Apo-E genotypes and cardiovascular diseases: A sensitivity study using cross-validatory criteria

Running title: Apo-E and cardiovascular diseases

Mark E. Glickman*
Department of Health Services
Boston University

Mei-Fang Kao
Department of Mathematics and Statistics
Boston University

Abstract

The Apolipoprotein-E (Apo-E) gene, a gene that produces proteins which help to regulate lipid levels in the bloodstream, is of interest in the study of cardiovascular diseases. An approach to making inferences about the genetic effects of the Apo-E gene has been developed by Glickman and Gagnon (2002). The framework describes the role of genetic and risk factors on the onset ages of multiple diseases, and accounts for the possibility that an individual was censored for reasons related to the diseases of interest. The framework also allows for missing genetic information, so that subjects censored prior to genetic sampling, and therefore missing such information, may still be included in the analysis. We apply an extension to this framework to the original cohort of the Framingham Heart Study for measuring the effects of different Apo-E genotypes on the onset age of various cardiovascular disease events. In particular, we compare the fit of univariate versus multivariate onset age components to the model, whether to incorporate health covariates measured at baseline or at a point later in the study, and whether to assume a heritability model for Apo-E genotype frequencies. The results of the best fitting model are presented.

Keywords: Bayesian survival model, Causal inference, Cardiovascular disease, Framingham Heart Study, Multivariate survival model.

*Work supported by NIH grant R03 AG19875-01. Address for correspondence: Center for Health Quality, Outcomes & Economics Research, Edith Nourse Rogers Memorial Hospital (152), Bldg 70, 200 Springs Road, Bedford, MA 01730, USA. E-mail address: mg@bu.edu. Phone: (781) 687-2875, Fax: (781) 687-3106.

1 Introduction

Apolipoprotein-E (Apo-E) is a gene of interest in the study of various pathologies, including the study of Alzheimer's disease (Saunders et al. 1993), sleep apnea (Kadotani et al. 2001), and cardiovascular diseases (CVD) (Eichner et al. 1993, Wilson et al. 1994, Schachter et al. 1994, Stengard et al. 1995). In all these and many other papers, polymorphism of the Apo-E gene have been associated with varying levels of plasma concentrations of cholesterol, so it has been a natural focus of medical research concerning diseases that stem from elevated lipid levels in the bloodstream. Recent work on the connections between the Apo-E gene and cardiovascular diseases have produced mixed results. Heijmans et al. (2002) find that, among a cohort of subjects over 85 years of age, little evidence points to a relationship between Apo-E genotype and cardiovascular mortality. Lahoz et al. (2001) conclude differential risk in cardiovascular disease among subjects with different Apo-E genotypes. In a recent review study, Eichner et al. (2002) find modest associations to cardiovascular disease risk and Apo-E polymorphism, but not overwhelming evidence. It is clear that the entire story on the genetic effects of Apo-E genotypes on cardiovascular disease risk has not been fully told.

This paper examines sensitivity to various assumptions for a class of models to measure the effects of Apo-E genotypes on the onset age of four different cardiovascular disease events. We use data obtained from the Framingham Heart Study, a large cohort that has been followed since 1948 to study factors influencing the onset of cardiovascular diseases. Our approach, which involves Bayesian estimation of a parametric multivariate survival model in the presence of censored onset ages, extends the one described in Glickman and Gagnon (2002) by including a model component that accounts for disease-related death.

In the context of our framework, we consider various model assumptions to test. These include the choice of whether to use baseline covariates or covariates collected more recently, whether to assume a population equilibrium for the Apo-E alleles, and whether to assume common genetic effects for all four cardiovascular disease onset ages. The approach we adopt for model selection is in the form of cross-validation, in which we fit candidate models on portions of the data, and examine their predictability on the remainder. Popular Bayesian-justified information criteria to compare models, such as the deviance information criterion (Spiegelhalter et al., 2002), are more easily implementable alternatives to cross-validation, though these criteria are based on asymptotic arguments and may not work well for complex model structures, such as the ones we examine in this paper. Our exploration of the data is essentially a sensitivity analysis to various model assumptions. While our work is not intended to be a complete analysis of the Framingham Heart Study data to determine the genetic effects of the Apo-E genotypes on CVD onset, the goal here is to provide some clues about reasonable modeling assumptions that might lead to an improved understanding of the connection between Apo-E polymorphism and cardiovascular disease onset.

