Modeling the effects of genetic factors on late-onset diseases in cohort studies

Mark E. Glickman David R. Gagnon

Abstract

Many late-onset diseases are caused by what appears to be a combination of a genetic predisposition to disease and environmental factors. The use of existing cohort studies provides an opportunity to infer genetic predisposition to disease on a representative sample of a study population, now that many such studies are gathering genetic information on the participants. One feature to using existing cohorts is that subjects may be censored due to death prior to genetic sampling, thereby adding a layer of complexity to the analysis. We develop a statistical framework to infer parameters of a latent variables model for disease onset. The latent variables model describes the role of genetic and modifiable risk factors on the onset ages of multiple diseases, and accounts for right-censoring of disease onset ages. The framework also allows for missing genetic information by inferring a subject's unknown genotype through appropriately incorporated covariate information. The model is applied to data gathered in the Framingham Heart Study for measuring the effect of different Apo-E genotypes on the occurrence of various cardiovascular disease events.

1 Introduction

Medical researchers are increasingly interested in the role of genetic factors in the manifestation of late-onset diseases. Late-onset diseases, such as Alzheimer disease and hypertension, are triggered by what appears to be a combination of a genetic predisposition to disease and environmental factors. Understanding the genetic and potentially modifiable environmental mechanisms has important implications for genetic and medical counseling, as well as further research activities. Research in establishing genetic contributions to the occurrence late-onset diseases has been on the rise. Among many such studies, recent ones that have

found a relationship between genetic factors and age of disease onset include Lautenschlager et al. (1996), Claus et al. (1990), Payami et al. (1997), and Blacker et al. (1997). It is now accepted that environmental factors are not the only determinants of disease onset; the inheritance of deleterious genes, at least in part, is causally responsible.

Existing prospective cohort studies offer opportunities to make inferences about the role of genetics in late-onset diseases. These studies allow for genetic information to be collected on a sample that is chosen to be representative of an appropriate study population. Incomplete information is usually not as pervasive as in other designs, such as retrospective family studies (see, for example, Thomson 1995, and references within). In these retrospective family studies, incident cases (called probands) of the disease are identified, and then retrospective information about the proband and the proband's family is collected. The affected status of members of the proband's family provide information on incidence and age at disease onset. In cohort studies, in contrast, the ability to follow individuals for long periods results in more incident cases and fewer censored cases. Furthermore, cohort studies typically eliminate certain forms of selection bias, such as ascertainment bias, that are present in retrospective family studies. The use of available databases often provides extensive information on risk factors and disease events which results in decreased study costs that allow for more informative data analyses. Statistical development of methods for determining disease onset age in cohort studies includes Cupples et al. (1989) and Cupples et al. (1991). Their work uses likelihood-based life table models to account for informative censoring. More recently, Gauderman and Thomas (1994) examined the genetic effects in proportional hazard models using Markov chain Monte Carlo posterior simulation. Iversen et al. (2000) developed a framework to model the effect of carrying mutations of particular genes on post-breast cancer survival using a latent variables approach, incorporating family

information. Frailty models, in which each family's overall risk can be modeled as a random effect, have also been explored in the context of cohort studies on disease onset (Petersen et al. 1996).

While cohort studies allow for representative samples from the population about which inferences are desired, a difficulty in analyzing the relationship between genetic information and disease onset in this context is that a substantial number of individuals may have died due to the disease prior to the time DNA sampling begins. Ignoring these individuals would normally lead, without proper adjustment, to biased inferences due to a survivor effect. Furthermore, health studies are often concerned with multiple manifestations of a disease process. For example, studies on cardiovascular diseases, including the Framingham Heart Study, may be concerned with different cardiovascular events (ischemia, myocardial infarction, etc.) that may be triggered by common physical causes. The possibility for multiple diseases and missing genetic information motivate the need for a more complete statistical framework to assess genetic factors for disease onset in prospective cohort studies.

This article develops a statistical framework for inferring genetic and environmental effects on disease onset in cohort studies when DNA sampling may have occurred substantially after study inception. Our framework multivariately models disease onset ages conditional on genetic and environmental covariates. Our framework also accounts for missing genetic information that may have arisen from subjects dying prior to DNA sampling. Furthermore, our framework can be applied to studies involving multiple diseases in which the disease processes are suspected to have a common physiological etiology. We discuss the model development and approach to inference in Section 2. We explore sensitivity to large amounts of missingness in the data in Section 3 based on the analysis of simulated data. In Section 4, we apply our approach using data from the Framingham Heart Study to measure the effect

of Apo-E genotypes on the development of certain cardiovascular disease (CVD) events. We discuss the limitations and extensions of our framework in Section 5.

2 Multivariate statistical model for disease onset

Consider a cohort of n subjects who are free of K genetically-related and irreversible diseases under investigation. We assume that covariate information is collected at baseline, but that genetic information may be obtained a substantial time beyond the study inception for subjects remaining in the cohort. For subject i, let $A_{i1}, \ldots, A_{iK}$ be the onset ages of the K diseases. Our framework assumes that all subjects would eventually contract all diseases, though censoring due to death or study termination prevents some or all of these events from being observed. In contrast to cure rate models considered by Berkson and Gage (1952) and, more recently, Chen et al. (1999), our assumption may be reasonable if it is believed that other factors besides the genes and measured covariates contribute to disease onset. Let $\boldsymbol{W}_i = (W_{i1}, \ldots, W_{iM})$ be a vector of M binary indicators for the genotype of subject i. A typical application would involve one of several genes or their mutations at a particular gene locus, or combinations of genes at several loci. In the latter case, only one of the W_{im} would be 1, and the rest 0 for subject i. Our framework also allows the W_{im} to represent binary indicators of gene presence or absence at M different loci, if interest centers on the effect of several particular genes in combination. Furthermore, let $\boldsymbol{X}_i = (X_{i1}, X_{i2})$ denote a vector of covariates measured at baseline, partitioned into two sets. One set, X_{i1}, consists of environmental factors, comorbidities, and other health factors that are believed to be causally unrelated to the genetic factors considered in the study. The second set, X_{i2}, consists of health factors that may be causally related to the genetic factors. It is assumed that prior information exists that allows the investigator to separate covariates into these

two categories. Because the goal of our framework is to model the direct effects of genes, it is important to treat these "causal" covariates separately in our model to avoid potential confounding.

As is common with parametric survival models, we consider a Weibull distribution for disease onset age given by

$$A_{ik} \mid \boldsymbol{X}_i, \boldsymbol{W}_i, \beta_k, \lambda, \gamma \sim \text{Weibull}(\mu_{ik}, \lambda) \tag{1}$$

with density

$$f(A_{ik} \mid \mu_{ik}, \lambda) = \frac{\lambda}{\mu_{ik}} A_{ik}^{\lambda-1} \exp(-A_{ik}^{\lambda}/\mu_{ik}).$$

where

$$\log(\mu_{ik}) = (X_1)_i' \beta_k + \boldsymbol{W}_i' \gamma \tag{2}$$

Our onset age distribution depends on the disease and subject only through Weibull scale parameter μ_{ik}. The parameter μ_{ik} is composed of the additive effects of covariates X_{i1} through a linear term with β_k, and of genetic factors $\boldsymbol{W}_i$ through a linear term depending on parameters γ.

