ik^{th} individual:

$$\lambda_{ik}(t|X_{ik}, g_{ik}, \epsilon_k) = \lambda_0(t)\epsilon_k \exp(\beta' X_{ik} + \mu_{g_{ik}}), \tag{5}$$

where $\mu_{g_{ik}} = 0$ if $g_{ik} = aa$ or μ if $g_{ik} = Aa$ or AA, is the genetic effect. The vector parameter β specifies risk ratios associated with the covariates X_{ik}, and $\lambda_0(t)$ is an unspecified baseline hazard function; $\Lambda_0(t) = \int_0^t \lambda(s)ds$ is the cumulative hazard function. The family effect, ϵ_k, is assumed to be *i.i.d.* gamma variate with mean 1 and unknown variance ν. This model incorporates the dependencies due to gene segregation and to shared environment. It is appropriate only for data on many families; variance ν is estimable only with a set of at least 3 families. The full model is specified by $\Theta = (\mu, q, \theta, \beta, \Lambda_0(t))$. If $\theta = 0$, $\epsilon_k = 1.0$ with probability 1 for all families, and model (5) reduces to the Cox-Gene model (4). If $\mu = 0$ or $q = 0$, model (5) reduces to the gamma frailty model. The parameters associated with the frailties are $\{\mu, q, \theta\}$. The genetic effects are measured by two parameters q and μ, where q measures the frequency of genetic susceptibility and μ measures the extent of genetic effects. The family-specific effects are measured by parameter ν; a larger value of ν corresponds to a stronger familial dependence due to common environmental effects and greater heterogeneity between families. Mendelian dependence of $\mu_{g_{ik}}$ makes it possible in theory to separate the genetic effects from the shared environmental effects and thus to identify and estimate the model parameters.

Li *et al.* (1997) developed an MCEM algorithm for estimating the model parameters and applied this model to analysis of a breast cancer family data set. The null hypothesis of interest includes $H_0 : \nu = 0$; when rejected, it implies that the major gene segregation cannot explain all the correlation of disease risks within families and additional genes or shared environmental factors may exist.

4 Survival Analysis Methods for Linkage Analysis

Linkage analysis examines the co-segregation of disease locus and markers or genomic loci within a family. Model-based linkage analysis often assumes a penetrance function and a specific mode of inheritance and tests whether the recombination fraction between the candidate disease locus and the marker locus is 0.5. Model-free allele-sharing-based linkage analysis is based on testing whether the probability distribution of identity-by-descent (IBD) among affected sib pairs deviates from the null probability or whether the distribution of the inheritance vector at a putative disease locus deviates from the null distribution under Mendelian segregation among the affected relatives (Kruglyak *et al.*, 1996). Incorporating age of onset or covariate data into parametric model-based linkage analysis is easy, simply by introducing age-dependent and covariate-dependent penetrance functions. In the following, I only review the survival analysis methods for allele-sharing-based linkage analysis based on the inheritance vectors.

4.1 Construction of genetic frailties for sibship

In order to adequately model the within-family dependency of age of onset variable to segregation of genes, the genetic frailties should be defined according to the law of Mendelian segregation. Li (2002) gave the following construction of the genetic frailties based on the concept of inheritance vectors (Kruglyak *et al.*, 1996; Lander and Green, 1987). Consider a sibship with n sibs, $1, 2, \cdots, n$, and denote their parents as F for the father and M for the mother. Assuming that the father and mother are unrelated, there are only four unique alleles that are distinct by descent at a given locus. Consider the setting of Kruglyak *et al.* (1996), where we have a series of markers

on a chromosomal region that may harbor the disease-causing locus/loci. Suppose d is a point in this test chromosomal region. We are interested in testing whether there is a disease-susceptible (DS) gene linked to locus d. Arbitrarily label the paternal chromosomes containing the locus of interest by (1,2), and label 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 the vector

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

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

Li and Zhong (2004) define the additive genetic frailties due to the gene linked to locus d for the father and mother as

$$\begin{aligned} Z_{dF} &= U_{d1} + U_{d2}, \\ Z_{dM} &= U_{d3} + U_{d4}, \end{aligned}$$

where U_{d1} and U_{d2} are used to represent the genetic frailties due to part of the genome on the two chromosomes of the father at locus d; U_{d3} and U_{d4} are analogous though for the mother. For a given inheritance vector v_d at the d locus for a sibship, we define the frailty for the jth sib as

$$Z_{dj} = U_{dv_{2j-1}} + U_{dv_{2j}}$$

for $j = 1, 2, \cdots, n$. This definition is based on the fact that it is the parts of the genome of the parents that are transmitted to the sibs, and the inheritance vectors indicate which parts are transmitted. We further assume that the U_{d1}, U_{d2}, U_{d3} and U_{d4} are independently and identically distributed across different families as $\Gamma(\nu_d/2, \eta)$, where the parameter η is the inverse scale parameter, and ν_d is the shape parameter. Then Z_{dj} is distributed as $\Gamma(\nu_d, \eta)$, for $j = 1, 2, \cdots, n$.

Taking into account possible genetic contributions to the disease not due to the single disease locus linked to d, e.g., due to loci unlinked to locus d, or contributions to shared familial effects, we add another random frailty term, U_p, to the genetic frailty, and define the genetic frailty for the jth sib as

$$Z_j = Z_{dj} + U_p$$

$$= \quad U_{dv_{2j-1}} + U_{dv_{2j}} + U_p.$$

We assume that U_p is distributed as $\Gamma(\nu_p, \eta)$ over different sibships. Then Z_j follows a $\Gamma(\nu_d+\nu_p, \eta)$ distribution. It is easy to verify that both the conditional (on V_d) and the marginal means of the frailties are

$$E(Z_1) = E(Z_2) = \cdots = E(Z_n) = \frac{\nu_d + \nu_p}{\eta},$$

and both the conditional and the marginal variances of the frailties are

$$Var(Z_1) = Var(Z_2) = \cdots = Var(Z_n) = \frac{\nu_d + \nu_p}{\eta^2}.$$

So the parameter ν_d can be interpreted as the proportion of the variance of the genetic frailty that can be explained by the gene linked to the locus d.

