Latent Class Models for Joint Analysis of Longitudinal Biomarker and Event Process Data: Application to Longitudinal Prostate-Specific Antigen Readings and Prostate Cancer

Haiqun LIN, Bruce W. TURNBULL, Charles E. MCCULLOCH, and Elizabeth H. SLATE

A retrospective substudy of the nutritional prevention of cancer (NPC) trials investigated the utility of longitudinally measured prostate-specific antigen (PSA) as a biomarker for subsequent onset of prostate cancer (PCa). Serial PSA levels were determined retrospectively from frozen blood samples that had been collected from all patients at successive clinic visits with the timing and the number of these visits highly variable. Diagnosis dates of all incident cases of PCa were recorded. Heterogeneity in PSA trajectories was observed that could not be fully explained by the usual linear mixed-effects model and measured covariates. Latent class models that incorporate both a longitudinal biomarker process and an event process offer a way to handle additional heterogeneity, to uncover distinct subpopulations, to incorporate correlated nonnormally distributed outcomes, and to classify individuals into risk classes. Our latent class joint model can aid the prediction of PCa probability given the longitudinal biomarker information available on an individual up to any date. The proposed model easily accommodates highly unbalanced longitudinal data and recurrent events. There are two levels of structure in the latent class joint model. First, the uncertainty of latent class membership is specified through a multinomial logistic model. Second, the class-specific marker trajectory and event process are specified parametrically and semiparametrically, under the assumption of conditional independence given the latent class membership. We use a likelihood approach to obtain parameter estimates via the EM algorithm. We fit the latent class joint model to the data from the NPC trials; four distinct subpopulations are identified that differ with regard to their PSA trajectories and risk for prostate cancer. Higher PSA level is significantly associated with increased risk of PCa, but appears to be conditionally independent once the latent classes are taken into account. Among the covariates, selenium supplementation and age at entry are statistically significant for various parts of the model. Assumptions—in particular the conditional independence between the longitudinal PSA biomarker and time to PCa diagnosis—are assessed.

KEY WORDS: Biomarker; EM algorithm; Joint longitudinal and event process model; Latent class; Risk classification; Semiparametric maximum likelihood estimator.

1. INTRODUCTION

Prostate cancer (PCa) is the most common cancer and the second leading cause of cancer death (after lung cancer) in U.S. men, with approximately 180,000 new PCa cases and 32,000 deaths from PCa in the United States (Greenlee, Murray, Bolden, and Wingo 2000). Early detection of PCa can often lead to cure; however, one characteristic of PCa is its lack of early symptoms. The introduction of a prostate-specific antigen (PSA) test has allowed diagnosis of PCa at a much earlier stage than was previously possible. The blood PSA concentration varies with the amount of normal and malignant prostate tissue, the location and type of cancer, and the extent of any existing infection or inflammation of the prostate. A normal blood PSA value is usually below 3–4 ng/ml and may vary with age, but about 20% of men diagnosed with PCa have measured PSA level below 4 ng/ml (Catalona, Smith, and Ornstein 1997). With an elevated PSA reading, ultrasonography, biopsy, or other test may be recommended by physicians. However, because other conditions, such as benign prostatic hyperplasia or prostatitis, can also increase PSA levels, a single high PSA measurement is not a highly specific indicator of PCa. This leads to consideration of whether and how serial PSA measurements taken over time can be used to aid more accurate prediction of PCa onset. Describing the joint behavior of longitudinal PSA readings and time to onset of prostate cancer is thus of great clinical and scientific interest. Clinical investigators (Morrell, Pearson, Carter, and Brant 1995) postulate that there may be different disease patterns; once the disease occurs, it proceeds aggressively in some men, whereas in others it progresses slowly and is unlikely to cause death. This may be because of genetic factors, such as the abl gene (Padilla-Nash et al. 2001).

Therefore, it is natural to consider both the PSA readings and PCa onset as response variables, rather than, say, using PSA as a time-varying predictor for onset of PCa in a hazard regression model. (The latter approach also has technical problems, because the time-varying predictor is not continuously observed and must be interpolated or extrapolated, which causes bias.) Tsiatis, DeGruttola, and Wulfsohn (1995), Faucett and Thomas (1996), and Wulfsohn and Tsiatis (1997) considered a joint modeling approach for a longitudinal biomarker and an event–time process in the context of study-

Haiqun Lin is Assistant Professor, Division of Biostatistics, Department of Epidemiology and Public Health, Yale University, New Haven, CT 06520 (E-mail: *haiqun.lin@yale.edu*). Bruce W. Turnbull is Professor and Chair, Department of Statistical Science, Cornell University, Ithaca, NY 14853 (E-mail: *bruce@orie.cornell.edu*). Charles E. McCulloch is Professor and Head, Division of Biostatistics, Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94143 (E-mail: *chuck@biostat.ucsf.edu*). Elizabeth H. Slate is Associate Professor, Department of Biometry and Epidemiology, Medical University of South Carolina, Charleston, SC 29425 (E-mail: *slateeh@musc.edu*). This article is based in part on the thesis of the first author in partial satisfaction for Ph.D. degree requirements at Cornell University. The research was supported by grants from the U.S. National Institutes of Health. The authors acknowledge Larry Clark of the Arizona Cancer Center for making the NPC trials data available to them and for his support of this research. Larry Clark passed away on March 20, 2000 from complications of prostate cancer before he could see the results. He was an esteemed colleague who highly valued the contributions of statistics to medical research. He will be dearly missed. The authors thank the editor, the associate editor, and the three referees for their constructive and thoughtful comments.

