Published in final edited form as:
Lifetime Data Anal. 2014 January ; 20(1): 16–22. doi:10.1007/s10985-013-9262-8.

Survival analysis with incomplete genetic data

D. Y. Lin
Department of Biostatistics, University of North Carolina at Chapel Hill

Abstract

Genetic data are now collected frequently in clinical studies and epidemiological cohort studies. For a large study, it may be prohibitively expensive to genotype all study subjects, especially with the next-generation sequencing technology. Two-phase sampling, such as case-cohort and nested case-control sampling, is cost-effective in such settings but entails considerable analysis challenges, especially if efficient estimators are desired. A different type of missing data arises when the investigators are interested in the haplotypes or the genetic markers that are not on the genotyping platform used for the current study. Valid and efficient analysis of such missing data is also interesting and challenging. This article provides an overview of these issues and outlines some directions for future research.

Keywords

Case-cohort design; Case-control design; Censoring; Genome-wide association studies; Haplotypes; Next-generation sequencing; Nonparametric likelihood; Single nucleotide polymorphisms; Two-phase study; Women’s Health Initiative

1. INTRODUCTION

Thanks to the availability of detailed genetic maps across the human genome (e.g., The International Human Genome Sequencing Consortium, 2001; The International HapMap Consortium, 2005) and precipitous drops in genotyping costs, there has been a worldwide proliferation of genetic association studies, which explore population relationships between disease phenotypes and genetic variants, particularly single nucleotide polymorphisms (SNPs). Many of these studies survey the entire genome with high-density genotyping chips containing several million SNPs; such studies are referred to as genome-wide association studies (GWAS). Recent technological advances have made it possible to conduct sequencing studies, which determine the entire DNA sequence of individual genes, all exomes or whole genomes.

Most genetic association studies employ the case-control design, which genotypes a sample of diseased individuals and a sample of disease-free individuals. This design is particularly appealing for rare diseases. Cohort studies offer several advantages over case-control studies. First, the age of onset carries more information about the etiology of a disease than the disease status. Secondly, selection and information biases inherent in case-control studies can usually be removed in cohort studies. Thirdly, the cohort design enables one to investigate a wide range of diseases and related phenotypes in a single study.

Address for correspondence: D. Y. Lin, Department of Biostatistics, CB#7420, University of North Carolina, Chapel Hill, NC 27599-7420, USA. lin@bios.unc.edu.

Cohort studies are major undertakings, requiring long-term follow-up of many individuals, especially for rare diseases. Fortunately, there are a number of cohort studies that have already been assembled for other purposes and have repositories of stored specimens that allow the individuals to be genotyped for the genes of interest. Examples of well-established cohorts include the Cardiovascular Health Study (Fried et al. 1991), the Women's Health Initiative (Johnson et al. 1999) and the Atherosclerosis Risk in Communities (ARIC) Study (The ARIC Investigators 1989).

Despite the continuing improvement in genotyping efficiency, it is still expensive to genotype a large cohort. A cost-effective strategy is to employ the case-cohort or nested case-control design (Prentice, 1986; Thomas, 1977), so that only a subset of the cohort members need to be genotyped. Such designs pose a challenging missing-data problem in that the subjects who are not selected for genotyping have no genotype data.

A different type of missing data arises in the analysis of haplotype-disease association. A haplotype is a specific sequence of nucleotides on the same chromosome of a subject. Because current genotyping technologies do not separate a subject's two homologous chromosomes, haplotypes are not directly observed. A related form of missing data arises when investigators are interested in untyped SNPs, i.e., the SNPs that are not even on the genotyping chip used in the study and are thus missing on all study subjects.

In this article, we describe statistical methods for handling the aforementioned missing-data problems. The methods are based on nonparametric maximum likelihood estimation. Kalbfleisch and Prentice (2002) provided extensive coverage of likelihood construction for survival data. In particular, they presented a general form of likelihood for missing covariate data (Kalbfleisch and Prentice, 2002, p. 344). In this article, we show how to apply the general principle of Kalbfleisch and Prentice (2002) to the two specific missing-data problems mentioned above. We will discuss some related open problems.

2. TWO-PHASE SAMPLING

Let T denote the failure time, G denote the genetic variable, and X denote a set of environmental factors. We specify that the hazard function of T conditional on G and X satisfies the proportional hazards model (Cox, 1972)

$$\lambda(t|G, X)=\lambda_0(t)e^{\theta^{\mathrm{T}} Z(G,X)},$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard function, Z is a vector-function of G and X, and θ is a set of unknown regression parameters. Under the commonly assumed additive mode of inheritance, G pertains to the number of the "risk" allele the subject carries at the locus of interest.