The paper is organized as follows. We discuss the data used in our study in Section 2. This is followed in Section 3 with the statistical model for the genetic effects on CVD onset ages. We explain our approach to examining the sensitivity to modeling and data assumptions in Section 4, the results of which are described in Section 5. We conclude the paper in Section 6 with a discussion of our methods and results.

2 Framingham Heart Study data

The data we use to study the genotypic effects of Apo-E polymorphism on cardiovascular disease onset come from the Framingham Heart Study. 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.

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 four CVD events. These included the age of first occurrence of angina pectoris, recognized acute myocardial infarction (MI), unrecognized ("silent") acute myocardial infarction, and congestive heart failure (CHF). Unrecognized MI 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. For subjects that died, we obtained the age of death, as well as whether death was due to reasons related to a CVD (not just restricted to the four CVD events in the study). If a subject lived or was lost to follow-up, we recorded the age at which the subject was last observed 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.

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 produces proteins that 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), and at higher risk for sleep apnea (Kadotani et al. 2001) than individuals without. 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.

Health covariates were also obtained for use in our statistical models. These included smoking frequency, forced vital capacity (cl/s), serum glucose level (mg/dl), systolic blood pressure (mm Hg), hemoglobin level (mg/ml), body mass index (kg/m^2), serum cholesterol level (mg/dl) and phospholipid level (mg/dl). Information on these health covariates were obtained at every exam in the Framingham Heart Study, and could be expected to be related to CVD onset.

For the 1266 subjects with recorded Apo-E genotypes, Table 1 shows the mean age of disease onset for the four different CVDs we examine, and the average age of CVD-related and non-CVD-related death ages. The average disease onset ages are generally in the late 60s

Genotype	Angina Pectoris	Unrecognized MI	Recognized MI	CHF	CVD Death	non-CVD Death
e2/e2	67.9 (NA)	67.36 (NA)	NA (NA)	NA (NA)	NA (NA)	90.9 (0.2)
e2/e3	68.0 (0.8)	68.9 (1.0)	71.1 (1.1)	77.9 (0.9)	84.0 (0.5)	81.8 (0.6)
e2/e4	65.2 (2.7)	NA (NA)	74.2 (2.3)	86.9 (0.3)	89.2 (0.2)	85.6 (2.3)
e3/e3	67.9 (0.3)	72.4 (0.3)	72.6 (0.4)	78.7 (0.3)	81.7 (0.2)	84.0 (0.3)
e3/e4	66.3 (0.6)	78.5 (0.6)	72.1 (0.7)	74.6 (0.8)	81.4 (0.5)	83.6 (0.5)
e4/e4	69.8 (NA)	67.3 (1.5)	83.8 (0.8)	76.8 (2.5)	84.3 (1.1)	86.0 (0.2)

Table 1: Sample means (and standard errors) for the ages of disease onset, and ages of CVD-related death and non-CVD-related death, grouped by Apo-E genotype. With only one observation in a category, the standard error is missing (and denoted as "NA").

to early 80s. Average ages of death are slightly higher, tending to be in the eighth decade. No clear pattern emerges of how average disease onset ages vary by Apo-E genotype, certainly without accounting for health covariates.

3 Multivariate statistical model for disease onset

The multivariate model we adopt for measuring the effects of Apo-E genotypes based on the Framingham Heart Study data extends the one presented in Glickman and Gagnon (2002). We assume that $n = 4804$ subjects in the cohort are free of the $K = 4$ genetically-related and irreversible cardiovascular diseases under investigation. The model assumes that covariate information is collected on subjects prior to their developing any of the diseases, but the genetic information is obtained a substantial time beyond the beginning of the study. For subject i, let $A_{i1}, \ldots, A_{i4}$ be the onset ages of the 4 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 the variability of disease onsets. Let $\boldsymbol{W}_i = (W_{i1}, \ldots, W_{iM})$ be a vector of $M = 6$ binary indicators for the Apo-E genotype of subject i. Because our application involves gene mutations at a single locus, only one of the W_{im} would be 1, and the rest 0 for subject i.

Let $\boldsymbol{X}_i = (X_{i1}, X_{i2})$ denote a vector of covariates 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.

Table 2 displays health covariates 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 and phospholipid levels. Health covariates are non-missing for all subjects.