ing serial CD4 counts for progression to AIDS in HIV-positive patients. They all used growth curve models for longitudinal CD4 counts and then modeled the hazard rate using the imputed value of the marker as a time-dependent covariate. In a second approach, the CD4 marker process is modeled conditionally on the event times, which is then modeled marginally (Pawitan and Self 1993). A third approach to joint modeling uses latent variables to simplify the mutual dependence between a longitudinal marker and an event process. Schluchter (1992) and DeGruttola and Tu (1994) adopted shared random effects approaches, modeling the joint distribution of log-survival time and biomarker as multivariate normal conditional on a shared random variable. Henderson, Diggle, and Dobson (2000) modeled event times and longitudinal measurements separately, conditional on a bivariate latent stochastic process, so that the relationship between the longitudinal and event–time responses depends on assumptions concerning the relationship between the two latent processes.

In these approaches, event times and marker trajectories follow a single pattern. This may not be the case for our highly heterogeneous population. Another limitation of these joint models is the assumption of a common baseline hazard in the event process model, which could be violated for a potentially heterogeneous study population. Anderson and Fleming (1995) showed that in the proportional hazards model, performing regression adjustment for a predictive covariate (e.g., a biomarker), as well as omitting it entirely, will produce biased estimates of relative risk parameters. One way to correct for the bias is stratification, which is natural with discrete variables such as treatment group. However, the situation may become complicated when the stratification is based on more than one variable or involves a continuous variable, situations that require arbitrary discretization. Permitting subpopulation structure via latent classes handles heterogeneous baseline hazards. In this article we propose and apply a latent class joint model of the longitudinal PSA and the PCa onset outcome that accommodates all the preceding considerations. Our conceptually simpler latent class joint model permits distinct behavior within each subpopulation and more flexibility to model the association between the longitudinal and event–time responses.

Although there is a large literature on latent class models and "mixtures of experts" models (e.g., Rosen and Tanner 1999), there do not appear to be any applications of such an approach to the joint modeling of a longitudinal biomarker and an event–time outcome. This article generalizes the latent class models of Muthén and Shedden (1999) and Lin, McCulloch, Turnbull, Slate, and Clark (2000), who proposed latent classes in a fully parametric model to characterize different patterns of a longitudinal biomarker and a *binary* outcome in the setting of complete follow-up. Here the binary outcome is replaced by a censored survival outcome, which is modeled semiparametrically. Recent advances in semiparametric maximum likelihood inference (Parner 1998; Murphy and van der Vaart 2000) have laid the theoretical foundation for the semiparametric joint model proposed in this article. Additionally, we incorporate covariates into the model for survival outcome, whereas the models of Muthén and Shedden and Lin et al. for binary outcomes did not. Advantages of modeling disease outcome as a censored survival outcome rather than as incidence in a fixed time, say T, include incorporation of time-dependent covariates, estimates of survival over the entire time period, and inclusion of all PCa cases even beyond T and inclusion of patients with follow-up times less than T. For some datasets, the gain using a survival endpoint could be substantial. In both models, conditional independence between the binary/survival outcome and the longitudinal response given the latent classes are assumed. In this article we provide empirical methods to check this assumption.

This article is arranged as follows. In Section 2 we describe the Nutritional Prevention of Cancer (NPC) trials (Clark et al. 1996), which provide a good resource for studying the relation between longitudinally measured PSA and PCa (Slate and Clark 1999; Slate and Turnbull 2000). In Section 3 we present the latent class joint model, and in Section 4 we describe how maximum likelihood and the EM algorithm can be used to fit this model. In Section 5 we describe the results from fitting the NPC data; we compare and contrast our findings with some simpler analyses and with the analysis where PCa is treated as a binary outcome. Our results indicate that there are indeed distinct subpopulations of subjects that can be identified. In Section 6 we assess the fit and consider the validity of model assumptions. We end with a discussion in Section 7.

2. THE NUTRITIONAL PREVENTION OF CANCER TRIALS DATA

The NPC trials consist of two multicenter, double-blind, randomized trials designed to evaluate the cancer prevention potential of a daily nutritional supplement of selenium (Se). Patients were entered into the trial between September 15, 1983 and April 5, 1992. We consider only the 1,229 males so randomized. This dataset was closed on July 25, 1997. The time origin is taken as the date of entry (i.e., randomization) into the study. The event of interest is diagnosis of PCa, the date of which was recorded for all 82 incident cases during the study. Serial PSA levels were determined from frozen blood samples been collected from all patients at successive clinic visits, with the timing and number (range 1–20, median 4) of these visits highly variable among the participants. Both the PSA trajectories and times to diagnosis exhibit substantial subject-to-subject variability.