In practice, T is subject to right censoring. Let C denote the censoring time. For each T, we observe Y and Δ, where $Y = \min(T, C)$, $\Delta = I(T \le C)$, and $I(\cdot)$ is the indicator function. For a cohort of n subjects, the data potentially consist of $(Y_i, \Delta_i, X_i, G_i)$ $(i = 1, \ldots, n)$.

For a large cohort, it may be prohibitively expensive to genotype all cohort members. A cost-effective strategy to employ a two-phase design. In the case-cohort design (Prentice, 1986), all the cases (i.e., the subjects with $\Delta_i = 1$) and a random sample of controls (i.e., the subjects with $\Delta_i = 0$) are selected for genotyping. In the nested case-control design (Thomas, 1977), a small number of controls, typically between 1 and 5, are selected for each case. In the original case-cohort and nested case-control designs (Prentice, 1986; Thomas, 1977), both the G_i's and X_i's are measured only on the selected subjects. We consider the more realistic scenario in which the X_i's are available on all cohort members. We allow the selection to depend on any aspects of the data (Y_i, Δ_i, X_i) $(i = 1, \ldots, n)$. In particular, some cases may not be selected, and the selection probabilities may depend on the X_i's. For $i = 1, \ldots, n$, let R_i indicate, by the values 1 versus 0, whether G_i is measured or not. Then the observed data consist of $(Y_i, \Delta_i, X_i, R_i, R_iG_i)$ $(i = 1, \ldots, n)$.

The likelihood function takes the form

$$\prod_{i=1}^{n}\{P(Y_i,\Delta_i|G_i,X_i)P(G_i|X_i)\}^{R_i}\left\{\sum_{g}P(Y_i,\Delta_i|g,X_i)P(g|X_i)\right\}^{1-R_i}, \quad (1)$$

where $P(\cdot|\cdot)$ denotes conditional density functions (Kalbfleisch and Prentice, 2002, p. 344). Under the assumption that C is independent of T and G conditional on X, $P(Y, \Delta|G, X)$ is proportional to

$$\left\{\lambda_0(Y)e^{\theta^{\mathrm{T}}Z(G,X)}\right\}^{\Delta}\exp\left\{-\Lambda_0(Y)e^{\theta^{\mathrm{T}}Z(G,X)}\right\}, \quad (2)$$

where $\Lambda_0(t)=\int_0^t\lambda_0(u)\,du$ (Kalbfleisch and Prentice, 2002, p. 54). The likelihood function given in (1) is not tractable if X is continuous and correlated with G. We assume that X is independent of G or is discrete such that $P(G|X)$ is a discrete probability function.

Note that (1) is a nonparametric likelihood in that $\lambda_0(\cdot)$ is an infinite-dimensional parameter. To maximize (1), we let the estimator for Λ_0 be a step function with jumps only at the observed Y_i with $\Delta_i = 1$ and replace $\lambda_0(Y)$ in (2) by the jump size of $\Lambda_0(\cdot)$ at Y. The maximization can be carried out by an EM algorithm (Zeng et al., 2006). The resulting estimator of θ is consistent, asymptotically normal, and asymptotically efficient (Zeng et al., 2006). The limiting covariance matrix can be estimated by the profile likelihood method (Murphy and van der Vaart, 2000).

3. HAPLOTYPES

We consider a set of SNPs that are correlated. We may have a direct interest in the haplotypes of these SNPs or wish to use the haplotype distribution to infer the unknown value of one SNP from the observed values of the other SNPs. Let H and G denote the diplotype (i.e., the pair of haplotypes on the two homologous chromosomes) and genotype, respectively. We write $H = (h, h')$ if the diplotype consists of h and h', in which case $G = h + h'$. Note that H cannot be determined with certainty on the basis of G if the two constituent haplotypes differ at more than one position.

We specify that the hazard function of T conditional on X and H satisfies the proportional hazards model

$$\lambda(t|X,H)=\lambda_0(t)e^{\theta^{\mathrm{T}}Z(X,H)},$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard function, $Z(X, H)$ is a vector-function of X and H, and the θ is a set of unknown regression parameters (Lin, 2004; Lin and Zeng, 2006). If we are interested in the additive genetic effect of a risk haplotype h^* and its interactions with X, then we may specify

$$Z(x,(h,h^{'}))=\left[\begin{array}{c} I(h{=}h^{*})+I(h^{'}{=}h^{*}) \\ x \\ \left\{I(h{=}h^{*})+I(h^{'}{=}h^{*})\right\}x \end{array}\right].$$

If we are interested in the additive effect of a particular SNP, then we replace $I(h = h^*) + I(h' = h^*)$ by the value of $(h + h')$ at that SNP position.