As is common and conventional with parametric survival models, we consider a Weibull distribution for onset age A_{ik} for subject i, disease $k = 1, \ldots, 4$, given by

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

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 2: 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.

with density

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

Here, we assume a log-link function for μ_{ik}

$$\log(\mu_{ik}) = (X_1)_i'\beta_k + \boldsymbol{W}_i'(\gamma_{0k} + Z_i\gamma_{1k}) \qquad (2)$$

where Z_i is 1 if subject i is female, and 0 if male, γ_{0k} is a vector of the six Apo-E genotype effects for males, and γ_{1k} is a vector of six incremental Apo-E genotypic effects for females. The parameter μ_{ik} is composed of the additive effects of covariates X_{i1} not causally related to the Apo-E genes through a linear term with β_k, and of genetic factors $\boldsymbol{W}_i'(\gamma_{0k} + Z_i\gamma_{1k})$. For each disease k, the genetic factors comprise 12 possible effects, each one indexed by a gender and Apo-E genotype combination. In the actual model implementation, only 11 free parameters per disease is assumed to avoid parameter aliasing. Lahoz et al. (2001) were able to infer a relationship between Apo-E genotype and gender on the onset of cardiovascular diseases. Before constructing our model for onset ages of the four CVD events, we examined scatter plots which suggested transformations of several covariates. 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. As explained in Glickman and Gagnon (2002), avoiding the incorporation X_{i2} into the onset age component of the model is to prevent potential confounding of cholesterol and phospholipid levels with the Apo-E genotypes, as these covariates would be in the causal pathway to the effect on onset ages.

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 and Dey (1997). We argue for the use of parametric hazards 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. With a large fraction of missing and censored data, precise inferences about model parameters is not realistic. The effectiveness of parametric hazards is argued by Efron (1988), and more recently by Gelfand et al. (2000), for such situations.

We further model cholesterol and phospholipid levels, the covariates that are possibly causally related to the genetic factors, X_{i2}, as a function of the $\boldsymbol{W}_i$. We assume that $(X_2)_{i1}$, cholesterol level for subject i, and $(X_2)_{i2}$, the phospholipid level for subject i, can be modeled as a function of genotype,

$$(X_2)_{ij} \sim \mathrm{N}(\boldsymbol{W}_i \alpha_{Xj},\ \sigma^2_{Xj}) \tag{3}$$

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. From examining distribution plots of cholesterol and phospholipid levels, these variables remained untransformed. 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. 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.

In the Framingham Heart Study data, we measure not only disease onset ages, but also age of CVD-related death, disease death unrelated to CVD (e.g., due to cancer), and accidental death or censoring due to study termination. For each subject, exactly one of these three ages is observed, and is therefore assumed to censor the other two events that could have otherwise occurred. We extend the framework in Glickman and Gagnon (2002), which does not distinguish among the different events that censor disease onset ages, to specifically include a model component that accounts for cardiovascular disease-related death. Letting R_i be the age at which a subject dies for reasons arguably related to cardiovascular disease, we assume

$$R_i \mid \boldsymbol{X}_i, \boldsymbol{W}_i, Z_i, \delta_R, \mu_{iR} \sim \text{Weibull}(\mu_{iR},\ \delta_R) \tag{4}$$

with

$$\log(\mu_{iR}) = (X_1)_i'\beta_R + (X_2)_i'\alpha_R + Z_i\eta_R + \boldsymbol{W}_i'\theta,$$

where δ_R is the Weibull shape parameter specific to the distribution of the R_i, β_R and α_R are the non-causal and causal covariate effects, η_R is the gender effect, and θ are the genotypic effects on CVD age of death. The distribution of R_i depends on covariates both causally and non-causally related to the Apo-E genotypes, as well as to the genotypes themselves. The inclusion of this component of the model is intended to improve inferences on the missing $\boldsymbol{W}_i$.

Because the $\boldsymbol{W}_i$ are missing for some i, our model specification is completed by assuming a model for the $\boldsymbol{W}_i$. In our subsequent analyses, we consider two possible model components. The most general model assumes

$$\boldsymbol{W}_i \mid \boldsymbol{\pi} \sim \text{Multinomial}(1, \boldsymbol{\pi}) \tag{5}$$