We further model the covariates that are possibly causally related to the genetic factors, X_{i2}, as a function of the $\boldsymbol{W}_i$ through the general location model of Olkin and Tate (1961). This model assumes that the X_{i2} are partitioned into continuous and categorical covariates. The categorical covariates are modeled multinomially (usually with log-linear constraints), as in a contingency table. Conditional on the categorical covariates and parameters, the continuous covariates appropriately transformed are assumed to follow a multivariate normal distribution. A more complete discussion of general location models, including inference in the presence of missing data, can be found in Schafer (1997). This model is sufficiently flexible to be able to describe a wide class of covariate collections without introducing undesired

computational complexity.

Because the $\boldsymbol{W}_i = (W_{i1}, \ldots, W_{iM})$ are missing for some i, our model specification is completed by assuming a model for the $\boldsymbol{W}_i$. In the most general situation where the W_{im} are indicators of different genotypes, we assume

$$\boldsymbol{W}_i \mid \boldsymbol{\pi} \sim \text{Multinomial}(1, \boldsymbol{\pi}) \tag{3}$$

for M-vector parameter $\boldsymbol{\pi}$. Depending on the nature of the problem, constraints can be imposed on the multinomial model. For example, log-linear constraints may be appropriate if modeling genotypes which are the composition of genes at multiple loci. If the $\boldsymbol{W}_i$ are indicators of particular genes of interest at M loci, then we may assume, independently,

$$W_{im} \mid \pi_m \sim \text{Bernoulli}(\pi_m) \tag{4}$$

for $m = 1, \ldots, M$.

When disease onset is not observed for all K diseases, we assume for our modeling framework that censoring was noninformative (see, for example, Kalbfleisch and Prentice, 1980). If censoring is due to dropout or study termination, then clearly the potentially observable disease onset is not affected. If health-related death is the cause of censoring, then one might contemplate that, given a subject's health at the time of death, onset age for the censored diseases would have likely occurred soon after censoring. However, as noted by Lagakos (1979), typical structural assumptions about informative censoring mechanism could be incorporated but that they would be untestable without further assumptions. In our framework, informative censoring assumptions are not only untestable, but rely on making modeling assumptions of disease onset after death. It should be noted that in some applications noninformative censoring is not a sensible assumption, and our framework would therefore not be appropriate.

The specific choice of a prior distribution must be made on a case-by-case basis. It is sensible, for example, to choose a prior that factors into independent densities of each model parameter. A natural choice for the contribution to the prior by the π_m is a non-informative Dirichlet density, if the $\boldsymbol{W}_i$ are assumed to come from an unconstrained multinomial distribution. Alternatively, if information about the mode of gene inheritance is known, possibly through genetic information obtained on family members, this can be incorporated into modeling the π_m.

The framework described in (1)–(3) jointly models disease onset ages and genetic factors accounting for the possibility that a non-trivial fraction of subjects are censored prior to genetic testing. A feature of our framework is that each subject has simultaneously any number between 0 through K disease onset ages $A_{i1}, \ldots, A_{iK}$ observed, though any or all may be censored. Our framework can be viewed as a type of model for multiple survival outcomes, but with potentially a large amount of missing information.

To aid in causal interpretation of our model, a conditional independence assumption is made. In particular, the models for age of disease onset only depend on genetic factors, $\boldsymbol{W}$, and covariates that do not relate to genetic factors, X_1. Conditional on genetic factors, the onset ages do not depend on health factors, X_2, that are believed to be causally related to the genes. Because our model is constructed to measure genetic effects in an observational framework, the only allowable covariates that may be incorporated are ones that are unrelated to the "treatment" (genotype). If X_2 were included in the onset age model along with $\boldsymbol{W}$, then the likely result is that the estimated effect of genotype would be inappropriately lessened in magnitude as some of the genotypic effects would be absorbed into the estimated effect of X_2. The covariates, X_1, may be viewed as risk adjusters that do not interfere with causal inferences but are incorporated to increase precision.

The model for X_2 conditional on $\boldsymbol{W}$ is included in our framework to improve the accuracy of inferring the multinomial probabilities of $\boldsymbol{W}$. Our framework recognizes that a substantial number of subjects may have missing values for $\boldsymbol{W}$, so that our model allows the missing $\boldsymbol{W}$ to be inferred via Bayes rule conditional on the X_2. This in turn improves inferences for the onset age parameters because the missing $\boldsymbol{W}$ are more precisely described.

In modeling the effect of the genetic and environmental factors, we assume that the health/environmental parameters, β_k, are random effects from a common distribution. This aspect of the framework allows the parameters to differ for each disease, but the random effects assumption recognizes the potential similarity of the effects so that information across diseases is pooled. If the onset of different diseases in a disease cluster stem from similar environmental and genetic etiology, then it is not unreasonable to assume that their effects on the disease onset are similar. The pooling of information by assuming a random effects distribution also can increase precision of the inferences. In contrast, we assume that genetic effects are common to the different onset age models. By assuming identical contribution of the genetic factors to the different onset age distributions, we recognize that the effects of genes are common to diseases with similar disease etiology. While it is not necessary in modeling the effects of genotypes to assume a common form (the genotypic effects can be modeled as separate effects), precision in inferences can be gained by assuming identical effects. The extra precision can be important when a substantial fraction of a cohort has missing genetic information.

The approach to modeling identical genetic effects for related diseases is similar to problems that involve summarizing health effects from many variables to produce a single overall summary health score, which can then be used in subsequent analyses. Two such commonly used scores are the APACHE-II score (Knaus et al. 1985) and the Medicare Mortality

Prediction Study score (Daley et al. 1988). These types of severity scores have been used successfully in health outcome models, such as in Normand et al. (1996). Our approach, in contrast, allows all effects to be inferred simultaneously. Because much of our interest is on the effects of the genetic factors, the contribution of $(X_1)_i'\beta_k$ to the onset age models may be understood as a covariance adjustment.

Instead of modeling onset age distributions as Weibull, a wide variety of models could be considered, such as log-normal and Gamma if interest lies in parametric modeling, or semi-parametric models such as Cox's proportional hazards model. A recent review of semi-parametric survival modeling techniques, appropriate for use in our framework, can be found in Sinha (1997). We argue for the use of parametric modeling in our context because disease onset ages are of crucial interest, and that potentially a large amount of missing data often requires stronger modeling assumptions. The effectiveness of parametric hazards is argued by Efron (1988), and more recently by Gelfand et al. (2000), for such situations.