Various prognostic variables for each participant, (e.g., occupational exposures, smoking status) were also recorded. Of particular interest were the baseline Se and PSA blood levels taken at the initial entry visit, age at entry, and, of course, treatment group assignment. In the data from the NPC trials, we drop 29 subjects with a missing Se value and 18 for whom no PSA reading is available. For those diagnosed with PCa, we include only PSA readings before diagnosis, so that the effect on the biomarker of any resulting treatment is omitted. The resulting dataset that we analyze thus consists of 1,182 subjects, 82 of whom were diagnosed with PCa.

3. THE MODEL

The search for subpopulation structure leads us to a model that contains K latent classes, with each class representing a subpopulation that has its own pattern of longitudinal and event–time responses. Because PSA is acting as a biomarker for PCa, the hope is that, after the distinct PCa disease states

(as represented by the latent classes) are identified, PSA and PCa will be approximately conditionally independent within a class.

3.1 Submodel Specification

Our latent class model has three ingredients: class membership, the longitudinal biomarker trajectories, and the hazard for the time-to-event process. Suppose that we have n subjects ($n = 1{,}182$ in our PSA application), labeled $i = 1, \ldots, n$, and K latent classes labeled by $k = 1, \ldots, K$. We define $c_{ik} = 1$ if subject i is a member of class k and 0 otherwise. The probability that subject i belongs to latent class k is described through the multinomial distribution of the class membership vector for subject i, $c_i = (c_{i1}, \ldots, c_{iK})^{\mathrm{T}}$, modeled via a logit model with covariate vector $v_i = (v_{i1}, \ldots, v_{im})^{\mathrm{T}}$ and associated class-specific coefficient vector η_k,

$$\pi_{ik} = P(c_{ik} = 1) = \frac{\exp\left(v_i^{\mathrm{T}}\eta_k\right)}{\sum_{j=1}^{K}\exp\left(v_i^{\mathrm{T}}\eta_j\right)}, \tag{1}$$

where π_{ik} denotes the probability that subject i belongs to latent class k and η_k is the coefficient vector for class k with $\eta_1 = 0$.

Each subpopulation has its own model for the longitudinal biomarker readings, y, which is represented by a linear mixed model with subpopulation differences entering the mean,

$$y_i = X_i\beta + Z_i b_i + W_i(Mc_i) + \varepsilon_i. \tag{2}$$

Here $y_i = (y_{i1}, \ldots, y_{in_i})^{\mathrm{T}}$ is the n_i vector of biomarker readings for subject i. Fixed covariates from subject i are represented by the $n_i \times p$ matrix $X_i = (X_{i1}^{\mathrm{T}}, \ldots, X_{in_i}^{\mathrm{T}})^{\mathrm{T}}$, with associated p vector of coefficients β. The tth row of X_i, denoted by X_{it}^{T}, is thus a p vector of covariate values measured at occasion t. Covariates for random effects and for class-specific effects are denoted by the $n_i \times q_1$ matrix Z_i and the $n_i \times q_2$ matrix W_i. Both have a structure similar to X_i. There may be an overlap of the covariate effects in X_i, Z_i, and W_i. We use polynomial terms of time in W_i and Z_i to model class-specific and individual trajectories. The q_1 vector of random effects b_i is taken to be multinormally distributed with mean 0 and covariance D. The class-specific parameters are in the $q_2 \times K$ matrix M, where $M = (\mu_1, \ldots, \mu_K)$, with μ_k a q_2-dimensional column vector containing the parameters specific to class k. Given $c_{ik} = 1$, we have $Mc_i = \mu_k$ for $k = 2, \ldots, K$ and take $\mu_1 = 0$ to make the model identifiable. The error ε_i is an n_i vector that is uncorrelated with b_i and is multinormally distributed with mean 0 and covariance matrix $\sigma^2 \mathrm{I}_{n_i}$. Given the latent classes and random effects, the timings of covariates and y are assumed to be noninformative.

Model (2) captures common characteristics of biomarker trajectories within a subpopulation through latent classes while accommodating the variability among subjects in the same class through random effects. The use of a mixture of multivariate normal distributions for the longitudinal response y provides flexibility that allows nonnormal distributions for random effects.

Each class also has its own model for the event process. We use a frailty model that accommodates possibly time-dependent covariates, discontinuities in observation periods, and recurrent events. In counting process notation, the event process for subject i is written as $(N_i(t), Y_i(t))$, with $N_i(t)$ the number of events for subject i by time t and $Y_i(t)$ a left continuous at-risk process with $Y_i(t) = 1$ if subject i is at risk at time t and 0 otherwise. In the context of our NPC trials data described in the previous section, the counting process $N_i(t)$ for each subject i remains at 0 unless and until PCa is diagnosed in that subject, when it jumps to 1. We model the intensity process for subject i as

$$\alpha_i(t|c_{ik} = 1, \omega_i) = \omega_i Y_i(t)\lambda_k(t)\exp\{\gamma^{\mathrm{T}} x_i(t)\} \tag{3}$$

for $0 \le t \le \tau$, where τ is the end of the observation period for all subjects, $\lambda_k(t)$ describes the class-specific baseline hazard, and γ is a q_3 vector of coefficients for the covariates $x_i(t)$. The frailty distribution for ω_i, $g(\omega_i; \theta)$, is assumed to be gamma with mean 1 and variance θ. Model (3) can be viewed as a variation of a stratified model. Instead of stratifying on observable discrete covariates or combinations of the covariate values, we stratify on latent class membership.