for 6-vector parameter of probabilities $\boldsymbol{\pi}$. The second model we consider places constraints on the multinomial model by assuming that the allele frequencies in the population are in Hardy-Weinberg equilibrium (which is crudely supported by the distribution of the observed data). Letting p_2, p_3 and p_4 represent the allele frequencies in the study population of Apo-E e2, e3 and e4, respectively, the genotype frequencies 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, and the cholesterol and phospholipid parameters in (3) were assumed normal with mean 200 and variance 10000. For the Apo-E genotype probabilities, we assumed for the Hardy-Weinberg equilibrium model that the parameters (p_2, p_3, p_4) had a Dirichlet prior 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. For the unconstrained model, we translated these Dirichlet parameters to genotype parameters (that is, the Dirichlet parameter for π_1 was 0.05^2, for π_2 was $2(0.05)(0.85)$, and so on). 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 noteworthy feature of our model is that each subject has simultaneously any number between 0 and 4 disease onset ages $A_{i1}, \ldots, A_{i4}$ observed, though any or all may be censored. Furthermore, the age of cardiovascular disease death, R_i, which can be viewed as a type of disease onset age, can also be "censored" in the sense that study termination or non-cardiovascular disease death prevents R_i from being observed. In our model, when R_i is not observed, we act as though it would have been observed eventually as if it were a censored event, so that the contribution of a missing R_i to the likelihood is the right-tail probability of the Weibull distribution.

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 sets of conditional posterior distributions: the distribution of the missing A_{ik} and R_i 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 each of the conditional distributions is described in detail in Glickman and Gagnon (2002). For the current study, the Gibbs sampler was implemented in the statistics software package BUGS (Spiegelhalter et al., 1996).

4 Data and modeling assumptions

Using the framework in Section 3, we examine various modeling and data assumptions to obtain the best-fitting model. Specifically, we vary three different factors relevant to modeling the onset age distributions, and then evaluate the fit using cross-validatory criteria. The resulting model fit is described in Section 5.

The first modeling assumption we explored was to examine the different diseases univariately, versus multivariately, within one model. Modeling individual disease onset ages separately makes no assumptions about connections between the effects of genotypes and covariates. The tradeoff, however, is that inferences may be less precise because of the abundance of missing genetic information and censored CVD event ages. A multivariate model allows for the possibility of pooling information across the four cardiovascular diseases. The pooling was carried out by assuming that the effects of gender and Apo-E genotype are identical for all four diseases. This corresponds to assuming that $\gamma_{ik} = \gamma_i$ in Equation (2), that is, the γ_{ik} are identical for $k = 1, \ldots, 4$. An alternative approach, not explored in this paper, would be to assume a random effects distribution for the γ_{ik} rather than assuming them equal across the four diseases.

Because the covariates are measured at least every two years in the Framingham Heart Study, several alternative methods exist for their inclusion. The two approaches we consider in our study consist of modeling CVD onset ages as a function of covariates measured at baseline, and incorporating the most recently measured covariate information prior to any onset of disease. For the second approach, we use the covariate information for each subject from an exam that occurred no more than two years prior to the onset of one of the four diseases. A clear advantage of using the more recent covariate information is that the temporal relationship to disease onset should be more pronounced and predictive, so that the variability in onset ages due to health covariates will be adjusted to full effect. A disadvantage of this approach, however, is that temporal changes in health covariates may be related to unobserved or unmeasured diseases that may have a cardiovascular origin. For example, suppose a subject in the study has hypertension onset before the occurrence of any of the four measured diseases. Then hypertension may affect covariate levels, which in turn may result in the onset of one of the observed diseases. If there is an impact of the Apo-E genotypes on hypertension, then the effect of the recent covariates are confounded with the effect of the genotypes. Alternatives to the choices we consider include methods for tracking changes in covariates over time. For example, models with time-dependent covariates are possible, but this approach has the aforementioned disadvantages connected with using recent covariate values. Also, incorporating covariates measured temporally into a model has the difficulty that the magnitude of temporal changes in covariates may have varying impact on disease onset age across subjects.

The third factor is the assumed distribution on genotype frequencies. The most natural assumption is that the Framingham population is in genetic equilibrium and that the allelic distribution follows Hardy-Weinberg equilibrium. This assumption induces constraints on the genotype frequencies. We model the frequencies as arising from a Dirichlet distribution that has three parameters (representing the allele frequencies). While clearly population genetics arguments would lead to an allelic distribution that is close to that assumed under Hardy-Weinberg equilibrium, it is not necessarily exactly so due to other potential selection sources. In addition to considering Hardy-Weinberg equilibrium, we model genotype frequencies as following an unconstrained Dirichlet distribution with six parameters instead of three.

We therefore consider $2^3 = 8$ modeling assumptions to explore:

1. Univariate onset age model versus multivariate model.

2. Including covariate information measured at study baseline, versus including covariate information measured most recently before any CVD onset age.