Inference for our model can be performed using Markov chain Monte Carlo (MCMC) methods, such as the Gibbs sampler (e.g., Gelfand and Smith 1990). The algorithm can be implemented by alternately sampling from the following three conditional posterior distributions: the distribution of the missing A_{ik} given model parameters, observed data, and complete genetic information; the distribution of the missing genetic information $\boldsymbol{W}_i$ given model parameters and disease onset ages; and the distribution of the model parameters given complete genetic information and disease onset ages.

Sampling from the conditional posterior distribution of missing onset ages is straightforward. Let Q_i be the (observed) censoring age. Then the conditional posterior distribution of missing A_{ik} follows (1) over the range $(Q_i,\ \infty)$, appropriately renormalized. With our

Weibull model, values may be simulated using standard inverse-cdf methods.

Sampling from the conditional posterior distribution of the missing genetic factors, $\boldsymbol{W}_i$, for subject i is also a straightforward procedure. The conditional posterior mass function for W_{im} given observed data and the current draws of other parameters and missing data can be obtained through Bayes rule by evaluating the unnormalized posterior density, and then normalizing by their sum over the M values. Similar conditional posterior sampling can be carried out when the $\boldsymbol{W}_i$ are assumed, for example, product binomial.

The conditional posterior distribution of model parameters can be accomplished following Dellaportas and Smith (1993), who discuss fitting survival models with Weibull hazards. The problem fits naturally into their framework because the problem of simulating model parameters within an iteration of Gibbs sampling has been reduced to simulating parameters from a survival model with no censoring.

3 Sensitivity analysis

Our framework can be demonstrated on simulated data in order to examine the sensitivity to the large amount of missing onset age and genotypic information in typical applications. Suppose interest lies in measuring the effect of different genotypes at two different loci on the onset of four related diseases. For simplicity, assume that at each locus one of two genotypes can be observed, independently. A simulated sample of 1000 subjects were generated as follows. Let W_{i1} and W_{i2}, for $i = 1, \ldots, 1000$, each be 0 or 1, depending on the genotype at the locus. A value of 0 indicates the presence of a mutation that is likely to lower disease onset age, while a value of 1 is protective. We generated $W_{i1} = 1$ and $W_{i2} = 1$ with probabilities 0.3 and 0.2, respectively. Three non-causal covariates (X_{i1}, X_{i2}, X_{i3}) and one

causal covariate (X_{i4}) were also simulated. The first two covariates (X_{i1}, X_{i2}) were binary (0 or 1), with $X_{i1} = 1$ and $X_{i2} = 1$ generated with probabilities 0.3 and 0.45 respectively. The third covariate, X_{i3}, was simulated from a normal distribution with mean 250 and standard deviation 35. The fourth covariate, X_{i4}, was simulated from a normal distribution with variance 10, and with a mean, $\mu_{X_{i4}}$, that depended on the genotypes,

$$\mu_{X_{i4}} = \gamma_0 + \gamma_1 W_{i1} + \gamma_2 W_{i2}$$

where $\gamma_0 = 20$, $\gamma_1 = 10$, and $\gamma_2 = 5$.

Disease onset ages for the four diseases, A_{ij}, for $j = 1, \ldots, 4$, were generated from a Weibull model where

$$A_{ik} \sim \text{Weibull}(\mu_{ik}, \lambda)$$

with $\lambda = 1.5$ and

$$\log \mu_{ik} = \beta_{k0} + \beta_{k1} X_{i1} + \beta_{k2} X_{i2} + \beta_{k3} X_{i3} + \beta_{k4} W_{i1} + \beta_{k5} W_{i2}.$$

The generating values of the β_{kj} are displayed in Table 1.

Disease censoring ages (due to death or study termination) were simulated from a N(70, 5^2) distribution. Independently, ages at which DNA samples were obtained were simulated from a N(76, 5^2) distribution. When the minimum censoring age was less than the simulated DNA sampling age, the DNA sampling was assumed not to have occurred, and the genetic information was set to be missing. This resulted in only 187 of the 1000 original simulated subjects to have observed DNA information.

Two analyses were performed. The first analysis was based on only using the 187 of the 1000 subjects that had observed genetic information. The second analysis is the same as the first, except incorporating the partial information (incomplete genetic information,

censored disease onset ages) for all 1000 simulated subjects. This is the situation that our modeling framework demonstrates potential benefits. Both models were fit using MCMC simulation from the posterior distribution, using a burn-in of 5,000 iterations and summarizing inferences based on every 20th iteration of a subsequent 10,000 iterations. Table 1 presents the results of the simulations. The table shows that both analyses produce similar point estimates for the βs and the γs. The posterior intervals for the reduced sample analysis are uniformly wider, reflecting a smaller overall sample. Specifically, the posterior standard deviations of the β_{kj} for the reduced-sample analysis were on average 2.11 times as large as the corresponding posterior standard deviations in the full-sample analysis. The modeling framework for the full sample analysis takes advantage of information contained in X_4 in making inferences about the missing W_1 and W_2, so that the estimates of the onset age parameters are more accurate than in the reduced sample analysis.

It is also worthwhile to note that among the sample of 1000 subjects, 30.3% were generated to have $W_{i1} = 1$ and 19.0% to have $W_{i2} = 1$. Among the 187 subjects who were not censored prior to randomly generated DNA sampling ages, the proportion with $W_{i1} = 1$ was 32.1%, and with $W_{i2} = 1$ was 27.9%. These percentages, particularly for W_{i2}, are larger than those of the original 1000 because the subjects with either $W_{i1} = 0$ or $W_{i2} = 0$ tended to die before DNA testing, so proportionately fewer of such individuals remained in the smaller sample. The modeling framework that allows for censored genotypes successfully estimated the generated proportions; the estimated proportion (that is, the posterior mean) with $W_{i1} = 1$ was 28.3%, and with $W_{i2} = 1$ was 21.0%, with 95% posterior intervals covering the generating values of 0.3 and 0.2, respectively.