3.2 Conditional Independence Assumption

The conditional independence (CI) assumption for the latent class joint model is summarized as

$$\begin{aligned}&\left[y_i, N_i, Y_i|\check{X}_i, c_{ik} = 1\right]\\ &\qquad = \left[y_i|\check{X}_i, c_{ik} = 1\right]\left[N_i, Y_i|\check{X}_i, c_{ik} = 1\right],\end{aligned} \tag{4}$$

where $\check{X}_i$ denotes the combined vector of covariates including v_i, X_i, Z_i, W_i, and x_i and $[\mathcal{A}|\mathcal{B}]$ denotes the conditional density of $\mathcal{A}$ given $\mathcal{B}$. In the foregoing model specification, the longitudinal response and time-to-event process are related only through latent classes. The motivation for the CI assumption comes from two factors: (1) The evolution of PSA and the PCa risk that characterize the disease process may be distinct across latent classes, and (2) once the disease risk class has been identified, diagnosis time should be independent of any surrogate biomarker. The shared random-effects model of DeGruttola and Tu (1994) and the latent process models of Henderson et al. (2000) also assume CI given the shared latent variables, except that their shared latent variables are continuous, which limits the form of association between the longitudinal and event process responses. CI is an important feature of the latent class model and greatly simplifies the modeling and estimation procedures. In Section 5.5 we discuss a method to check this assumption.

4. ESTIMATION

4.1 A Likelihood Approach for Latent Class Joint Model

We use semiparametric maximum likelihood methods to estimate the parameters in our model. For a fixed K, let Θ denote the combined parameter vector comprising $\eta, \beta, \sigma, \mu_1, \ldots, \mu_K, \mathrm{vec}(D)$ and γ, and let $\Lambda_k(t) = \int_0^t \lambda_k(s)\,ds$, $\Lambda = \{\Lambda_1, \ldots, \Lambda_K\}$, and $Y_i^{\gamma}(t) = Y_i(t)\exp\left\{\gamma^{\mathrm{T}} x_i(t)\right\}$. The log-likelihood of the observed data

$\{y_i, N_i, Y_i;\ i = 1, \ldots, n\}$ can be written as

$$l_n(\Theta, \theta, \Lambda; y, N, Y, \check{X}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \left[c_{ik} = 1 | \check{X}_i\right]\left[y_i | \check{X}_i, c_{ik} = 1\right] \times \left[N_i, Y_i | \check{X}_i, c_{ik} = 1\right]. \quad (5)$$

The first term within the inner summation of (5) is given by (1). Using the properties of multivariate normal (MVN) variables, the second term in (5) is

$$\left[y_i | \check{X}_i, c_{ik} = 1\right] = \mathrm{MVN}_{n_i}\left(X_i\beta + W_i\mu_k, Z_i D Z_i^{\mathrm{T}} + \sigma^2 \mathrm{I}_{n_i}\right). \quad (6)$$

Using the product-integral notation $\prod$, the third term can be expressed as

$$\int_0^\infty \left[N_i, Y_i | \check{X}_i, \omega_i, c_{ik} = 1\right] g(\omega_i; \theta)\, \mathrm{d}w_i = \frac{\prod_{t\in[0,\tau]} \left\{\left(1 + \theta N_i(t-)\right) Y_i^{\gamma}(t) \lambda_k(t)\right\}^{\Delta N_i(t)}}{\left\{1 + \theta \int_0^\tau Y_i^{\gamma}(t)\, \mathrm{d}\Lambda_k(t)\right\}^{1/\theta + N_i(\tau)}}. \quad (7)$$

The semiparametric maximum likelihood estimator (SPMLE), denoted by $(\hat{\Theta}, \hat{\theta}, \hat{\Lambda})$, is defined as the maximizer of (5). For a given number K of latent classes, under regularity conditions, the identifiability, existence, consistency, and weak convergence of the SPMLE $(\hat{\Theta}, \hat{\theta}, \hat{\Lambda})$ can be established by arguments similar to those of Parner (1998) (for details, see Lin 2000).

4.2 The EM Algorithm

Direct maximization of (5) is difficult with the sum over latent classes within the logarithm and with potentially hundreds of parameters, mainly arising from the $\Delta\Lambda_k$, the changes of Λ_k at the observed event times for each class k. Viewing the class membership, random effects, and frailties as missing data, we use a version of the EM algorithm to obtain parameter estimates.