The frailties for a sibship can be written into a matrix form as

$$Z = HU, \tag{6}$$

where

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

$$H = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & 1 \\ & & \vdots & & \\ a_{n1} & a_{n2} & a_{n3} & a_{n4} & 1 \end{pmatrix},$$

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

where $a_{j1} = I(v_{2j-1} = 1), a_{j2} = I(v_{2j-1} = 2), a_{j3} = I(v_{2j} = 3), a_{j4} = I(v_{2j} = 4)$ for $j = 1, 2, \cdots, n$, where $I(.)$ is the indicator function.

4.2 The additive genetic gamma frailty model for sibship data

Consider a sibship with n 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 modeled by the proportional hazards model with random effect Z_j,

$$\lambda_j(t_j|Z_j) = \lambda_0(t)\exp(X_j'\beta)Z_j, \text{ for } j = 1, 2, \cdots, n, \tag{7}$$

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 unobserved genetic frailty constructed by equation (6) in the previous section. Since $Z_1, Z_2, \cdots, Z_n$ are dependent due to gene segregation and shared frailty, $T_1, T_2. \cdots, T_n$ are therefore dependent. Finally, to make the baseline hazard $\lambda_0(t)$ identifiable, we let $\nu_d + \nu_p = \eta$, which sets $E(Z_j) = 1, j = 1, 2, \cdots, n$, and prevents arbitrary scaling in model (7). Under this restriction, there are two free parameters, ν_d and ν_p, and $Z_{dj} \sim \Gamma(\nu_d, \nu_d+\nu_p)$, and $Z_p \sim \Gamma(\nu_p, \nu_d+\nu_p)$. We may also consider reparameterization in terms of the two frailty variances, $\sigma_d = Var(U_{dj}) = \nu_d/\eta^2$, and $\sigma_p = Var(U_p) = \nu_p/\eta^2$. Let $\sigma_{dp} = \sigma_d + \sigma_p$ denote the variance of Z_j. We then have $Z_{dj} \sim \Gamma(\sigma_d\sigma_{dp}^{-2}, \sigma_{dp}^{-1})$, and $Z_p \sim \Gamma(\sigma_p\sigma_{dp}^{-2}, \sigma_{dp}^{-1})$.

Based on this additive genetic gamma frailty model, Li and Zhong (2004) derived the joint survival function of age of onset data within a family as a function of the baseline hazard function, the covariate effects and the parameters related to the frailty. In addition, the null hypothesis that the disease locus is not linked to the candidate locus d can be reformulated as

$$H_0 : \nu_a = 0,$$

which is equivalent to assuming that the additive variance due to the gene linked to locus d is zero. In order to test this hypothesis, an estimate of the baseline hazard function is often required. However, the data collected for linkage analysis such as affected sib pairs or affected relatives do not often provide enough information for estimating such population-level baseline hazard functions. Instead, Li and Zhong (2004) developed a retrospective likelihood ration test for this null hypothesis assuming that the baseline hazard function can be estimated from external data such as the SEER database for various types of cancers.

Zhong and Li (2004) further extended the additive genetic gamma frailty model to simultaneously consider linkage to two unlinked loci and demonstrated that simultaneously searching for

two loci can result in increased power to detect linkage when the disease risk is affected by two unlinked genes. Instead of assuming gamma frailty models, one can also assume a log-multivariate normal frailty where the variance-covariance matrix is specified by kinship coefficients and pairwise IBD-sharing proportions (Pankratz *et al.*, 2005). Further, Pankratz *et al.* (2005) proposed a procedure using Laplace approximation for estimating model parameters and for testing linkage, i.e., testing whether the additive variance due to a given locus is zero.

5 Survival Analysis Methods for Family-based Genetic Association Analysis

Association studies look for specific alleles at a marker locus that are more frequent in affected individuals (cases) than in the unaffected population (controls). Population-based studies compare allele frequencies in cases and controls, but this methodology has been criticized as prone to false positives due to population admixture. To eliminate the effect of disequilibrium created by population stratification, and therefore to eliminate the false positive mapping results, family-based association methods such as haplotype relative risk (Falk and Rubinstein, 1987), the transmission disequilibrium test (TDT) (Spielman and Ewens, 1995; Spielman and Ewens, 1996), and a likelihood-based method (Schaid, 1996; Schaid and Li, 1997) using affected and family-based controls are often used. Li and Hsu (2001) demonstrated the importance of incorporating age of onset data into family-based genetic association analysis.

5.1 Survival analysis methods for family-based association tests

There are several approaches that extend the TDT to handle age of onset or age at censoring. Li and Fan (2000) proposed a linkage disequilibrium-based Cox (LDCox) model for nuclear family data and used a robust Wald's test for association. Mokliatchouk *et al.* (2001) and Shih and Whittemore (2002) developed likelihood-based score statistics to test for association between a disease and a genetic marker. The score statistic can be written as a weighted sum over family members of their observed minus expected genotypes. Age of onset data can be used

in the weight, which is the difference between the observed and expected value, $\delta_i - \Lambda_0(t_i)$ for individual i, where $\Lambda_0(t_i)$ is the cumulative hazard function at age t_i, which is assumed to be known from external data sources. Both methods of Li and Fan (2002) and Shih and Whittemore (2002) assume that the genetic effects on the risk of onset are proportional in the framework of the Cox regression model. Jiang *et al.* (2006) developed a family-based association test for time-to-onset data assuming time-dependent differences between the hazard functions among different genotype groups by using the weighted logrank approach of Flemming and Harrington (1981).