		Reduced Sample Analysis		Full Sample Analysis	
Parameter	Generating Value	Posterior Mean	Posterior 95% Interval	Posterior Mean	Posterior 95% Interval
β_{11}	0.05	–0.145	(–0.547, 0.207)	0.0936	(–0.0716, 0.251)
β_{21}	–0.10	0.057	(–0.435, 0.528)	0.0268	(–0.161, 0.230)
β_{31}	–0.15	–0.283	(–0.714, 0.131)	–0.249	(–0.424, –0.0559)
β_{41}	–0.2	–0.388	(–0.915, 0.113)	–0.392	(–0.623, –0.189)
β_{12}	–0.05	–0.135	(–0.521, 0.215)	–0.120	(–0.268, 0.0303)
β_{22}	0.05	0.121	(–0.231, 0.536)	0.174	(–0.0117, 0.337)
β_{32}	0.00	0.137	(–0.240, 0.489)	–0.0724	(–0.235, 0.0993)
β_{42}	0.03	0.023	(–0.369, 0.367)	–0.0210	(–0.182, 0.161)
β_{13}	0.0005	0.00403	(–0.000895, 0.00907)	0.00134	(–0.000625, 0.003352)
β_{23}	0.0007	0.00194	(–0.00431, 0.00846)	0.000323	(–0.00239, 0.00299)
β_{33}	0.0007	0.00306	(–0.00140, 0.00796)	0.000112	(–0.00216, 0.00240)
β_{43}	0.0005	0.00432	(–0.000535, 0.00973)	0.00195	(–0.000550, 0.00439)
β_{14}	0.3	0.496	(0.109, 0.835)	0.542	(0.351, 0.733)
β_{24}	0.5	1.042	(0.635, 1.475)	1.057	(0.856, 1.248)
β_{34}	0.6	0.871	(0.502, 1.276)	0.950	(0.743, 1.137)
β_{44}	0.7	1.065	(0.660, 1.469)	1.068	(0.852, 1.259)
β_{15}	–0.03	–0.189	(–0.595, 0.227)	–0.370	(–0.683, –0.0146)
β_{25}	0.03	–0.0980	(–0.625, 0.385)	–0.0763	(–0.458, 0.267)
β_{35}	0.05	0.519	(0.0870, 0.926)	0.402	(0.110, 0.711)
β_{45}	–0.01	0.324	(–0.134, 0.791)	0.224	(–0.148, 0.572)
γ_0	20	20.01	(19.35, 20.64)	20.09	(19.76, 20.42)
γ_1	10	9.47	(8.22, 10.61)	9.70	(9.16, 10.26)
γ_2	5	5.10	(3.86, 6.23)	5.12	(4.24, 5.96)
λ	1.5	1.54	(1.42, 1.65)	1.53	(1.47, 1.59)

Table 1: Estimates of model parameters from simulation analyses. The "Reduced Sample" analysis was based on using only the 187 simulated subjects for whom genetic information was observed. The "Full Sample" analysis was based on using available information on all 1000 subjects. Summaries were computed from 500 simulated values from the posterior distribution using MCMC simulation.

4 Effect of Apo-E genotype on the Onset of Cardiovascular Disease

We demonstrate the application of the modeling framework in Section 2 to data obtained from the Framingham Heart Study. Interest of our analysis centers on measuring the frequency of the common Apolipoprotein-E (Apo-E) alleles in the study population, and on the effects of different Apo-E allelic combinations on the onset of various CVD events stratified by gender. In particular, we examine the genetic effects on the age of first occurrence of angina pectoris, recognized acute myocardial infarction, unrecognized ("silent") acute myocardial infarction, and congestive heart failure. Unrecognized myocardial infarction is usually identified upon EKG reading during a physical, and in the Framingham Heart Study such physical exams were conducted at most every two years. Identification of the risk-adjusted effects of different Apo-E genotypes has important practical implications for genetic counseling and preventive medicine. Also, understanding the impact of the Apo-E genotypes on the onset of CVD events allows for further understanding and exploration of disease etiology.

The Framingham Heart Study, which began in 1948 and initially funded by the National Heart Institute, is one of the largest ongoing studies conducted to learn about the causes of heart disease and stroke. Details of the design and methods of the study can be found in Dawber et al. (1951). The study recruited 5209 men and women between the ages of 28 and 62 from Framingham, Massachusetts, almost all Caucasian, gathering extensive information every two years from physical exams and interviews. Diagnosis of CVD events were based on clinical information obtained during study visits, records obtained from personal physicals, and hospitalizations (Dawber et al. 1951). To increase certainty of diagnosis, a panel of three physicians reviewed all suspected CVD events to ascertain their occurrence.

Disease	Percent Missing	Average Age
Angina Pectoris	81.4	65.06
Unrecognized MI	93.1	70.18
Recognized MI	83.5	69.49
Congestive Heart Failure	83.8	76.04

Table 2: Descriptive summaries of disease onset ages. For each disease, the percentage of missing (censored) onset ages and the mean onset age among the observed ages are reported.

The cohort we examine consists of 4804 subjects who were free of CVD symptoms at study entry. For each subject, we collected information on the ages of onset of the four mentioned CVD events, and the censoring age if any of the diseases did not occur. Table 2 displays observed information about the four diseases in the study. The frequency of missingness of the disease onset ages is high in the cohort. For all four diseases, over 80% of the onset ages are missing, and therefore treated as censored. The average censoring age for this cohort was 76.55.

Table 3 displays health covariates measured at baseline that we have incorporated into our analysis. We divide the variables into ones that can be viewed as being causally related to the Apo-E gene (serum cholesterol levels, phospholipid levels), and ones that are assumed causally unrelated to the Apo-E genes. Because the Apo-E genes are responsible for cholesterol transport in the blood, we would anticipate a direct relationship between gene presence and average cholesterol levels. Health covariates were non-missing for all subjects and reflect information available at the start of the study.

Blood samples for DNA in the Framingham Heart Study were obtained between 1987 and 1991. Apo-E genotyping was performed as described in Hixson and Vernier (1990). The Apo-E gene, a gene of current interest in the study of cardiovascular diseases (Eichner et al. 1993, Wilson et al. 1994, Schachter et al. 1994, Stengard et al. 1995), produces proteins that

Covariates not causally related to to Apo-E alleles	Covariates possibly causally related to Apo-E alleles
Cigarettes per day	Serum Cholesterol level (mg/dl)
Forced vital capacity (cl/s)	Phospholipid level (mg/dl)
Serum glucose level (mg/dl)	
Systolic blood pressure (mm Hg)	
Hemoglobin level (mg/ml)	
Body mass index (kg/m^2)	

Table 3: List of health-related covariates. The first column contains covariates that are assumed causally unrelated to the Apo-E alleles, and the second column contains covariates in our model that are assumed casually related to the Apo-E alleles.

Genotype	Number of Males	Number of Females	Total Number	Percent of Non-missing
e2/e2	1	4	5	0.39
e2/e3	51	85	136	10.74
e2/e4	8	14	22	1.74
e3/e3	320	526	846	67.82
e3/e4	85	153	238	18.80
e4/e4	10	9	19	1.50

Table 4: Frequency distribution of genotypes, stratified by gender, in sample of 1266 individuals with measured genetic information.

are believed to help metabolize certain plasma lipoproteins in the circulation. The three most common allelic forms of the Apo-E gene are referred to as e2, e3 and e4, which code for proteins Apo-E2, Apo-E3 and Apo-E4. Consequently, we consider six possible genotypes formed by pairs of alleles: e2/e2, e2/e3, e2/e4, e3/e3, e3/e4, and e4/e4. Carriers of an e4 allele are understood to be at higher risk for Alzheimer's disease than individuals without (e.g., Saunders et al. 1993), though no conclusive evidence exists about the impact on CVD.