The complete-data log-likelihood l_c based on the complete data (y, b, c, N, Y, ω) is given by

$$\sum_{i=1}^{n} \left\{\log[c_i | \check{X}_i] + \log[b_i] + \log[y_i | b_i, c_i, \check{X}_i] + \log[\omega_i] + \log[N_i, Y_i | \omega_i, c_i, \check{X}_i]\right\}. \quad (8)$$

Expansion of each term in (8) is straightforward and thus is omitted here. Let $\tilde{a} = \mathrm{E}(a | y, N, Y)$ denote expectation of the random variable a conditional on y, N, and Y. The conditional expectation for the terms involving c_i, b_i, and ω_i given the observed data are calculated in the E step as in Appendix A.1. Inserting the conditional expectations calculated in (A.3a)–(A.3c) and (A.4a) and (A.4b) into (8), the parameters β, σ^2, μ_k, D, η_k, and $\Delta\Lambda_k$ are updated in the M step as follows:

$$\hat{D} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{bb}_i,$$

$$\hat{\beta} = \left(\sum_{i=1}^{n} X_i^{\mathrm{T}} X_i\right)^{-1} \sum_{i=1}^{n} X_i^{\mathrm{T}}(y_i - Z_i\tilde{b}_i - W_i M \tilde{c}_i),$$

$$\hat{\sigma}^2 = \frac{1}{\sum_{i=1}^{n} n_i} \sum_{i=1}^{n} \left\{(y_i - X_i\hat{\beta})^{\mathrm{T}}(y_i - X_i\hat{\beta}) - 2Z_i\tilde{b}_i - 2W_i M\tilde{c}_i) + 2\,\mathrm{trace}(Z_i^{\mathrm{T}} W_i M \widetilde{bc}_i^{\mathrm{T}}) + \mathrm{trace}\left\{M^{\mathrm{T}} W_i^{\mathrm{T}} W_i M \mathrm{diag}(\tilde{c}_i)\right\} + b\widetilde{ZZ}b_i\right\},$$

$$\hat{\mu}_k = \left(\sum_{i=1}^{n} W_i^{\mathrm{T}} W_i\right)^{-1} \sum_{i=1}^{n} W_i^{\mathrm{T}}\left\{(y_i - X_i\beta)\tilde{c}_{ik} - Z_i\widetilde{b_i c_{ik}}\right\},$$

and

$$\Delta\hat{\Lambda}_k(t_j) = \frac{\sum_{i=1}^{n} \tilde{c}_{ik}\Delta N_i(t_j)}{\sum_{i=1}^{n} \tilde{c}_{ik}\tilde{\omega}_i Y_i^{\gamma}(t_j)} = \frac{\tilde{c}_{(j)k}}{\sum_{i=1}^{n} \tilde{c}_{ik}\tilde{\omega}_i Y_i^{\gamma}(t_j)} \equiv \hat{\lambda}_k(t_j). \quad (9)$$

Here $\tilde{c}_{(j)k}$ denotes the $\tilde{c}_{ik}$ for a subject i who experiences the event (possibly recurrent) at time t_j. Note that (9) is a weighted version of the usual Breslow (1972) estimator of the baseline hazard.

The updates of η_k, γ, and θ have no closed-form expression. Let superscript (j) denote the jth iteration *within* each M step; we use Newton–Raphson iterations until a tolerance on the relative change of the parameters within each M step is achieved:

$$\eta_k^{(j+1)} = \eta_k^{(j)} + \mathrm{I}^{-1}(\eta_k^{(j)})\mathrm{S}(\eta_k^{(j)}), \quad (10a)$$

$$\frac{1}{\theta^{(j+1)}} = \frac{1}{\theta^{(j)}} + \mathrm{I}^{-1}\left(\frac{1}{\theta^{(j)}}\right)\mathrm{S}\left(\frac{1}{\theta^{(j)}}\right), \quad (10b)$$

and

$$\gamma^{(j+1)} = \gamma^{(j)} + \mathrm{I}^{-1}(\gamma^{(j)})\mathrm{S}(\gamma^{(j)}), \quad (10c)$$

where $\mathrm{S}(\cdot)$ represents the score and $\mathrm{I}(\cdot)$ the information, which are given in Appendix A.2. A useful result is that the maximum likelihood estimator (MLE) for γ given in (10c) is the same as the maximum partial likelihood estimator (MPLE); see Appendix A.2.

This EM algorithm depends on a predetermined value of K; in practice, K will be unknown. We propose that the model be fit separately for several values of K in a plausible range. The choice of K for the PSA data is discussed in Section 5.2. For each K, we start the algorithm from multiple initial values and in each case run to convergence, which is determined by tolerances on consecutive parameter values. To obtain the initial parameter values, we randomly assign class memberships to the subjects, then fit a multinomial logistic regression to obtain η, a linear mixed-effects model, to get β, M, D, and σ^2, and a frailty model stratified on class to get the initial values for λ_k, γ, and θ. Other starting values are obtained by perturbing various of these initial parameters. In our case, all different starting points that we tried resulted in the same set of parameter estimates for each $K = 1, \ldots, 6$.