3. Assuming an unstructured prior distribution for the genotype frequencies, versus assuming a prior distribution consistent with Hardy-Weinberg equilibrium.

To assess and compare the adequacy of the different models, two forms of validation are considered. Each form involves a cross-validatory criterion in which the proposed models are fit on a portion of the data, and then we determine how well each fitted model predicts the data left out.

The first model assessment approach involves predicting onset ages from randomly deleted cases, and comparing the predictive discrepancy between predictions and actual onset ages. We use squared deviations as the form of the discrepancy, though clearly other choices are possible (e.g., those based on the Weibull deviance function). For each model considered, a randomly chosen subset of 400 cases out the 4804 are left out as a validation sample. The model is then fit to the remaining 4404 subjects. For the validation sample of 400 subjects, posterior simulated values of the onset ages are computed using the results of MCMC posterior simulation. The average squared deviation of the simulated onset ages from the observed onset age is computed for each subject having an observed onset age, and the results are averaged across the entire validation sample. The process of randomly choosing a subset of 400 to leave out, fitting the model, and calculating the predictive goodness of fit is repeated three more times, and the resulting goodness of fit summary is averaged across the four analyses. The repeated analyses are carried out to reduce the dependence of the results on a single validation sample. Note that the same sets of splits between development and validation samples are used for each model we considered. The final value averaged over the four analyses is a measure of predictive goodness of fit, with smaller values indicating better fits. This value can be viewed as a variance measure, so it has an interpretation as such.

The second model assessment makes use of all subjects, and focuses on how well the model infers genotypes. The statistic we compute is an average of probabilities of correctly predicting an artificially missing genotype. The final value of the statistic is an estimate of the average posterior probability of correctly identifying genotypes from an external sample. An alternative would be to calculate the log of the probability, and average these across subjects and diseases – this particular statistic is closer to the form of a multinomial deviance. The calculation is carried out as follows. A randomly chosen subset of 800 subjects is selected. The genotype information, if observed, is then treated as missing. This process mimics what would happen if the age of DNA sampling for these individuals is artificially increased, so that genotyping never takes place for that subject while in the study. The model is then fit by the modified data, and posterior simulated values are produced for the deleted genotypes. For each subject with an observed genotype that was deleted, the proportion of simulated genotypes that match the observed genotype is recorded, and averaged across all subjects. Again, this process of selecting 800 subjects at random, deleting the genotype information, fitting the model, and recording the average proportion of matches is repeated three more times. The final value is averaged over the four analyses to produce an overall measure of the probability of recovering missing genotypes. Larger values indicate better model performance.

To assess the variability of the cross-validatory criteria, we performed standard two-way ANOVA on the cross-validation statistics for each of the eight models across the four validation sets. That is, in the two-way ANOVA, one factor consisted of the eight model choices, and the second factor consisted of the four validation sets. The results of the ANOVA provide an indication for whether the models differ on the cross-validation statistic, accounting for the different validation sets, and the magnitude of the variation of the cross-validation statistics through the mean-squared error.

5 Results

For each of the eight models that was fit, single series Gibbs sampler with over-dispersed starting values were 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. 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 cross-validation model diagnostics described in Section 4 are displayed in Tables 3 and 4. In Table 3, we compute the approximate posterior means for the mean squared deviations between the predicted and actual onset ages. From the two-way ANOVA performed on these statistics across the four validation samples, the residual standard deviation was 0.90, and that the differences among the models was not compellingly strong (the p-value for the difference among models was 0.27). Carrying out the ANOVA on the logarithms of the statistics did not substantively change the conclusions. However, examining the individual assumption-based factors on the cross-validation statistics leads to some noteworthy comments. The one factor that distinguishes among models through the predictive cross-validation measure is whether Hardy-Weinberg equilibrium is assumed for the allele frequencies. When the genotype distribution is unrestricted, the posterior variances are consistently lower than assuming Hardy-Weinberg equilibrium for the alleles. To a lesser extent, assuming different genetic effects on the onset ages across diseases produces lower variances on average, so that assuming different genetic effects is slightly preferred on this measure. The comparison of baseline versus recent data does not reveal any notable difference using predictive standard deviations. On the basis of design issues, using the baseline health data seems slightly preferable because we are more certain that unobserved confounding health variables are not as relevant at baseline. The single best model according to the posterior mean predictive variance is the one that uses baseline covariates, different genotypic effects, and an unrestricted prior distribution on genotype frequencies.