5.2 Test of genetic association in the presence of linkage

It is well known that genetic linkage induces within family association of phenotypes such as disease onset or age at disease onset. A limitation of most family-based association tests is that, although they remain valid tests of linkage, they are not valid tests of association if related nuclear families and or sibships from larger pedigrees are used. The allele-sharing-based linkage analysis only considers allele sharing by descent pattern among the sibs within a sibship. However, it does not differentiate which allele they share as long as they share it by descent. In other words, linkage analysis does not consider which particular allele is shared by the sibs. On the other hand, the association that we are interested in is the association due to LD. For association analysis and LD analysis, the particular allele that an individual carries determines his/her risk of developing disease, since different marker alleles have different coupling frequencies with the disease variant if LD exists. In typical tests of association, it is very rare that the genetic marker itself is the disease susceptible locus (DSL). When the marker locus is not the DSL but is in LD with it, all sibling resemblance or lack of resemblance and within sibship correlation of age of onset cannot be fully accounted for by the genotypes at the marker locus. Motivated by this key difference between linkage and LD, Zhong and Li (2004) defined a joint model for the risk of disease to account for both the allele sharing information and the genotype information at the candidate marker locus by including the genetic frailties derived from the inheritance vector. Specifically, consider a candidate marker d in the linked region and let $g = (g_1, \cdots, g_n)$ denote the vector of genotypes at marker locus d of the n sibs of known age at onset or censoring. Zhong

and Li (2004) assume that the hazard function of developing disease for the jth individual at age t_j is modeled by the proportional hazards model with random effect Z_j,

$$\lambda_j(t_j|Z_j) = \lambda_0(t_j)\exp(X_{g_j}\beta)Z_j, \text{ for } j = 1, 2, \cdots, n, \tag{8}$$

where $\lambda_0(t)$ is the unspecified baseline hazard function and X_{g_j} denotes some function of the jth offspring's marker genotype in the family. For example, for additive model, $X_{g_j} = l$, $l = 0, 1, 2$, counts the number of the putative high-risk marker allele and corresponds to the genotype of jth member in the family who carries l copies of the putative high-risk marker allele. Z_j is the unobserved genetic frailty, which is defined as in equation (6).

When $\beta = 0$, the hazard function (8) and the joint survival function for a sibship do not depend on the genotype at the marker locus d; therefore, tests of allelic association between locus d and the disease or the null hypothesis that the genotype at the marker locus is not associated with the risk of the disease can be formulated as testing

$$H_0 : \beta = 0.$$

Zhong and Li (2004) developed a score test for H_0 based on a retrospective likelihood function, which is a weighted sum over family members of their observed minus expected genotypes, where weights depend on both age of onset and also the IBD sharing between the sibs within a family. Different from score tests for linkage and association, the score test for testing association in the presence of linkage is also a function of allele-sharing IBD among the sib pairs or the inheritance distribution among the sibships. Zhong and Li (2004) demonstrated by simulations that such a score test indeed results in correct a type 1 error rate when testing for association in the linked region.

6 Survival Models for Haplotype Effects Based on Cohort, Case-Cohort and Nested Case-Control Designs

The most commonly used design for population-based haplotype analysis is the case-control design. Although case-control studies can potentially identify disease-predisposing variants, such

studies have certain limitations. This includes the tendency for clinically diagnosed cases to represent more severe ends of the whole disease spectrum and difficulty in selecting unbiased controls. In addition, such designs may suffer recall bias in disease status and other covariates such as family history (Doll, 1964; Collins, 2004). In contrast, large-scale population -based cohort studies can overcome these limitations. The prospective population cohorts can also enable the studies of many complex diseases in the same cohort. Large cohort studies are designed to learn about gene and environmental effects for relatively rare diseases. However, since many of the environmental covariates of interest are expensive to obtain, to reduce the cost of large cohort studies, several alternative sampling schemes within the framework of cohort studies have been suggested and well-studied and widely applied in the traditional epidemiological investigations. Among these, the most popular ones are the case-cohort design proposed by Prentice (1986) and the nested case-control design proposed by Thomas (1977) and Liddell *et al.* (1977). I review some recent methods for haplotype association analysis for cohort data, case-cohort data and nested case-control data. In order to account for variable age of onset, survival analysis methods are required for testing the haplotype effects for cohort, case-cohort and nested case-control designs.

6.1 Survival model for haplotype inference based on cohort data

Assume that N individuals are collected from a cohort and are type over K SNP markers. Consider the proportional hazards model to relate the disease risk to haplotype. Specifically, for the ith individual in the cohort, we assume the following Cox proportional hazards model

$$\lambda(t_i|X_i, H_i) = \lambda_0(t_i) \exp(\beta' \mathcal{F}(X_i, H_i)) \tag{9}$$

to relate the hazard function to the covariates vector X_i and the haplotype H_i, where $\mathcal{F}(X_i, H_i)$ is a known function to parameterize the covariates and the haplotype. Here the haplotype H_i can be over a set of SNPs in a candidate gene or SNPs in a sliding window in the whole-genome study. Depending on the model we choose, there are many different ways to parameterize the function $\mathcal{F}(X_i, H_i)$. For example, if h_0 is a particular haplotype of interest, we can assume the

following multiplicative model with haplotype and covariate interaction,

$$\mathcal{F}(X_i, H_i) = \beta_1(I(h_l = h_0) + I(h_m = h_0)) + \beta_2 X_i + \beta_3 X_i(I(h_l = h_0) + I(h_m = h_0)),$$

where (h_l, h_m) are the pair of the haplotype of H_i.

Lin (2004) proposed a likelihood-based approach and EM algorithm for estimating the parameter β and for haplotype inference for the proportional hazards model (9) in full cohort studies of unrelated individuals. Chen *et al.* (2004) derived a score test based on the partial likelihood function for testing the null hypothesis $H_0 : \beta = 0$, which is much easier to implement than the likelihood-based approach of Lin (2004). However, although the method of Lin provides estimate of the haplotype risk ratio parameters and the baseline hazard function, the method may suffer computational instability due to possible many rare haplotypes.