4.3 Approximation of the Information Matrix

By using an argument similar to that of Parner (1998), it can be shown that the inverse of the observed information matrix is a consistent estimator of the covariance process of the parameters evaluated at the observed failure times. We calculate the standard errors of the parameters through the approximation of the inverse of the observed information matrix. Let s denote the observed score vector and let $\Psi = (\Theta, \theta, \Lambda)$. The observed information matrix can be approximated by the observed empirical information matrix denoted by $\mathrm{I}_e(\Psi; y, N, Y, \breve{X})$,

$$\sum_{i=1}^{n} s(y_i, N_i, Y_i, \breve{X}_i; \Psi) s^{\mathrm{T}}(y_i, N_i, Y_i, \breve{X}_i; \Psi) - n^{-1}\left\{\sum_{i=1}^{n} s(y_i, N_i, Y_i, \breve{X}_i; \Psi)\right\}\left\{\sum_{i=1}^{n} s^{\mathrm{T}}(y_i, N_i, Y_i, \breve{X}_i; \Psi)\right\}.$$

The observed score equals the conditional expectation of the complete-data score (McLachlan and Krishnan 1997, result 3.42). Therefore, on evaluation at $\Psi = \hat{\Psi}$, the second term of $I_e(\hat{\Psi}; y, N, Y, \breve{X})$ is 0 and

$$s(y_i, N_i, Y_i, \breve{X}_i; \hat{\Psi}) = \mathrm{E}_{\Psi}\left[\partial\{\log L_c(y_i, N_i, Y_i; \Psi)\}/\partial\Psi | y_i, N_i, Y_i\right]|_{\Psi=\hat{\Psi}}.$$

4.4 Predicting Event Probability With Longitudinal Measurements to Date

Let T_i denote the time of PCa diagnosis for subject i. The following calculation allows us to predict PCa probability given the longitudinal measurements y_i and covariates $\breve{X}_i$ up to time t, denoted by y_i^t and $\breve{X}_i^t$:

$$P(T_i \le t | y_i^t, \breve{X}_i^t) = \sum_{k=1}^{K} P(T_i \le t | c_{ik} = 1, \breve{X}_i^t) P(c_{ik} = 1 | y_i^t, \breve{X}_i^t), \quad (11)$$

where $P(T_i \le t | c_{ik} = 1)$ can be obtained using model (3) with the frailty ω_i integrated out and $P(c_{ik} = 1 | y_i^t)$ is calculated through the Bayes formula. These calculations are straightforward and thus are omitted. In practice, (11) is computed using estimated values for the parameters.

5. FITTING THE LATENT CLASS JOINT MODEL TO THE NUTRITIONAL PREVENTION OF CANCER TRIALS DATA

5.1 Variables Used in the Model

We include the following in our model: The covariate vector v used to predict class membership in (1) contains an intercept, the treatment assignment indicator of Se supplementation group, and age, PSA, and Se level at randomization. The biomarker value y in (2) is the vector of longitudinal PSA readings transformed as log(PSA+1). The transformation log(PSA + 1) was also used by Pearson, Morrell, Landis, Carter, and Brant (1994), Whittemore et al. (1995), Slate and Clark (1999), Slate and Turnbull (2000), and Lin et al. (2000) and implies growth on the log scale, which might be justified both biologically, by the exponential growth of malignancies, and statistically, as an approximate variance stabilizing transformation. The addition of one is to diminish the influence of extremely small PSA readings. The fixed-effects covariate vector X in (2) contains the treatment assignment indicator, both age and Se level at randomization, and linear and quadratic terms of visit time expressed in years since entry into the trial. The covariates for the random effects and class-specific effects in (2), Z and W, both contain an intercept and linear and quadratic terms of years since entry. The survival time T was calculated from the date of randomization for each subject until the date of censoring due to death from unrelated cause, the date of PCa diagnosis, if any, or the end of the observation time (July 25, 1997). The covariate vector $x(t)$ in (3) comprises the indicator of Se supplementation group and both age and Se level at randomization. Here, in fact, the covariates in $x(t)$ are time independent. Thus we use a proportional hazards model for the event process submodel in (3), and verify the qualitative results using a more general nonproportional hazards model described in Section 6.

We use all potentially useful information from predictors available in all three pieces of the model. Baseline PSA is, of course, an important predictor; it is also known that the probability of PCa increases with age. Information on Se is also useful to help in determining its role in cancer prevention. Baseline PSA is not used in the longitudinal submodel (2), because, it is treated as y at time 0, and also is not used in the survival submodel (3), we assume, because that the dependence of time to PCa diagnosis on PSA is induced only through latent classes. Likelihood ratio tests of the effect of baseline PSA on time to event are performed in Section 6. The effect of a predictor in a piece of the model can be considered to be adjusted for the effects of the predictor in other pieces of the model.

5.2 Fitting the Model

Latent class models are special cases of finite mixture models; if K is not known in advance, then the model is nonstandard (Titterington, Smith, and Makov 1998). In fitting such models, the number of component densities may be drastically overestimated when there is a lack of model fit (Leroux 1992). This can have serious consequences and can encourage substantive interpretations for spurious subgroups. Nonetheless, the possibility of underestimating K is less likely in theory (Chen and Kalbfleisch 1996). For this reason, as well as for parsimony, we seek the solution with the fewest classes (smallest K) that still provides a satisfactory fit to the data. Chen and Kalbfleisch proved that the number of components in the finite mixture model can be consistently estimated using a class of minimum distance estimators in which the fit is penalized with small mixing weights. They found that their method gives values of K similar to those found using the Bayesian information criterion (BIC) (Schwartz 1978). Also, BIC gives a relatively larger penalty on the number of parameters used than other criteria, such as the Akaike information criterion (AIC). We thus adopt BIC to guide selection of K.