Because H is not directly observed, it is not possible to make statistical inference without constraining the joint distribution between the two constituent haplotypes in H. Let K be the total number of haplotypes in the population. For $k = 1, \ldots, K$, denote the kth haplotype by h_k. Define $\pi_{kl} = \Pr(H = (h_k, h_l))$ and $\pi_k = \Pr(h = h_k)$, $k, l = 1, \ldots, K$. Under Hardy-Weinberg equilibrium (HWE),

$$\pi_{kl}{=}\pi_k\pi_l, \quad k,l{=}1,\ldots,K.$$

It possible to allow parametric departures from HWE (Lin and Zeng, 2006), but HWE is expected to hold reasonably well in most human populations.

Assume that C is independent of T and H conditional on X and G. The likelihood takes the form

$$\begin{array}{l} \prod_{i=1}^{n}\sum_{H=(h_k,h_l)\in S(G_i)}\left\{\lambda_0(Y_i)e^{\theta^{\mathrm{T}}Z(X_i,(h_k,h_l))}\right\}^{\Delta_i} \\ \times\exp\left\{-\Lambda_0(Y_i)e^{\theta^{\mathrm{T}}Z(X_i,(h_k,h_l))}\right\}P\left(X_i|H{=}(h_k,h_l)\right)\pi_k\pi_l, \end{array} \qquad (3)$$

where $S(G)$ is the set of diplotypes that are compatible with genotype G. If X and H are independent, then $P(X_i|H)$ does not depend on H and can be dropped from the likelihood (Lin and Zeng, 2006). If X and H are not independent, we characterize their dependence through a generalized odds ratio function (Hu et al., 2010).

To maximize the nonparametric likelihood given in (3), we treat $\Lambda_0(\cdot)$ as a right-continuous function and replace $\lambda_0(Y)$ with the jump size of $\Lambda_0(\cdot)$ at Y. The maximization can be carried out through EM algorithms (Lin and Zeng, 2006; Hu et al., 2010). The resulting estimator of θ is consistent, asymptotically normal, and asymptotically efficient (Lin and Zeng, 2006; Hu et al., 2010). The limiting covariance matrix can be estimated by using the profile likelihood function (Murphy and van der Vaart, 2000).

When one of the SNPs in G is untyped, i.e., missing on all study subjects, the haplotype probabilities $(\pi_1, \ldots, \pi_K)$ cannot be estimated from the study data alone. Fortunately, external databases, such as the HapMap, can be used to estimate these probabilities. Suppose that the external sample consists of $\tilde{n}$ trios. For the jth trio, the genotype data consist of $G_j \equiv (GF_j, GM_j, GC_j)$, where GF_j, GM_j and GC_j denote the genotypes for the father, mother and child, respectively. Then the likelihood function for the external sample is

$$\prod_{j=1}^{\tilde{n}}\sum_{(h_k,h_l,h_{k'},h_{l'})\in S(G_j)}\pi_k\pi_l\pi_{k'}\pi_{l'},$$

where $(h_k, h_l, h_{k'}, h_{l'}) \in S(G_j)$ means that (h_k, h_l) is compatible with GF_j, $(h_{k'}, h_{l'})$ is compatible with GM_j, and $(h_k, h_{k'})$, $(h_k, h_{l'})$, $(h_l, h_{k'})$ or $(h_l, h_{l'})$ is compatible with GC_j. We multiply (3) by this likelihood and maximize the combined likelihood function. The resulting estimator of θ is consistent, asymptotically normal, and asymptotically efficient.

4. REMARKS

We have focused on the proportional hazards model. All the results described in this article can be extended to semiparametric linear transformation models (Lin and Zeng, 2006; Zeng et al., 2006; Hu et al., 2010). It would be more difficult to extend to the semiparametric accelerated failure time model (Kalbfleisch and Prentice, 2002, Ch. 7). We have assumed that X is time-invariant. It would be challenging to allow X to be time-varying in the likelihood approach.

There is a large body of statistical literature on case-cohort designs. Most of the estimators are based on inverse probability weighting. The interested readers are referred to Kalbfleisch and Prentice (2002, §11.4) and Zeng et al. (2006) for references.