6.2 Test of haplotype association for case-cohort and nested case-control designs

Liddell *et al.* (1977) and Thomas (1977) suggested an alternative design called nested case-control design, in which a cohort is followed to identify cases of some disease of interest and then controls are selected for each case from within the cohort (i.e., controls are a random sample of unaffected individuals from the risk set in the cohort at the event time). Cases and controls can be matched on some covariates. In a such design, the covariates of interest are only measured for the cases and controls. For nested case-control data, Chen *et al.* (2004) developed a score test for $H_0 : \beta = 0$ in model (9). Alternatively, if the disease onset information is available for the full cohort, one can develop an EM algorithm for obtaining the NPMLE for the parameters associated with the model (9). An alternative design to the nested case-control design is the case-cohort design, as proposed by Prentice (1986) for large survey studies such as the Women's Health Study, where the population size makes it infeasible to collect data on all of the individuals in the cohort. If there is a concurrent registry that can be used to identify all of the subjects who experience an event, then it is possible to collect covariates data on only a sub-cohort of the subjects, randomly sampled from the population at large, and (perhaps at a later date) on those

subjects who experience an event. The sub-cohort in a given stratum constitutes the comparison set of cases occurring at a range of failure times (Prentice, 1986).

Since detailed procedures for haplotype analysis for case-cohort data have not been seen in literature, I provide some details on estimating the parameters under the case-cohort setting for the haplotype-disease risk model (9). For a case-cohort design, the data for individuals in R^+ (including case set R_1 and controls in the sub-cohort) are $D_i = (t_i, \delta_i, M_i, Z_i)$. However, the haplotype H_i may not be known for all individuals in R^+. Let $S(M_i)$ be the set of haplotype pairs consistent with genotype M_i. For individuals in R^- (those in the cohort but not in R^+), we only observe $D_i = (t_i, \delta_i)$, and for these individuals, let $S(M_i)$ be the set of all possible haplotypes. Denote $D = \{D_1, \cdots, D_N\}$ as the observed data, N is the number of individuals in the full cohort.

Let $f(Z)$ be the marginal distribution of the covariates Z in the population, and $G(t) = Pr(T > t)$. The likelihood function of the observed data is given by

$$\begin{aligned} L(\theta) &= \prod_{i \in R^+} \Big\{ \sum_{l,m} I(H_i(l,m) \in S(M_i)) \Big(\lambda_0(t_i) e^{\beta' \mathcal{F}(Z_i, H_i(l,m))} \Big)^{\delta_i} \exp\Big(-\Lambda_0(t_i) e^{\beta' \mathcal{F}(Z_i, H_i(l,m))} \Big) \pi_l \pi_m \Big\} \\ &\quad \times f(Z_i) \times \prod_{i \in R^-} G(t_i) \end{aligned} \tag{10}$$

where π_l, π_m are the haplotype frequencies of the haplotypes h_l and h_m, $H_i(l,m)$ represents the two haplotypes h_l and h_m for the ith individual, and $\theta = \{\Lambda_0(t), f(Z), \pi_l, \pi_m, \beta\}$. Note that here we assume that the Hardy-Weinberg equilibrium holds for the haplotypes, although this can be relaxed by introducing additional parameters. Instead of assuming a particular distribution of the covariates, we propose to deal with the distribution of Z nonparametrically as in Wellner and Zhan (1997) and Scheike and Juul (2004).

Since there are two nonparametric terms in this likelihood function, it is difficult to maximize it directly. We can develop an EM-algorithm instead. To write down the full data likelihood, we define W_j as the observed jth distinct combinations among the set $\{Z_i : i \in R^+\}$ for $j \in j = 1, \cdots, J$, and the corresponding point mass as $p_1, \cdots, p_J$ such that $\sum_j p_j = 1$. Then the missing data are the phases of the haplotypes for some individuals in R^+ and both the haplotypes and

covariates for individuals in R^-. The corresponding log full-data likelihood is

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{N}\Big\{\delta_i(\log\lambda_0(t_i)+\beta^{'}\mathcal{F}(Z_i,H_i))-\Lambda_0(t_i)e^{\beta^{'}\mathcal{F}(Z_i,H_i)}+\log f(Z_i)+\log Pr(H_i)\Big\} \\ &= \sum_{i=1}^{N}\Big\{\delta_i(\log\lambda_0(t_i)+\beta^{'}\mathcal{F}(Z_i,H_i))-\Lambda_0(t_i)e^{\beta^{'}\mathcal{F}(Z_i,H_i)}\Big\} \\ &+ \sum_{j=1}^{J}\sum_{i=1}^{N} I(Z_i=W_j)\log p_j+\sum_{l,m}\sum_{i=1}^{N} I(H_i=(h_l,h_m))\log(\pi_l\pi_m). \end{aligned} \tag{11}$$

E-Step: To implement the EM algorithm, we need to obtain the expectation of the equation (11), which requires the following expectations. First, for an individual $i \in R^+$, (M_i, Z_i) are known, but H_i may not be known,

$$E[I(H_i=(h_l,h_m))|D_i]=\frac{I(H_i(l,m)\in S(M_i))\exp\{\delta_i(\beta^{'}\mathcal{F}(Z_i,H_i(l,m)))-\Lambda_0(t_i)e^{\beta^{'}\mathcal{F}(Z_i,H_i(l,m))}\}\pi_l\pi_m}{\sum_{H_i(l',m')\in S(M_i)}\exp\{\delta_i(\beta^{'}\mathcal{F}(Z_i,H_i(l',m')))-\Lambda_0(t_i)e^{\beta^{'}\mathcal{F}(Z_i,H_i(l',m'))}\}\pi_l'\pi_m'},$$

and with this probability, $E(\mathcal{F}(Z_i,H_i))$ and $E(\exp(\mathcal{F}(Z_i,H_i)))$ can be derived. For an individual $i \in R^-$, we only observe $(t_i, \delta_i = 0)$,

$$E[I(Z_i=w_j,H_i=(h_l,h_m))|T_i>t_i]=\frac{p_j\exp\{-\Lambda_0(t_i)e^{\beta^{'}\mathcal{F}(W_j,H_i(l,m))}\}\pi_l\pi_m}{\sum_{j'=1}^{J}p_j'\sum_{H_{i'}(l,m)}\exp\{-\Lambda_0(t_i)e^{\beta^{'}\mathcal{F}(W_{j'},H_{i'}(l'm'))}\}\pi_{l'}\pi_{m'}},$$

with this probability, $E(\mathcal{F}(Z_i,H_i)$ and $E(\exp(\mathcal{F}(Z_i,H_i)))$ can be derived.