Covariates	Genetic effects across diseases	Genotype distribution	Posterior mean predictive variance
Baseline	Different	Unrestricted	6.86
Baseline	Different	Hardy-Weinberg	8.12
Baseline	Same	Unrestricted	7.34
Baseline	Same	Hardy-Weinberg	7.79
Recent	Different	Unrestricted	7.04
Recent	Different	Hardy-Weinberg	7.88
Recent	Same	Unrestricted	7.38
Recent	Same	Hardy-Weinberg	8.33

Table 3: Onset age cross-validation results. For each of the eight models, the posterior mean of the cross-validated predictive variance is computed. The root MSE across the validation samples is 0.90.

The posterior mean genotype matching probabilities are displayed in Table 4. The larger the reported probability, the greater the frequency that the model recovers the genotype of randomly deleted genes. From the two-way ANOVA associated with the matching probabilities, differences among the models accounting for the variation due to the different validation samples are more compelling than in the predictive squared deviation analysis. Here, the p-value for the model factor is 0.026, and the standard deviation of the probability statistic adjusted for the validation samples is 0.018. We therefore view the information in Table 4 as more reliable in distinguishing among the different modeling assumptions. The pattern of probabilities reveals that a greater chance at recovering randomly deleted genes when the prior distribution on genotype frequencies is unrestricted. This is true for all four combinations of the other two factors. Whether baseline covariates are used, or whether the genotypic effects are assumed the same across diseases has little impact on the matching probabilities. The best model suggested by this measure uses baseline covariates, unrestricted prior distribution on genotype frequencies, and common genotypic effects across disease.

We first examined the results of fitting the model that is most preferred by the posterior

Covariates	Genetic effects across diseases	Genotype distribution	Posterior mean matching frequency
Baseline	Different	Unrestricted	0.429
Baseline	Different	Hardy-Weinberg	0.383
Baseline	Same	Unrestricted	0.430
Baseline	Same	Hardy-Weinberg	0.400
Recent	Different	Unrestricted	0.414
Recent	Different	Hardy-Weinberg	0.413
Recent	Same	Unrestricted	0.416
Recent	Same	Hardy-Weinberg	0.402

Table 4: Genotype matching cross-validation results. For each of the eight models, the posterior mean of the genotype match probabilities is computed. The root MSE across the validation samples is 0.018.

predictive variance criterion, where genetic effects are different for distinct diseases. In doing so, we refit the model with a burn-in period of 20,000 MCMC iterations, and then calculated posterior summaries based on 5,000 subsequent iterations. While slight differences existed in the genetic effects between diseases, the relative magnitudes were similar enough to prefer the more parsimonious model having identical genetic effects. This is the model for which the matching probability criterion indicates the greatest preference. The reliance on the matching probability analysis is also suggested by the greater precision of the cross-validation statistic in the two-way ANOVA. Summaries of the genetic effects are displayed in Table 5.

For two genotypic effects, γ_1 and γ_2, and Weibull shape parameter λ, the relative effect on median onset age is $\exp(-(\gamma_1 - \gamma_2)/\lambda)$. Thus larger values in Table 5 correspond to earlier onset ages. The genotype most associated with early CVD onset age is e2/e4, particularly for males, and this result is consistent with that described in Glickman and Gagnon (2002). With a posterior mean shape parameter of 7.34 (the 95% central posterior interval is (7.11, 7.44)), we can estimate that, for example, males with e2/e4 have CVD onset ages typically

Genotype	Posterior Mean Effect	95% Central Posterior Interval
Male, e2/e2	−1.136,	(−3.331, 0.982)
Male, e2/e3	−0.880,	(−1.406, −0.359)
Male, e2/e4	2.048,	(1.769, 2.372)
Male, e3/e3	−0.203,	(−0.473, 0.121)
Male, e3/e4	−0.847,	(−1.207, −0.494)
Male, e4/e4	0.935,	(0.609, 1.313)
Female, e2/e2	−1.212,	(−2.910, −0.040)
Female, e2/e3	−1.074,	(−1.570, −0.612)
Female, e2/e4	1.322,	(1.071, 1.678)
Female, e3/e3	−0.909,	(−1.154, −0.573)
Female, e3/e4	−1.583,	(−2.011, −1.156)
Female, e4/e4	−0.341,	(−0.681, 0.045)

Table 5: Approximate posterior means and 95% central posterior intervals for the genotypic effects on age of CVD onset.