Of the 4804 subjects in our analysis, only 1266 (26%) had information recorded about their Apo-E genotype. Table 4 shows the genotype distribution for individuals with non-missing genetic information. The e3/e3 genotype is, by far, the most common among the six. The e2 and e4 alleles are much less frequent than e3.

Before constructing our model for onset ages of the four CVD events, we examined scatter plots which suggested transformations of several covariates. In particular, serum glucose level was incorporated into the model on the log scale, including a quadratic term (again on the log scale). A quadratic term was included for systolic blood pressure, and a quadratic term was also included for hemoglobin level. The remaining covariates in the model were included linearly. Both cholesterol and phospholipid levels remained untransformed.

For $k = 1, \ldots, 4$ cardiovascular disease events, we model the onset ages as

$$A_{ik} \sim \text{Weibull}(\mu_{ik}, \lambda) \tag{5}$$

where

$$\log(\mu_{ik}) = (X_1)_i' \beta_k + \gamma_{W_i, Z_i}. \tag{6}$$

For subject i, $(X_1)_i$ is the vector of the non-causal health factors (including a constant for an intercept term), β_k is a parameter vector specific to disease k, and γ_{W_i,Z_i} is the effect of genotype W_i and gender Z_i. The parameter vector γ therefore only takes on $6 \times 2 = 12$ values. To avoid parameter aliasing, we impose the constraint that $\gamma_{1,1} = 0$. Alternative linear constraints could be imposed as well, such as setting the sum of the $\gamma_{w,g}$ to be 0. In the current parameterization, the $\gamma_{w,g}$ can be interpreted as the the gender by Apo-E effect relative to males with the e2/e2 gene.

We assume that $(X_2)_{i1}$, the cholesterol level for subject i at baseline, and $(X_2)_{i2}$, the phospholipid level for subject i at baseline, can be modeled as a function of genotype,

$$(X_2)_{ij} \sim \text{N}(\boldsymbol{W}_i \alpha_{Xj},\ \sigma^2_{Xj}) \tag{7}$$

for $j = 1, 2$, where α_{Xj} is a vector of genotype effects on causal factor j, and similarly σ^2_{Xj} is the corresponding variance. Here, $\boldsymbol{W}_i$ is a vector of six indicators for Apo-E genotype. This

component of the model increases the precision of inferences about the missing $\boldsymbol{W}_i$ through their relationship to the cholesterol and phospholipid levels at baseline. Because the utility of this component of the model is used to improve inferences about the missing $\boldsymbol{W}_i$ as a simple linear model on the mean, rather than make inferences about the parameters of this model component, it is not necessary to model the correlation structure of X_2.

We model $\boldsymbol{W}_i$, the genotype of subject i, through a multinomial model

$$\boldsymbol{W}_i \mid \boldsymbol{\pi} \sim \text{Multinomial}(1, \boldsymbol{\pi}) \tag{8}$$

where $\boldsymbol{\pi}$ is a vector of length 6 indicating the probabilities of the six genotypes in the study population, with π_1 corresponding to e2/e2, π_2 corresponding to e2/e3, π_3 corresponding to e2/e4, π_4 corresponding to e3/e3, π_5 corresponding to e3/e4, and π_6 corresponding to e4/e4. We further assume that the genes obey Mendelian inheritance (each allele is inherited independently from each parent), and that the distribution of alleles in the population is in Hardy-Weinberg equilibrium (which is crudely supported by the distribution of the observed data). We assume no effects due to population admixtures. Letting p_2, p_3 and p_4 represent the allele frequencies in the study population of Apo-E e2, e3 and e4, respectively, these assumptions can be modeled as $\pi_1 = p_2^2$, $\pi_2 = 2p_2p_3$, $\pi_3 = 2p_2p_4$, $\pi_4 = p_3^2$, $\pi_5 = 2p_3p_4$, and $\pi_6 = p_4^2$.

To reflect our initial uncertainty, a non-informative proper prior factoring into locally non-informative independent densities was assumed for all model parameters. The linear parameters in the onset age models had normal densities with mean 0 and variance 1000, the variance parameters were assumed to have Gamma distributions with mean 1 and variance 10, the cholesterol and phospholipid parameters in (7) were assumed normal with mean 200 and variance 10000, and the parameters (p_2, p_3, p_4) have a Dirichlet distribution with

parameters $(0.05, 0.85, 0.10)$. The weakly informative Dirichlet parameters were chosen to be consistent with previous findings on Apo-E allele frequencies, while allowing the likelihood to dominate inferences. It should be noted, however, that it is not necessary to incorporate such information into the prior distribution; for example, assuming equal values for the p_j would be appropriate if no additional information were available. Estimates of these allele frequencies, which are known to vary in different populations, can be found, for example, in Hallman et al. (1991) and Louhija et al. (1994).

A single series Gibbs sampler with over-dispersed starting values was run for a burn-in period of 5,000 iterations. The sampler was continued another 4,000 iterations, subsampling all missing data and parameters every 8 iterations. The subsampling was performed both to reduce autocorrelation in the successive Gibbs draws, and to conserve disk space. The sampler was implemented in BUGS (Spiegelhalter et al. 1996). Convergence of the Gibbs sampler was assessed from these 4,000 simulated draws through trace plots, and simple diagnostics that compare parameter sample means from early parts of the series to later parts (Geweke 1992). Posterior summaries were computed from the 500 subsampled simulated values.

The results of the model fit indicate that the covariates unrelated to genes accounted for significant variability in onset ages. Marginal central posterior intervals of the components of the β_k in (6) reveal that smoking history, systolic blood pressure, and body mass index were uniformly important risk adjustment variables for onset age of all four disease models. Other covariates, such as forced vital capacity and serum glucose level, were important for some disease onsets but not others. These results were determined from 95% central Monte Carlo posterior intervals.

Genotype	95% Central Posterior Interval
e2/e2	(0.0039, 0.0052)
e2/e3	(0.0965, 0.1106)
e2/e4	(0.0204, 0.0244)
e3/e3	(0.5731, 0.6037)
e3/e4	(0.2424, 0.2647)
e4/e4	(0.0245, 0.0304)

Table 5: 95% central posterior intervals of, $\boldsymbol{\pi}$, the genotype frequencies.

Posterior summaries also revealed that the six genotypes were associated with differing levels of cholesterol and of phospholipid levels. Figure 1 shows the distribution of α_{X_1} and α_{X_2} in (7). The posterior simulated values shows clearly differing distributions, and that subjects with genotype e2/e4 are most strongly associated with increased cholesterol and phospholipid levels. The mean cholesterol and phospholipid levels for subjects with the e2/e2 genotype have large posterior variability, which is a reflection of the small number of such subjects in the data set. The information in Figure 1 suggests that subjects with missing genetic information will, with large probability, be inferred to have the e3/e3 genotype if their cholesterol and phospholipid levels are high (not e2/e4 because the frequency of this genotype is low), and e2/e3 if cholesterol and phospholipid levels are low.