M-Step: It is easy to see that the EM equations in the M-step are

$$\hat{p}_j=\frac{\sum_{i=1}^{N}E[I(Z_i=W_j)|D]}{N},\ \text{for } j=1,\cdots,J$$

$$\hat{\pi}_l=\frac{\sum_{i=1}^{N}\sum_{j=1}^{J}E[I(Z_i=W_j)|D]\sum_{H_i(l,m)\in S(M_i)}E[I(H_i=(h_l,h_m))|D]}{2N}$$

$$\hat{\Lambda}_0(t)=\sum_{i=1}^{N}\frac{I(t_i\le t)\delta_i}{\sum_{j\in Y(t_i)}\sum_{j'=1}^{J}E[I(Z_j=W_{j'})|D]\sum_{H_j(l,m)}E[H_j=(h_l,h_m)|D]e^{\beta^{'}\mathcal{F}(W_{j'},H_j(l,m))}}$$

where $Y(t_i)$ is the set of individuals who were at risk at time t_i. Finally, the estimator of β is the root of the estimating function,

$$\begin{aligned} U(\beta) &= \sum_{i=1}^{N}\delta_i\Big\{\sum_{j=1}^{J}E[I(Z_i=W_j)|D]\sum_{l,m}E[I(H_i=(h_l,h_m))|D]\mathcal{F}(Z_i,H_i)^T \\ &- \frac{\sum_{j\in Y(t_i)}\sum_{j'=1}^{J}E[I(Z_j=W_{j'})|D]\sum_{H_j(l,m)}E[H_{j'}=(h_k,h_l)|D]e^{\beta^{'}\mathcal{F}(W_{j'},H_j(h,k))}(\mathcal{F}(Z_j,H_j)^T}{\sum_{j\in Y(t_i)}\sum_{j'=1}^{J}E[I(Z_j=W_{j'})|D]\sum_{H_j(l,m)}E[H_j=(h_k,h_l)|D]e^{\beta^{'}\mathcal{F}(W_{j'},H_j(h,k))}}\Big\}. \end{aligned}$$

This is the score equation corresponding to a Cox model with an individual-specific offset term, which can be easily solved by using the Newton-Ralphson iteration.

Based on different ways of parameterizing the haplotype effects, the test of haplotype effects can be in general formulated as testing $H_0 : \beta_1 = 0$, where β_1 is a sub-vector of $\beta = \{\beta_1, \beta_2\}$. Similar to Lin (2004) and Scheike and Juul (2004), the likelihood ratio test can be applied for this null hypothesis.

7 Survival Analysis Methods for Gene Characterization

After the genetic variants related to the risk of disease are identified, it is important to estimate the penetrance of the variants and other population based parameters such as the allele frequencies. Cohort or case-control family designs can be used for gene characterization and for estimating population parameters such as genotype relative risk and age-dependent penetrance functions. For rare diseases, often a large cohort is required for estimating such population parameters. For case-control family designs, if the genotypes of the disease variants are available for all the family members, the methods by Li *et al.* (1998) and Shih and Chatterjee (2002) can be used for estimating the age-dependent penetrance functions.

When genotypes of the family members are not available, the kin-cohort design (Wacholder *et al.*, 1998) is a promising alternative to traditional cohort or case-control family designs for estimating penetrance of an identified rare autosomal mutation. In such a design, a suitably selected sample of participants provides genotype and detailed family history information on the disease of interest; however, the genotypes of the family members are not known. Gail *et al.* (1999) used the term "genotyped probands" to emphasize that the probands are genotyped in kin-cohort design. To estimate penetrance of the mutation, Chatterjee and Wacholder (2001) considered a marginal likelihood approach that is computationally simple to implement, more flexible than the original analytic approach proposed by Wacholder *et al.* (1998) and more robust than the likelihood approach considered by Gail *et al.* (1999) to the presence of residual familial correlation. Chatterjee *et al.* (2005) further extended the approach of Shih and Chatterjee

(2002) for data from the kin-cohort design with both cases and control probands and the kin-cohort design with only cases in order to account for residual correlations. In order to allow for residual familial aggregation given genotypes, Chatterjee *et al.* (2005) consider a copula models (Genest and MacKay, 1986) for specifying joint risks of the disease among the proband and his/her family members. The key of these various approaches is to make inference based on the likelihood function that corrects for ascertainment.

8 Ascertainment Correction

Different from traditional multivariate survival analysis, one of the most difficult problems in analyzing family data in genetic studies is that the families for genetic analysis are often not random samples from the population; rather, they are often ascertained because of one or more of the family members are affected with the disease of interest. This ascertainment problem makes statistical inferences for the proposed models in this paper difficult. For the ascertained family samples, estimating the population baseline hazard function becomes even more difficult. One way to go around this problem is by using a retrospective likelihood, which is defined as the probability of marker data given the observed age of onset data. In order to maximize such a likelihood function, the baseline hazard function is often assumed to be known or to follow some parametric form. Due to conditioning, one may expect loss of efficiency in parameter estimates. Sun and Li (2004) recently proposed and evaluated two approaches based on conditional prospective likelihood and conditional ascertainment corrected likelihood for the additive genetic gamma frailty model in order to estimate the baseline hazard function based on the family data collected for linkage analysis. However, such an ascertainment correction procedure requires knowledge of the population distribution of the family structures and family sizes, which can be difficult to obtain.

9 Survival Analysis Methods in the Genomics Era

Recent development of new high-throughput technologies for generating very high-dimensional genomic data such as microarray gene expression data raises other important and interesting problems that require development of new survival analysis methods. One such area is to link the microarray gene expression data to censored survival outcomes such as cancer recurrence. Due to high-dimensionality of the data, traditional survival analysis methods cannot be applied directly to such data sets or are expected to perform poorly.