Using BIC reveals that the four-class model is preferred (Table 1). This also agrees with visual inspection of Figure 1,

Table 1. Log-Likelihood and BIC by Number of Latent Classes, K

K	Maximized log $L^{(K)}$	BIC[a]	Deviance change[b]	No. of parameters, d_K
1	−2239	−2592	1,845	100
2	−1971	−2583	1,310	173
3	−1643	−2482	654	237
4	−1449	−2440	270	280
5	−1347	−2450	61	312
6	−1316	−2452	0	321

[a] BIC = $\log L^{(K)} - (1/2)\, d_K \log n$, where d_K is the number of parameters in the fitted model.
[b] Deviances are relative to the six-class model.

which contains the fitted PSA trajectories and survival curves for each class in the K-class solution and for each of $K = 1, \ldots, 6$ and shows fairly distinct changes as the classes are increased, up to four classes. From clinical practice, we might have expected two latent groups corresponding to high and low risk of PCa and a cutoff of 4 ng/ml for PSA. However, the latent class model suggests a more refined separation of the risk groups for PCa.

The four-class solution identifies fitted PSA trajectory classes that we label as "low I," "low II," "medium," and "high" (Fig. 2 (a); see the caption for the calculation of these fits). The majority classes "low I" and "low II" (Table 2) are characterized by a consistently low PSA level throughout the trial period. The "medium" class has a higher PSA level than the two "low" classes throughout the trial; the PSA level increases over time for this class. The minority class "high" has the highest PSA level at the beginning of the trial; the predicted level of PSA increases over time until the fourth year after randomization and then decreases. This quadratic trajectory may be partially explained by the tendency for subjects with high PSA to be removed because of the diagnosis of PCa and the PSA-decreasing effects of medications such as Proscar. To see how the fitted trajectories relate to the observed data, we calculate an "observed" trajectory for each class as the average on a 6-month interval weighted by the estimated class k probability $\hat{\pi}_{ik}$ for subject i (see Fig. 2 for a detailed explanation). The values of $\hat{\pi}_{ik}$ are obtained by inserting the covariate v_i and estimated parameter $\hat{\eta}_k$ from Table 3 into (1). It is seen that the observed trajectories agree well with the fits.

The class-specific survival probability can be calculated by integrating out the frailty ω as

$$S_i(t|c_{ik} = 1) = \left\{1 + \theta Y_i^{\gamma} \Lambda_k(t)\right\}^{-1/\theta}. \tag{12}$$

The four-class solution also identifies four fitted survival curves [Fig. 2 (d)] that assign decreasing estimated survival probabilities to the four PSA trajectories. Here the fitted curve is defined as the estimated survival curve averaged across the two treatment groups. We may also define observed survival curves obtained by a conventional Kaplan–Meier plot using $\hat{\pi}_{ik}$ as a case weight for subject i. Figure 2 (d), (e), (g), and (h) show that the fitted and observed survival curves across the four classes also agree well. We also plot the fitted and observed survival curves by the treatment group for each of the four classes in Figure 2 (g) and (h); the Se treatment group has a higher survival probability than the placebo group at all times.

To further explore the class structure, we examine the estimated proportion of subjects in class k, given by $\hat{p}_k = (1/n)\sum_{i=1}^{n} \hat{\pi}_{ik}$ (see the first row in Table 2). The estimated proportions of subjects in the two "low" classes are fairly large. The maximal values across subjects (see the second row of Table 2) of $\hat{\pi}_{ik}$ for each class k are extremely close to one, implying that each class has some unambiguously assigned subjects. The estimated or "fitted" PCa incidence for class k is calculated as

$$\frac{\sum_{i=1}^{n} \hat{\pi}_{ik}\left\{1 - \hat{S}(\tau|c_{ik} = 1)\right\}}{\sum_{i=1}^{n} \hat{\pi}_{ik}},$$

where $\hat{S}(\tau|c_{ik} = 1)$ is the estimate of $S(\tau|c_{ik} = 1)$ given in (12) with $t = \tau$ and parameters evaluated at their estimates. The "observed" PCa incidence for class k is calculated as $\sum_{i=1}^{n} \hat{\pi}_{ik} N_i(\tau) / \sum_{i=1}^{n} \hat{\pi}_{ik}$. It can be seen that the estimated and observed probabilities closely agree and that there is a monotone increasing relationship with class. The "low I" and "low II" classes have the lowest estimated probability of developing PCa (.7% and 4%). The estimated probability of developing prostate cancer in the "medium" class (19%) is between that in the "low" and "high" classes. More than 70% of the members in "high" class will develop prostate cancer. The "estimated" marginal PCa incidence is 6.7%, which was calculated by the average of the estimated class incidences weighted by estimated class proportions. This is very close to the observed PCa incidence in our data. The latent class approach allows us to separate groups of different size and risk for PCa, and thus we are able to identify the group with extremely high risk in the situation of overall low incidence. Tables 2, 3, and 4 report detailed calculations and the parameter estimates for the four-class model.

5.3 Characterization of Covariate Effects and the Latent Classes

We use the deviance change, (i.e., twice the difference in the maximized log-likelihood between models with and without inclusion of the effect parameters to be assessed) to calculate the significance of a given effect for our latent class models.