In the genetics community, the prevailing approach to analyzing untyped SNPs is single imputation (Marchini et al., 2007). Because of strong correlations among neighboring SNPs, untyped SNPs can often be imputed with high accuracy and thus single imputation works quite well in practice. However, the SNPs with low frequencies cannot be imputed accurately. In general, single imputation does not yield valid tests of genetic association or consistent estimators of genetic effects.

Case-cohort and nested case-control designs were used in several GWAS studies. Due to the substantial drops in genotyping costs, many large cohorts have genotyped all cohort members on GWAS chips. There is now a growing interest in next-generation sequencing. Such technologies are much more expensive than GWAS chips. Thus, two-phase sampling provides a cost-effective strategy to conduct sequencing studies with age-of-onset phenotypes.

Extreme-trait sampling, which is in the same vein as two-phase cohort sampling, was recently adopted in sequencing studies on quantitative traits. In the NHLBI Exome Sequencing Project, subjects with the highest or lowest values of body mass index, low-density lipoprotein or blood pressures were selected from several large cohorts for whole-exome sequencing. In the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) resequencing project, subjects with the highest values of several quantitative traits were selected from three major cohorts along with a random sample. Likelihood functions similar to (1) can be constructed for such two-phase sampling.

Case-control sampling is also a form of two-phase sampling in that the sampling is based on the disease status. Prentice and Pyke (1979) showed that case-control sampling can be ignored in the logistic regression analysis of case-control data. This remarkable result, however, does not apply to two-phase sampling for survival time or continuous trait. Furthermore, standard statistical methods cannot be used to analyze data on secondary traits, i.e., the traits that are correlated with the one used for sampling.

Much of the recent genetic research has focused on association studies, which are considered the first step toward the ultimate goal of personalized medicine. As genetic research shifts to personalized medicine, survival analysis of genetic data will become more prominent since the goal of personalized medicine is to prevent disease and prolong life and the time to disease occurrence or death is often subject to complex censoring. Thus, we expect survival analysis of genetic data to be a very active research area in the coming years.

Acknowledgments

It is my distinct honor to contribute to this special issue of LIDA in honor of Ross Prentice, whose pioneering work has greatly advanced the field of biostatistics and paved the way for the successful careers of many biostatisticians, including myself. This work was supported by the National Institutes of Health.

References

Cox DR. Regression models and life-tables (with discussion). J R Statist Soc B. 1972; 34:187–220.

Fried LP, Borhani NO, Enright P, Furberg CD, Gardin JM, Kronmal RA, Kuller LH, Manolio TA, Mittelmark MB, Newman A, O'Leary D, Psaty B, Rautaharju P, Tracy R. The Cardiovascular Health Study: design and rationale. Annals of Epidemiology. 1991; 1:263–276. [PubMed: 1669507]

Hu YJ, Lin DY, Zeng D. A general framework for studying genetic effects and geneenvironment interactions with missing data. Biostatistics. 2010; 11:583–598. [PubMed: 20348396]

Johnson SR, Anderson GL, Barad DH, Stefanick ML. The Women's Health Initiative: rationale, design, and progress report. Journal of the British Menopause Society. 1999; 5:155–159.

Kalbfleisch, JD.; Prentice, RL. The Statistical Analysis of Failure Time Data. 2. Hoboken: Wiley; 2002.

Lin DY. Haplotype-based association analysis in cohort studies of unrelated individuals. Genetic Epidemiology. 2004; 26:255–264. [PubMed: 15095385]

Lin DY, Zeng D. Likelihood-based inference on haplotype effects in genetic association studies (with discussion). Journal of the American Statistical Association. 2006; 101:89–118.

Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nature Genetics. 2007; 39:906–913. [PubMed: 17572673]

Murphy SA, van der Vaart AW. On profile likelihood. Journal of the American Statistical Association. 2000; 95:449–465.

Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika. 1986; 73:1–11.

Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. Biometrika. 1979; 66:403–411.

The ARIC Investigators. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. American Journal of Epidemiology. 1989; 129:687–702. [PubMed: 2646917]

The International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001; 409:860–921. [PubMed: 11237011]

The International HapMap Consortium. A haplotype map of the human genome. Nature. 2005; 437:1299–1320. [PubMed: 16255080]

Thomas, DC. Addendum to Methods for cohort analysis: Appraisal by application to asbestos mining. In: Liddell, FDD.; McDonald, JC.; Thomas, PC., editors. J Roy Statist Soc Ser A. Vol. 140. 1977. p. 469-491.

Zeng D, Lin DY, Avery CL, North KE. Efficient semiparametric estimation of haplotype-disease associations in case-cohort and nested case-control studies. Biostatistics. 2006; 7:486–502. [PubMed: 16500923]