Genotype	Posterior Mean Effect	95% Central Posterior Interval
e2/e2	−1.425,	(−5.279, 0.757)
e2/e3	−0.932,	(−2.371, −0.039)
e2/e4	2.347,	(1.243, 3.145)
e3/e3	−0.047,	(−1.193, 0.732)
e3/e4	−0.887,	(−2.107, −0.084)
e4/e4	0.944,	(−0.287, 1.813)

Table 6: Approximate posterior means and 95% central posterior intervals for the genotypic effects on age of CVD death.

about 14% lower than males with e4/e4, and about 26% lower than males with e3/e3. It is also notable that males have earlier onset of CVD compared to females, specifically about 10% earlier. The main exceptions are e2/e2 and e2/e3 where the onset ages are similar.

Interestingly, as shown in Table 6, in the component of the model predicting age of CVD-related death, R_i, the effects of the different genotypes are similar to those predicting onset ages. Specifically, the e2/e4 genotype is associated with earliest age of CVD death, followed by the e4/e4 genotype. The e2/e2 genotype appears most protective, though the

Apo-E Genotype	Percent observed in sample	95% Central Posterior Interval
e2/e2	0.39	(0.09, 0.66)
e2/e3	10.74	(6.84, 9.46)
e2/e4	1.74	(12.69, 15.63)
e3/e3	66.82	(49.57, 53.84)
e3/e4	18.80	(15.47, 19.09)
e4/e4	1.50	(7.32, 9.79)

Table 7: Frequency distribution of genotypes in sample of 1266 individuals with non-missing genetic information, and estimated population frequencies from the model.

large posterior interval casts uncertainty about this conclusion.

The health covariates in the model generally contribute substantially to the model fit. In the onset age components of the model, the coefficients for body mass index, cigarettes per day, and systolic blood pressure had central 95% posterior intervals that were positive for all four diseases (so that the greater the covariate values, the lower the average disease onset ages). The quadratic component for systolic blood pressure was not an important predictor here. Forced vital capacity was significantly associated only with lower onset age of angina and slightly higher onset age of congestive heart failure, and blood sugar was quadratically associated with age of recognized myocardial infarction and congestive heart failure. Hemoglobin levels had a slight quadratic association with age of recognized myocardial infarction. In the model component predicting age of CVD-related death, body mass index and cigarettes per day are associated with lower age of death, females tend to have higher age of CVD death, and blood sugar and systolic blood pressure were quadratically significant. Hemoglobin levels were not predictive of age of CVD death, nor were cholesterol and phospholipid levels.

Table 7 displays the observed genotype distribution in the cohort among the 1266 individuals who had DNA sampling, and the estimated distribution from the model (which extrapolates to the study population represented by the original 4804 subjects). It should be noted that a simple likelihood ratio chi-squared test for Hardy-Weinberg equilibrium has a p-value of 0.211, so that the observed frequencies are not inconsistent with Hardy-Weinberg equilibrium. According to the model, the frequency of the e4 allele is inferred to be much higher than in the fraction of sample where DNA information was obtained. This appears primarily to come through the e2/e4 genotype. Without heritability restrictions on the genotype distribution, the model predicts that of the 3538 subjects with missing DNA information, between roughly 15 and 19 percent had the e2/e4 genotype. The model also predicts that the e4/e4 genotype is underrepresented among subjects with observed genotypes. Because a large fraction of the cohort with unobserved genotypes evidenced CVD events earlier than the fraction with observed genotypes, and the observed e2/e4 individuals had early CVD onset ages, the model infers that a large number of individuals with missing DNA information who have low CVD onset ages are likely to have the e2/e4 genotype.

6 Discussion

The framework presented in this paper, which extends Glickman and Gagnon (2002) by including a model component that accounts for disease-related death, 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 via MCMC simulation from the posterior distribution, missing onset ages and genotypes can be addressed in a straightforward manner.

Our analysis of CVD events in the original Framingham Heart Study cohort using a cross-validatory approach for model selection has resulted in several noteworthy observations. First, imposing the distributional restriction on allele frequencies through the assumption of Hardy-Weinberg equilibrium results in a worse fit. Both cross-validatory criteria fared better (with one exception) when no restrictions were assumed for the genotype frequencies. Interestingly, the incorporation of health covariates measured shortly before CVD onset instead of baseline covariates does not improve the model's predictive features. According to the best-fitting model, the e2/e4 Apo-E genotype is most associated with earlier CVD onset age. This result is corroborated by Glickman and Gagnon (2002), and the deleterious effect of e2/e4 seems to persist in alternative models fit by the same data.