We also calculated estimated marginal posterior intervals for the three allele frequencies. The Monte Carlo 95% central posterior intervals for p_2, p_3, and p_4, respectively, are (0.0626, 0.0724), (0.7571, 0.7770), and (0.1565, 0.1744). The corresponding 95% central posterior intervals for genotype frequencies, $\boldsymbol{\pi}$, are given in Table 5. In comparison to Table 4, it is worth noting that a substantially smaller proportion of the subjects with missing genetic information are inferred to have the "normal" genotype e3/e3 than the observed 67.8%. Similarly, a great proportion of subjects with missing genetic information are inferred to have the e2/e4, e3/e4 and e4/e4 genotypes. This result may reflect the model's

ability to detect that the Apo-E e4 allele is related to early censoring.

The genotypic effects, represented by the marginal posterior distribution of the $\gamma_{w,g}$, are shown in Figure 2 as boxplots of MCMC samples stratified by gender. The parameters were transformed to sum to 0. Larger values of $\gamma_{w,g}$ correspond to later disease onset. The analysis shows that, for both males and females, the e2/e4 genotype is associated with earlier onset of cardiovascular disease. On average, men appear to experience earlier onset of cardiovascular diseases. To estimate the difference of the e2/e4 and e3/e3 genotypes on survival, we first note that the median onset age of disease k for subject i is $M_{ik} = (\mu_{ik} \ln 2)^{\lambda}$. For a subject with average covariates, the posterior mean of M_{i1}, the median onset age for angina, is computed to be 95.82 when the subject has genotype e3/e3, and is computed to be 73.43 when the subject has genotype e2/e4. The large median onset age for angina, particularly for "normal" e3/e3 subjects, is consistent with the low frequency of its occurrence in the Framingham population, as most people with the e3/e3 genotype would die before the potential onset of angina. The posterior mean of the ratio of medians is a 29% lower median for e2/e4 genotypes. Men appear at higher risk of early cardiovascular disease than men if they have the e2/e2 genotype, though the evidence suggested by the model fit is not strong due to the large variability in parameter inferences for the effect of the e2/e2 genotype. In fact, the effect of the e2/e2 genotype for males has very large variability because the sample contains only one male with the e2/e2 genotype who had no measured CVD events, and who died at age 80 for non-CVD related health causes.

5 Discussion

The framework presented in this paper offers a flexible approach to modeling the genetic impact on multivariate survival outcomes when genetic information is missing due to death prior to DNA sampling. Our approach allows for a substantial fraction of missing genetic information, as the genotyping is inferred from "causal" covariates measured at baseline. These covariates are used to reduce the bias due to dropouts that have been censored for reasons related to the disease process under investigation. Because we fit the models within a Bayesian framework, missing onset ages and genotypes can be addressed in a straightforward manner.

An aspect of our framework, which can be viewed as both a strength and a limitation, is the use of common genotypic effects on the collection of onset age distributions. This assumption is only appropriate if the diseases under study are believed to have a common underlying etiology. For measuring the onset of cardiovascular diseases, this assumption seems appropriate, as different cardiovascular events arguably have similar causes. The assumption of common effects might not be appropriate, for example, for different forms of cancer, or for multiple unrelated disease events. In cases where a moderate to large proportion of subjects have missing DNA information, assuming a common effect across diseases can also increase precision of inferences.

Our approach to inferring missing genetic information can be used to great advantage if covariates possibly causally linked to the genotypes are judiciously chosen and included in the model. For the cardiovascular disease example, the inclusion of baseline cholesterol and phospholipid levels allow greater probability of recognizing that subjects with missing genetic information are associated with the rarer genotypes. This occurs through the re-

lationship of genotypes among subjects with observed genetic information with cholesterol and phospholipid levels. It could be argued that, because cholesterol and phospholipid levels vary over time, other time-adjusted summaries of this information could be incorporated into the model. The inclusion of such time-varying factors is, however, outside the scope of this work, as measurements obtained beyond baseline could be confounded with a subject's disease status. Our framework for causal covariates does not directly address this situation.

In principle, our model could include a subject-specific frailty component by including a random effect per subject in (1). This approach was applied to model for recurrences of kidney infection in McGilchrist and Aisbett (1991). The difficulty with the inclusion of subject-specific frailties in our model is that many subjects have all censored onset ages for each disease, and missing genetic information, so that the frailty of such subjects is aliased with onset ages. The additional external structure required for identifiability is beyond the scope of this work.

The framework in this paper is flexible enough to allow a variety of extensions. For example, the model can be extended to include a frailty component to account for family effects, or for the inclusion of pedigree information. Extensions such as these can improve the precision of inferences on the genetic effects on disease occurrence, and can result in a greater understanding of disease etiology and, ultimately, disease treatment and prevention.

Bibliography

J. Berkson and R.P. Gage, "Survival curve for cancer patients following treatment," *J. Amer. Stat. Assoc.*, 47, pp. 501–515, 1952.

D. Blacker, J.L. Haines, L. Rodes, H. Terwedow, R.C. Go, et al., "ApoE-4 and age at onset of Alzheimer's disease: the NIMH genetics initiative," *Neurology*, 48, pp. 139–147, 1997.

M.H. Chen, J.G. Ibrahim, and D. Sinha, "A new Bayesian model for survival data with a surviving fraction," *J. Amer. Stat. Assoc.*, 94, pp. 909–919, 1999.

E.B. Claus, N.J. Risch, and W.D. Thompson, "Using age of onset to distinguish between subforms of breast cancer," *Ann Hum. Genet.*, 54, pp. 169–177, 1990.

L.A. Cupples, N.C. Terrin, R.H. Myers, and R.B. D'Agostino, "Using Survival Methods to Estimate Age-at-Onset distributions for Genetic Diseases with an Application to Huntington Disease," *Genetic Epidemiology*, 6, pp. 361–371, 1989.

L.A. Cupples, N. Risch, L.A. Farrer, and R.H. Myers, "Estimation of Morbid Risk and Age of Onset with Missing Information," *American Journal of Human Genetics*, 49, pp. 76–87, 1991.

J. Daley, S. Jencks, D. Draper, G. Lenhart, N. Thomas, and J. Walker, "Predicting hospital-associated mortality for Medicare patients: a method for patients with stroke, pneumonia, acute myocardial infarction, and congestive heart failure," *J. Amer. Med. Assoc.*, 260, pp. 3617–3624, 1988.

T.R. Dawber, G.F. Meadors, and R. Moore, "Epidemiological approaches to heart disease: The Framingham Heart Study," *Am. J Public Health*, 41, pp. 279–286, 1951.

P. Dellaportas, and A.F.M. Smith, "Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling," *Appl. Stat.*, 42, 443–59, 1993.