Currently, there are several classes of approaches for these type of censored data regression problems in the high-dimension and low sample-size settings. One class of approaches is dimension-reduction-based methods, such as extensions of the partial least square regression method for censored data regression problems (Park *et al.*, 2001; Li and Gui, 2004), extension of the slice inverse regression method (Li and Li, 2004) and supervised principal components analysis (Bair and Tibshirani, 2004). While these methods may perform well in prediction, they usually do not provide a direct way of selecting genes that are potentially related to time-to-event.

Another class of approaches is based on regularized estimation procedures such as L_2 penalized estimation (Li and Luan, 2004), the extension of the least absolute shrinkage and selection operator (Lasso) of Tibshirani (1996) to censored survival data using the least angle regression (LARS) (Efron *et al.*, 2004; Gui and Li, 2005; Segal, 2006), and the threshold gradient descent procedure (Friedman and Popescu, 2004; Gui and Li, 2005). These methods provide a way of selecting genes whose expression might be related to clinical outcome such as time-to-event. In addition, these methods can also be used for building a model for predicting future patients' time-to-event.

Survival ensembles, based on extensions of the random forests (Breiman, 2001) and the gradient descent boosting procedure (Friedman, 2001) to censored survival data, have also been developed recently (Li and Luan, 2005; Hothorn *et al.*, 2006). These procedures are more flexible and usually perform better in predicting future patients' time-to-event.

10 Conclusion and Future Directions

Since many complex diseases show large variation in age at disease onset, consideration of variable age of disease onset is an important aspect of genetic analysis of complex diseases. Methods in survival analysis provide a natural framework for incorporating age of onset and environmental risk factors into genetic analysis. In this paper, I have reviewed some recently developed survival analysis methods for aggregation, segregation, linkage and association analysis and gene characterization analysis in genetic epidemiology. Most of these methods were developed in the last ten years and have been shown to be able to offer additional insights into genetic studies of complex diseases. As user-friendly software packages implementing these methods become available, we should expect to see more applications of these methods in mapping genes for complex diseases.

With the completion of the Human Genome Project and the HapMap project, genome-wide association studies of complex traits are now possible and have already been proposed for several complex diseases. Under such studies, hundreds of thousands of SNPs are typed for a large set of patients and controls. In addition, large-scale cohort studies are under discussion or are already underway in the UK (UK Biobank), Iceland (Decode), Germany, Canada and Japan. The US is also considering to propose its own large-scale population cohort (Collins, 2004). We therefore expect that large amounts of data will be generated from these large cohort studies in the near future. Besides large cohort data, case-cohort and nested case-control designs offer alternatives to cohort and case-control designs. An important research question is how to identify SNPs, SNP-SNP interactions, gene-gene interaction, gene-environment interactions among hundreds of thousands of SNPs that may affect the disease risk based on case-cohort or nested case control data in the framework of survival analysis. In addition, many common diseases are known to be affected by certain genotype combinations; therefore, statistical methods to detect the influential genes along with their interaction structures are also required. Finally, new statistical methods are also required in order to fully utilize the genome-wide linkage disequilibrium patterns and the haplotype block structures available from the HapMap project. New ideas from statistical learning (Hastie *et al.*, 2001) hold great promise to address these important issues.

Since genes and proteins almost never work alone, they interact with each other and with

other molecules in highly structured but incredibly complex ways. Understanding this interplay of human genome and environmental influences is crucial to developing a systems understanding of human health and disease. An important venue for future research is to develop methods that can incorporate known biological knowledge such as pathways into statistical modeling in order to limit the search space for gene-gene and gene-environment interactions (Conti *et al.*, 2003; Wei and Li, 2006). Wei and Li (2006) proposed a non-parametric pathways-based regression model to incorporate pathways information into regression analysis. As biological knowledge accumulates, one should expect to see development of new methods and more applications of these models in identifying genes and environmental risk factors that are related to risk of developing disease.

Acknowledgments

This research was supported by NIH grant ES009911. I thank Mr. Edmund Weisberg, MS at Penn CCEB for editorial assistance.

References

Andrieu N, Demenais F (1997): Interaction between genetic and reproductive factors in breast cancer risk in a French family sample. *American Journal of Human Genetics*, 61: 678-690.

Bair E, Tibshirani R (2004): Semi-supervised methods for predicting patient survival from gene expression papers. *PLoS Biology*, 2:5011-5022.

Breiman L (2001): Random Forests. *Machine Learning*, 45:5-32.

Brennan P, Ollier B, Worthington J, Hajeer A, Silman A (1996): Are both genetic and reproductive associations with rheumatoid arthritis linked to prolactin. *Lancet*, 348: 106-109.

Carter BS, Beaty HB, Steinberg GD, Childs B, Walsh PC (1992): Mendelian Inheritance of Familial Prostate Cancer. *Proceedings of National Academy of Sciences USA*, 89: 3367-3371.

Chang IS, Hsiung CA, Wang MC, Wen CC (2005): An asymptotic theory for the nonparametric maximum likelihood estimator in the Cox gene model. *Bernoulli*,11(5): 863-892.

Chang IS, Wen CC, Wu YJ and Yang CC (2006): Fast algorithm for the nonparametric maximum likelihood estimate in the Cox-gene model. *Statistica Sinica*, in press.

Chatterjee N, Kalaylioglu Z, Shih JH and Gail MH (2005): CaseControl and Case-Only Designs with Genotype and Family History Data: Estimating Relative Risk, Residual Familial Aggregation, and Cumulative Risk. *Biometrics*, accepted.

Chatterjee N and Wacholder S (2001): A Marginal Likelihood Approach for Estimating Penetrance from Kin-Cohort Designs. *Biometrics*, 57 (1):245-252.

Chen J, Peters U, Foster C, Chatterjee N (2004): Haplotype-based test of genetic association using data from cohort and nested case-control epidemiologic studies. *Human Heredity*, 58:18-29.

Claus EB, Risch NJ, Thompson WD (1990): Using Age of Onset to Distinguish Between Subforms of Breast Cancer. *Annals of Human Genetics*, 54: 169-177.

Clayton DG (1978): A Model for Association in Bivariate Life Tables and Its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence. *Biometrika*, 65: 141-151.