A curious result of our analysis is that among the fraction of the cohort that did not survive to have their DNA sampled, a surprisingly large number of them are inferred (on average) to have the e4 allele as part of their Apo-E genotype. For example, the inferred fraction of the entire cohort with the e4/e4 Apo-E genotype is roughly 5 times as many as observed, and the inferred fraction of the cohort with the e2/e4 Apo-E genotype is about 8 times as many as observed. These inferred fractions are large enough to imply a serious violation of Hardy-Weinberg equilibrium, which would have the fraction of missing genotypes be closer in magnitude to the observed fractions. Of course, Hardy-Weinberg equilibrium is merely an assumption about the distribution of genotypes in the population, and is not verifiable from the observed data. Even though 15% of the original cohort possessing the e2/e4 Apo-E genotype seems like a big increase from the 2% observed in the sample, the validation criteria, especially the missing genotype recovery calculation, points towards this model being more consistent with the data than the models that assume Hardy-Weinberg equilibrium.

The analyses in this paper are not meant to be a conclusive examination of the genotypic effects of Apo-E on CVD onset, but instead an exploration of sensible modeling assumptions. With roughly 3/4 of the cohort having missing genetic information, it is difficult to report scientific conclusions without a large amount of uncertainty. While the original Framingham Heart Study cohort is a rich data set for an examination of genetic epidemiological models, more thorough modeling could involve, for example, the Framingham Heart Study offspring cohort to supplement the information in the original cohort (a crude analysis of the connection between Apo-E genotypes and CVD onset on the offspring cohort was carried out in Wilson et al., 1994). The sensitivity analyses in this paper serve to demonstrate that certain model choices can be considered by cross-validatory criteria, and that the results of such analyses can be used as the starting point for a more thorough investigation of genetic effects when a large fraction of a cohort has missing information.

References

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

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

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

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

Eichner, J.E., Dunn, S.T., Perveen, G., Thompson, D.M., Stewart, K.E., and Stroehla, B.C. (2002), "Apolipoprotein-E polymorphism and cardiovascular disease: A huge review," *Am. J. of Epidemiology*, **155**, 487–495.

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

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

Gelfand, A.E., Ghosh, S.K., and Christiansen, C., Soumerai, S.B., and McLaughlin, T.J. (2000), "Proportional hazards models: A latent competing risks approach," *Applied Stat.*, **49**, 385–397.

Glickman, M.E., and Gagnon, D.R. (2002), "Modeling the effects of genetic factors on late-onset diseases in cohort studies," *Lifetime Data Analysis*, **8**, 221–228.

Geweke, J. (1992), "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, UK.

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

Heijmans, B.T., Slagboom, P.E., Gussekloo, J., Droog, S., Lagaay, A.M., Kluft, C., Knook, D.L., and Wetendorp, R.G. (2002), "Association of Apo-E $\epsilon 2/\epsilon 3/\epsilon 4$ and promoter gene variants with dementia but not cardiovascular mortality in old age," *Am. J. Med. Gen.*, **107**, 201–208.

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

Kadotani, H., Kadotani, T., Young, T., Peppard, P.E., Finn, L., Colrain, I.M., Murphy Jr., G.M., and Mignot, E. (2001), "Association between Apolipoprotein-E $\epsilon 4$ and sleep-disordered breathing in adults," *J. Amer. Med. Assoc.*, **285**, 2888–2890.

Lahoz, C., Schaefer, E.J., Cupples, L.A., Wilson, P.W., Levy, D., Osgood, D., Parpos, S., Pedro-Botet, J., Daly, J.A., and Ordovas, J.M. (2001), "Apolipoprotein-E genotype and cardiovascular disease in the Framingham Heart Study," *Arteriosclerosis*, **154**, 529–537.

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

Saunders, A.M., Strittmatter, W.J., Schmechel, D., George-Hyslop, P.H., Pericak-Vance, M.A, Joo, S.H., Rosi, B.L., Gusella, J.F., Crapper-MacLachlan, D.R., and Alberts, M.J. (1993), "Association of Apolipoprotein-E allele epsilon-4 with late-onset familial and sporadic Alzheimer's disease," *Neurology*, **43**, 1467–1472.

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

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

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and Van der Linde, A. (2002), "On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Royal Stat. Soc., Ser. B*, **59**, 731–792.

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

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

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