B. Efron, "Logistic regression, survival analysis and the Kaplan-Meier curve," *J. Am. Stat. Assoc.*, 83, 414–425, 1988.

J.E. Eichner, L.H. Kuller, T.J. Orchard, G.A. Grandits, L.M. McCallum, R.E. Ferrel, et al., "Relation of apolipoprotein E phenotype to myocardial infarction and mortality from coronary artery disease," *Am. J. Cardiol.*, 71, pp. 160–165, 1993.

W.J. Gauderman, and D.C. Thomas, "Censored survival models for genetic epidemiology: A Gibbs sampling approach," *Gen. Epidem.*, 11, pp. 171–188, 1994.

A.E. Gelfand, S.K. Ghosh, C. Christiansen, et al., "Proportional hazards models: a latent competing risks approach," *Appl. Stat.*, 49, 385–397, 2000.

A.E. Gelfand, and A.F.M. Smith, "Sampling-based approaches to calculating marginal densities," *J Amer. Stat. Assoc.*, 85, pp. 972–985, 1990.

J. Geweke, "Evaluating the accuracy of sampling-based approaches to calculating posterior moments," in *Bayesian Statistics 4*, (ed. J.M. Bernardo, J.O. Berger, A.P. Dawid, and

A.F.M. Smith). Clarendon Press, Oxford, 1992.

D.M. Hallman, E. Boerwinkle, N. Saha, C. Sandholzer, et al., "The apolipoprotein E polymorphism comparison of allele frequencies and effects in nine populations," *Am. J. Hum. Genet.*, 49, pp. 338–349, 1991.

J.E. Hixson, and D.T. Vernier, "Restriction isotyping of human apolipoprotein E by gene amplification and cleavage with Hhal," *J Lipid Res*, 31, pp. 545–548, 1990.

E.S. Iversen, G. Parmigiani, D.A. Berry, and J.M. Schildkraut, "Genetic susceptibility and survival: application to breast cancer," *J. Amer. Stat. Assoc.*, 95, pp. 28–42, 2000.

D. Kalbfleisch and R.L. Prentice, *The Statistical Analysis of Failure Time Data.* Wiley, New York, 1980.

W.A. Knaus, E.A. Draper, D.P. Wagner, and J.E. Zimmerman, "APACHE-II: A severity of disease classification system," *Crit. Care Med.*, 13, p. 10, 1985.

S.W. Lagakos, "General right censoring and its impact on the analysis of survival data," *Biometrics*, 35, pp. 139–156, 1979.

N.T. Lautenschlager, L.A. Cupples, V.S. Rao, S.A. Auerbach, et al., "Risk of dementia among relatives of Alzheimer's disease patients in the MIRAGE study: What is in store for the oldest old?" *Neurology*, 46, pp. 641–650, 1996.

J. Louhija, H.E. Miettinen, K. Kontula, M.J. Tikkanen, et al., "Aging and genetic variation of plasma apolipoproteins. Relative loss of the apolipoprotein e4 phenotype in centenarians," *Arterioscler. Tromb.*, 14, pp. 1084–1089, 1994.

C. McGilchrist, and C. Aisbett, "Regression with frailty in survival analysis," *Biometrics*, 47, pp. 461-6, 1991.

S.L. Normand, M.E. Glickman, R.G. Sharma, and B.J. McNeil, "Using admission characteristics to predict short-term mortality from myocardial infarction in elderly patients: results from the cooperative cardiovascular project," *J. Amer. Med. Assoc.*, 275, pp. 1322–1328, 1996.

I. Olkin, and R.F. Tate, "Multivariate correlation models with mixed discrete and continuous variables," *Ann. Math. Stat.*, 32, 448–465, 1961.

H. Payami, H. Grimslid, B. Oken, R. Camicioli, G. Sexton, A. Dame, D. Howieson, J. and Kaye, "A prospective study of cognitive health in the elderly (Oregon Brain Aging Study): effects of family history and apolipoprotein E genotype," *Am J Hum. Genet.*,

60, pp. 948–956, 1997.

J.H. Petersen, P.K. Andersen, and R.D. Gill, "Variance components models for survival data," *Statistica Neerlandica*, 50, pp. 193–211, 1996.

A.M. Saunders, W.J. Strittmatter, D. Schmechel, et al., "Association of apolipoprotein E allele epsilon-4 with late-onset familial and sporadic Alzheimer's disease," *Neurology*, 43, pp. 1467–1472, 1993.

F. Schachter, L. Faure-Delanef, F. Guenot, H. Rouger, P. Froguel, L. Lesueur-Ginot, and D. Cohen, "Genetic associations with human longevity at the Apo-E and ACE loci," *Nat. Genet.*, 6, pp. 29–32, 1994.

J.L. Schafer, *Analysis of Incomplete Multivarite Data.* Chapman and Hall, London, 1997.

D. Sinha, and D. Dey, "Semiparametric Bayesian analysis of survival data," *J. Amer. Stat. Assoc.*, 92, pp. 1195–1212, 1997.

D.J. Spiegelhalter, A. Thomas, N.G. Best, and W.R. Gilks, BUGS: Bayesian inference Using Gibbs Sampling, Version 0.5, (version ii), 1996.

J.H. Stengard, K.E. Zerba, J. Pekkanen, C. Ehnholm, A. Nissinen, and C.F. Sing, "Apolipoprotein E polymorphism predicts death from coronary heart disease in longitudinal study of elderly Finnish men," *Circulation*, 91, pp. 265–269, 1995.

G. Thomson, "Mapping disease genes: Family-based association studies," *American Journal of Human Genetics*, 57, pp. 487–498, 1995.

P.W.F. Wilson, R.H. Myers, M.G. Larson, J.M. Ordovas, P.A. Wolf, and E.J. Schaefer, "Apolipoprotein E alleles, dyslipidemia, and coronary heart disease: The Framingham Offspring Study," *J. Amer. Med. Assoc.*, 272, 1666–1671, 1994.

Keywords

Causal inference, Cardiovascular disease, Censored data, Multivariate survival model, Observational study.

Affiliation of authors and acknowledgments

Mark E. Glickman: Department of Mathematics and Statistics, Boston University

David R. Gagnon: Department of Epidemiology and Biostatistics, Boston University

Partial funding provided by the Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC). The authors would like to thank Prof. Adrienne Cupples for her helpful suggestions.