Clayton DG, Cuzick J (1985): Multivariate generalizations of the proportional hazards model. *Journal of Royal Statistical Association Series A*, 148:82-117.

Collins FS (2004): The case for a US prospective cohort study of genes and environment. *Nature*, 429:475-477.

Conti DV, Cortessis V, Molitor J and Thomas DC (2003): Bayesian modeling of complex metabolic pathways. *Human Heredity*, 56: 83-93.

Doll R (1964): Retrospective and prospective studies. In *Medical Surveys and Clinical Trials* (L.J. Witts, ed), pp 96-97.

Efron B, Johnston I, Hastie T and Tibshirani R (2004): Least angle regression. *Annals of Statistics*,32:407-499.

Falk CT, Rubinstein P (1987): Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Annals of Human Genetics*, 51: 227-233.

Flemming TR and Harrington DP (1981): A class of hypothesis tests for one and two sample censored survival data. *Communications in Statistics, Theory and Methods*, 10: 763-794.

Friedman (2001): Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29: 1189-1232.

Friedman JH and Popescu BE (2004): Gradient Directed Regularization for Linear Regression and Classification. *Technical Report, Statistics Department, Stanford University.*

Gauderman WJ, Thomas D (1994): Censored survival models for genetic epidemiology: a Gibbs sampling approach. *Genetic Epidemiology*, 11:171-188.

Gail M, Pee D, Benechou J, Carroll R (1999): Designing studies to estimate the penetrance of an identified autosomal mutation: cohort, case-control and genotypes-proband design. *Genetic Epidemiology*, 16: 15-39.

Genest C and MacKay (1986): The joy of copulas: bivariate distributions with given margins. *American Statistician*, 40: 280-283.

Gliden DV (1999): Checking the adequacy of the gamma frailty model for multivariate failure times. *Biometrika*, 86: 381-393.

Glidden DV and Self SG (1999): Semiparametric Likelihood Estimation in the Clayton-Oakes model. *Scandinavian Journal of Statistics*, 26: 363-372.

Gui J and Li H (2005): Penalized Cox Regression Analysis in the High-Dimensional and Low-sample Size Settings, with Applications to Microarray Gene Expression Data. *Bioinformatics*, 21:3001-3008.

Gui J, Li H (2005): Threshold Gradient Descent Method for Censored Data Regression, with Applications in Pharmacogenomics. *Pacific Symposium on Biocomputing*, 10:272-283.

Guo SW (2000a): Gene-environment interactions and the affected-sib-pair designs. *Human Heredity* 50:271-285.

Guo SW (2000b): Gene-environment interaction and the mapping of complex traits: some statistical models and their implications. *Human Heredity*, 50:286-303.

Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King MC (1990): Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*, 250: 1684-1689.

Hasite T, Tibshirani R, Friedman J (2001): *The Elements of Statistical Learning.* Springer, New York.

Haynes C, Pericak-Vance MA, Dawson D (1986): Genetic analysis workshop IV, Analysis of Huntington's disease linkage and age of onset curves. *Genetic Epidemiology*, 1: 109-122.

Hothorn T, Buhlmann P, Dudoit S, Molinaro A, van der Laan MJ (2006): Survival Ensembles. *Biostatistics*, in press.

Houggaard P (1995): Frailty models for survival data. *Lifetime Data Analysis*, 1: 255-273.

Hsu L, Chen L, Gorfine M, Malone K (2004): Semiparametric Estimation of Marginal Hazard Function from CaseControl Family Studies *Biometrics*, 60: 936-944.

Hsu L, Li H , Houwing J (2002): A method for incorporating ages at onset in affected sib pair linkage studies, *Human Heredity*, 54: 1-12 .

Jiang H, Harrington D, Raby BA, Bertram L, Blacker D, Weiss ST and Lange C (2006): Family-based association test for time-to-event data with time-dependent difference between hazard functions. *Genetic Epidemiology*, 30: 124-132.

Klein JP (1992): Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, 48: 795-806.

Korsgaard IR, Anderson AH (1998): The additive genetic gamma frailty model. *Scandinavia Journal of Statistics*, 25: 255-269.

Kruglyak L, Daly MJ, Reeve-Daly MP and Lander ES (1996): Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics* 8: 1347-1363.

Kruglyak L and Lander ES (1995): Complete multipoint sib-pair analysis of qualitative and quantitative traits. *American Journal of Human Genetics*, 57: 439-454.

Lander E and Green P (1987): Construction of multilocus genetic maps in humans. *Proceedings of National Academy of Sciences USA*, 84: 2363-2367.

Li H (2002): An Additive Genetic Gamma Frailty Model for Linkage Analysis of Diseases with Variable Age of Onset Using Nuclear Families. *Lifetime Data Analysis*, 8:315-334.

Li H, Gui J (2004): Partial Cox Regression Analysis for High-Dimensional Microarray Gene Expression Data. *Bioinformatics*, 20:i208-i215.

Li H, Hsu L (2000): Effects of Ages at Onset on the Power of the Affected Sib Pair and Transmission/Disequilibrium Tests. *Annals of Human Genetics*, 64:239-254.

Li L, Li H (2004): Dimension Reduction Methods for Microarrays with Application to Censored Survival Data. *Bioinformatics*, 20:3406-3412.

Li H, Luan Y (2003): Kernel Cox regression models for linking gene expression profiles to censored survival data. *Pacific Symposium on Biocomputing*, 8: 65-76.

Li H , Luan Y (2005): Boosting Proportional Hazards Models Using Smoothing Splines, with Applications to High-Dimensional Microarray Data. *Bioinformatics*, 21: 2403-2409.

Li H, Thompson EA (1997): Semiparametric estimation of major gene and random familial effects for age of onset. *Biometrics* 53, 282-293.

Li H, Thompson EA, Wijsman EA (1998): Semiparametric estimation of major gene effects for age of onset. *Genetic Epidemiology*, 15:279-298.

Li H, Yang P, Schwartz AG (1998): Analysis of age of onset data from case-control family studies. *Biometrics,* 54,1030-1039.

Li H , Zhong X(2002): Multivariate survival models induced by genetic frailties, with application to linkage analysis. *Biostatistics*, 3: 57-75.

Liddell FDK, McDonald JC, Thomas DC (1977): Methods of cohort analysis: Appraisal by application to Asbestos Mining. *Journal of Royal Statistical Society*, 4:469-491.

Lin DY (2004): Haplotype-based association analysis in cohort studies of unrelated individuals. *Genetic Epidemiology*, 26:255-265.

McGilchrist CA (1993): REML estimation for survival models with frailty. *Biometrics*, 47: 461-466.

Mokliatchouk O, Blacker D and Rabinowitz (2001): Association tests for traits with variable age at onset. *Human Heredity*, 51: 46-53.

Morton LA, Kidd KK (1980): The effects of variable age-of-onset and diagnostic criteria on the estimates of linkage: An example using manic-depressive illness and color blindness. *Social Biology* 27, 1-10.

Murphy SA (1994): Consistency in a proportional hazards model incoporaring random effects. *Annals of Statistics*, 22: 712-731.

Nielsen GG, Gill RD, Andersen PK and Sorensen TIA (1992): A counting process approach to maximum likelihood estimation in frailty models. *Scandinavia Journal of Statistics*, 19: 25-43.

Oakes, D (1982): A Model for Association in Bivariate Survival Data. *Journal of the Royal Statistical Society, Series B* 44,414-422.

Ottman R (1990): An epidemiologic approach to gene-environment interaction. *Genetics Epidemiology* 11, 75-86.

Park PJ, Tian L, Kohane IS (2002): Linking expression data with patient survival times using partial least squares. *Bioinformatics* 18, S120-127.

Parner E (1998): Asymptotic theory for the correlated gamma-frailty model. *Annals of Statistics* 26: 183-214.

Pankratz VS , de Andrade M , Therneau TM (2005): Random-effects cox proportional hazards model: General variance components methods for time-to-event data. *Genetic Epidemiology*, 28(2):97-109.

Petersen JH (1998): An additive frailty model for correlated life times. *Biometrics*, 54: 646-661.

Prentice RL (1986): A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73:1-11.

Ripatti S, Larsen K, Palmgren J (2002): Maximum likelihood inference for multivariate frailty models using a Monte Carlo EM algorithm. *Lifetime Data Analysis* 8:349-360.

Ripatti S, Palmgren J (2000): Estimation of Multivariate Frailty Models using Penalized Partial Likelihood. *Biometrics*, 56:1016-22.

Schaid DJ (1996): General score tests for associations of hgenetic markers with disease using cases and their parents. *Genetic Epidemiology*, 13: 423-449.

Schaid DJ, Li H (1997): Genotype relative-risks and association tests for nuclear families with missing parental data. *Genetic Epidemiology*, 14:1113-1118.

Scheike TH and Juul A (2004): Maximum likelihood estimation for Cox's regression model under nested case-control sampling. *Biostatistics*, 5: 193 - 206.

Self SG and Prentice RL (1986): Incorporating Random Effects into Multivariate Relative Risk Regression Models. In *Modern Statistical Methods in Chronic Disease Epidemiology*, edited by H. Moolgavkar, and R.L Prentice , John Wiley & Sons.

Segal MR (2006): Microarray Gene Expression Data with Linked Survival Phenotypes: Diffuse Large-B-Cell Lymphoma Revisited. *Biostatistics*, in press.

Shih MC and Whittemore AS (2002): Tests for genetic association using family data. *Genetic Epidemiology* 22: 128-145.

Spielman RS, Ewens WJ (1996): The TDT and other family-based tests for linkage disequilibrium and association. *American Journal of Human Genetics* **59**, 983-989.

Spielman RS, Ewens WJ (1998): A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *American Journal of Human Genetics* **62**, 450-458.

Sun W, Li H (2004): Ascertainment-Adjusted Maximum Likelihood Estimation for the Additive Genetic Gamma Frailty Models. *Lifetime Data Analysis*, 10:229-245.

Thomas DC (1977): Addendum to: Methods of cohort analysis: Appraisal by application to Asbestos Mining, by Liddell FDK, McDonald JC, Thomas DC. *Journal of Royal Statistical Society*, 4:469-491.

Tibshirani R (1996): Regression shrinkage and selection via the Lasso. *Journal of Royal Statistical Society B* 58:267-288.

Tiret L, Rigat B, Visvikis S, Breda C, Corvol P, Cambien F, Soubrier F (1992): Evidence, from combined segregation and linkage analysis, that a variant of the angiotensin I-converting enzyme (ACE) gene controls plasma ACE levels. *American Journal of Human Genetics*, 51:197-205.

Hothorn T, Buhlmann P, Dudoit S, Molinaro A, and van der Laan MJ (2006): Survival Ensembles. *Biostatistics*, in press.

Towne B, Siervogel RM, Blangero J (1997): Effects of genotype-by-sex interaction on quantitative trait linkage analysis. *Genetic Epidemiology*, 14: 1053-1058.

Wacholder S, Hartge P, Sruewing JP, Pee D, McAdams M, Lawrance BC, Tucker MA (1998): The kin-cohort study for estimating penetrance. *American Journal of Epidemiology*, 148:623-630.

Wei Z, Li H (2006): Nonparametric Pathway-Based Regression Models for Analysis of Genomic Data. UPenn Biostatistics Working Papers. UPenn Biostatistics Working Paper Series. Working Paper 6. http://www.biostatsresearch.com/upennbiostat/papers/art6.

Wellner JA, Zhan Y (1997): A hybrid algorithm for computing the nonparametric maximum likelihood estimator from censored data. *Journal of American Statistical Association*, 92:945-959.

Zhong X, Li H (2002): An additive genetic gamma frailty model for two-locus linkage analysis using sibship age of onset data. *Statistical Applications in Genetics and Molecular Biology*, Vol 1, No 1, article 2.

Zhong X, Li H (2004): Score tests of genetic association in the presence of linkage based on the additive genetic gamma frailty model. *Biostatistics*, 5:307-327.