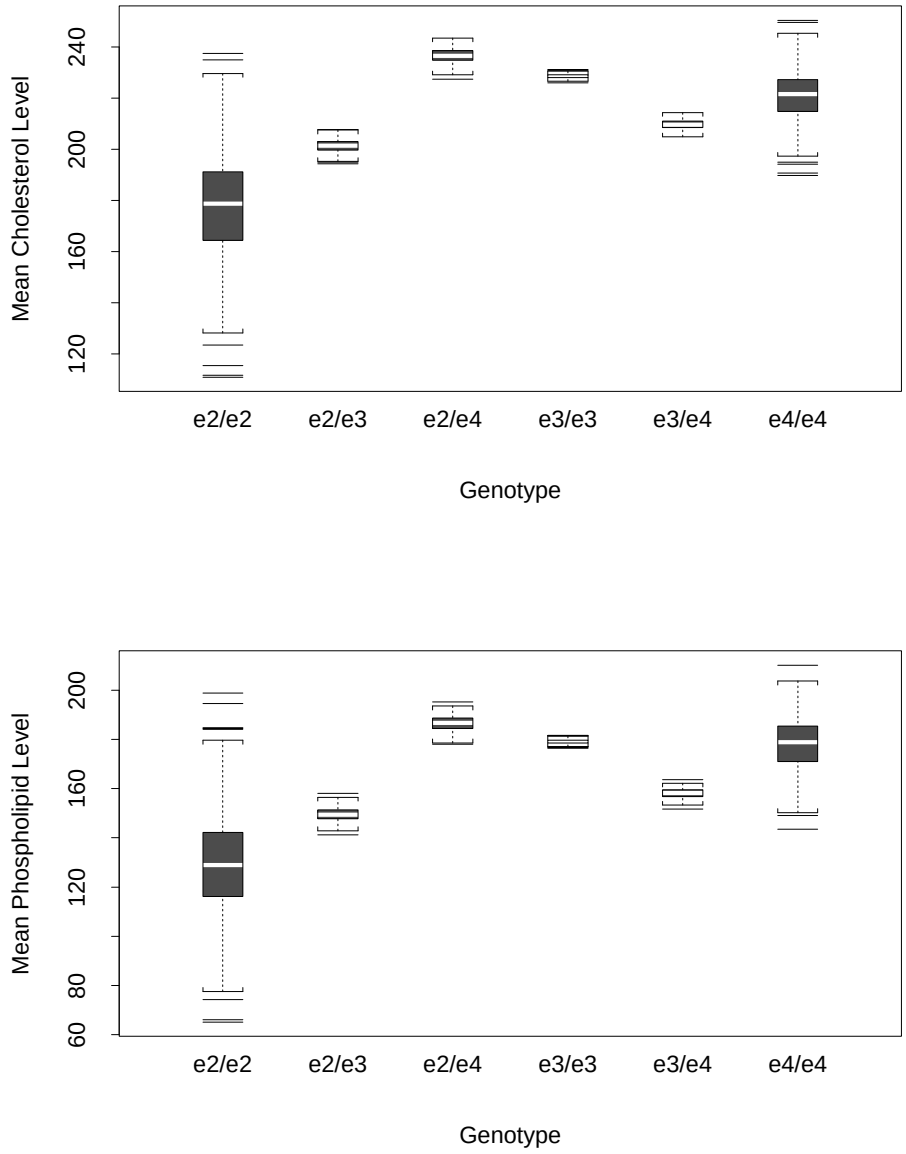

Figure 1: Estimated posterior distribution of α_{X_1} and α_{X_2}. Top: Posterior distribution of mean cholesterol level by genotype. Bottom: Posterior distribution of mean phospholipid level by genotype.

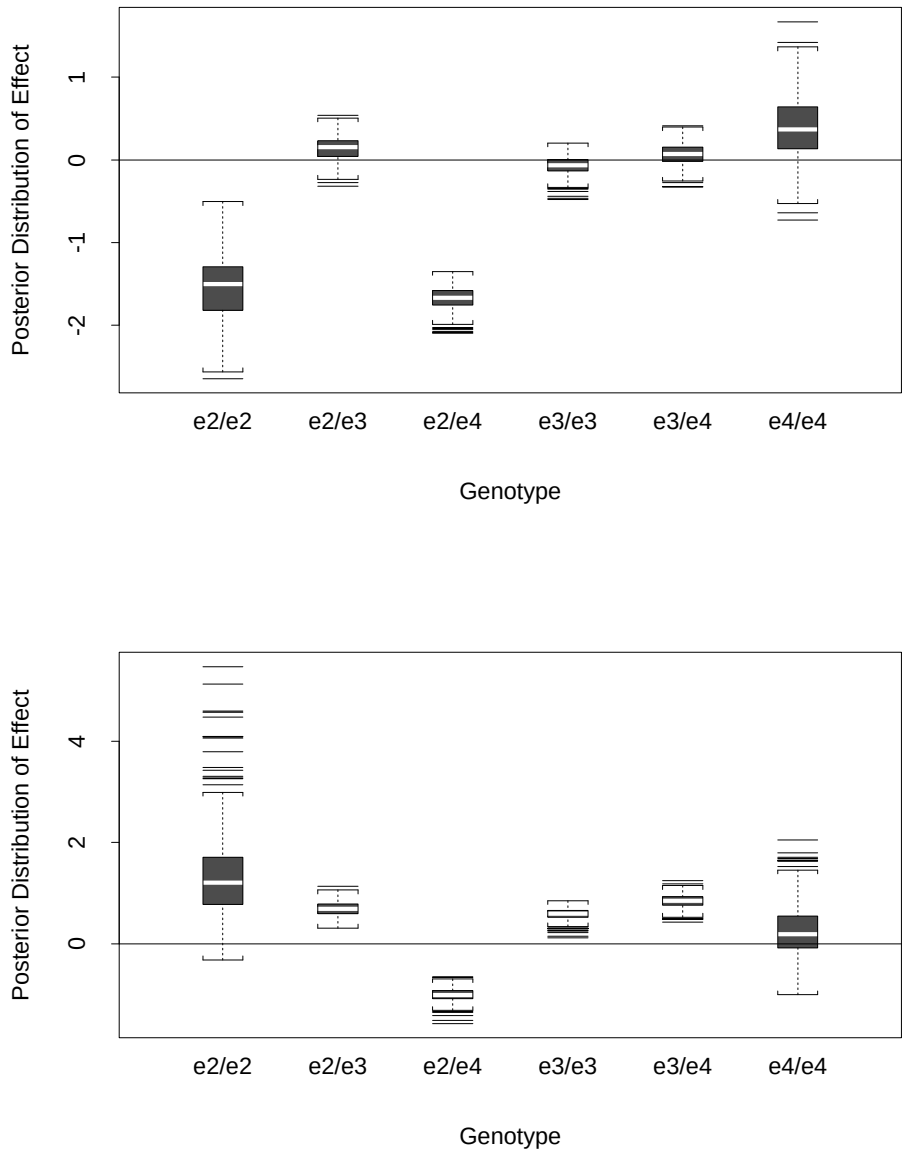

Figure 2: Estimated posterior distribution of $\gamma_{w,g}$, the effect of genotype on CVD onset age. The effects were normalized to sum to 0. Top: The distribution of $\gamma_{w,1}$, the effect of genotypes on CVD onset age for males. Bottom: The distribution of $\gamma_{w,2}$, the effect of genotypes on CVD onset age for females.