For the four-class solution, the upper part of Table 3 reports on those predictors used for the class membership model (1). It reveals that PSA and Se at entry are significant, with estimated coefficients in monotone agreement with the group ordering. Note that by exponentiating the coefficient estimates in the upper part of Table 3, we may obtain estimates of odds ratios for unit increases in the variables for each designated class relative to those for the baseline class, "low I." Testing for the fixed effects for the biomarker trajectory in model (2), Table 4 shows that the intercept and time since entry are significant. Baseline age is also significant; older men have higher PSA levels. To test the class-specific effects of the intercept and time on PSA trajectory, we used a reduced model in which we set all μ_k equal to 0 for $k = 1, \ldots, K$ and calculated the change in deviance of the reduced model as compared with the original model. The deviance changes in Table 3 for the longitudinal submodel reveal that the PSA level at entry is quite different, as are the patterns of PSA readings over time among the four classes. The estimated intercepts in $\hat{\mu}_k$ for

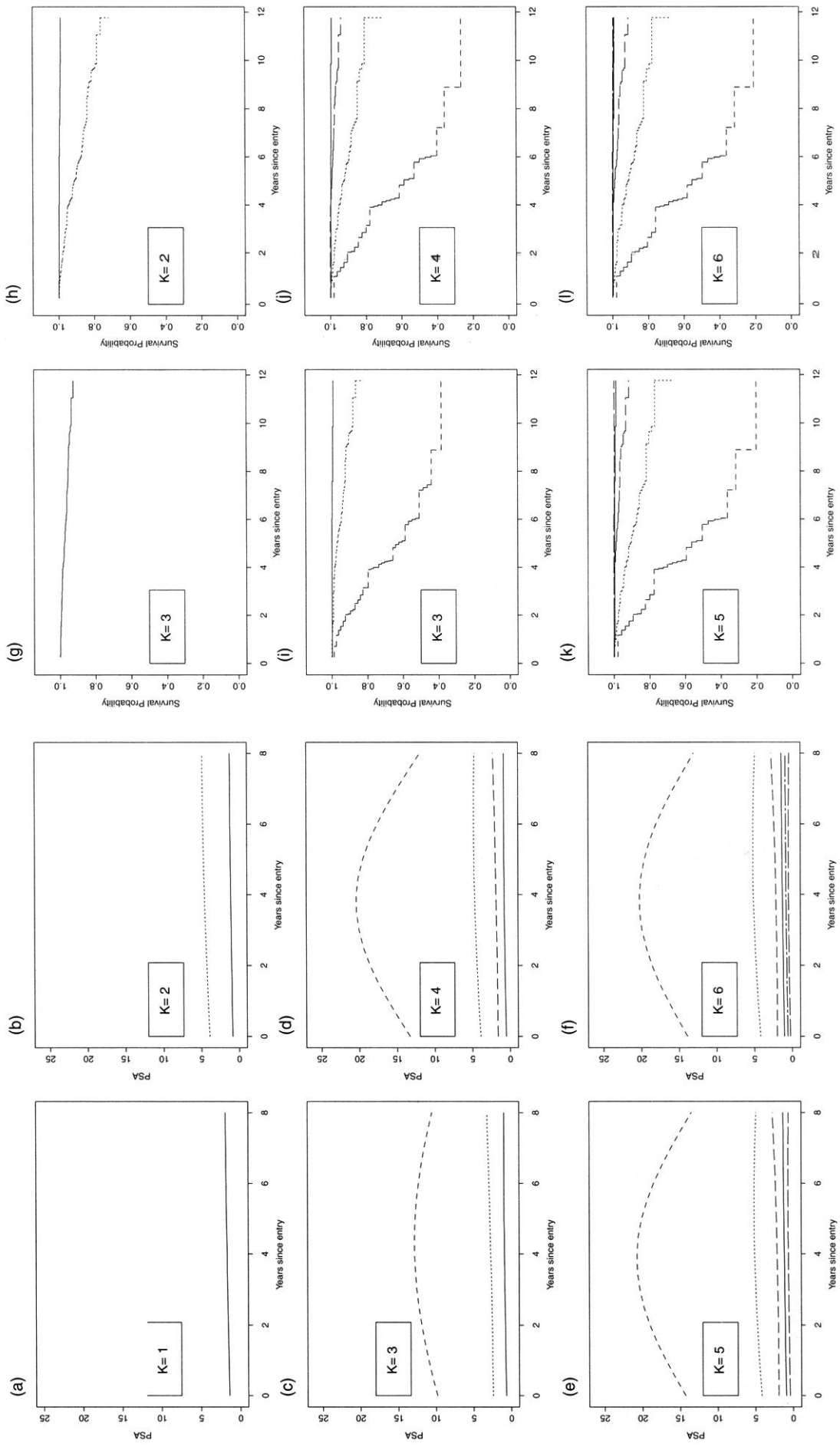

Figure 1. Fitted PSA Trajectories [(a)–(f)] and Survival Curves [(g)–(l)]. ——, Low I; - - - High; —— - Low III; . . . . . Medium; — – Low II; — — - Low IV.

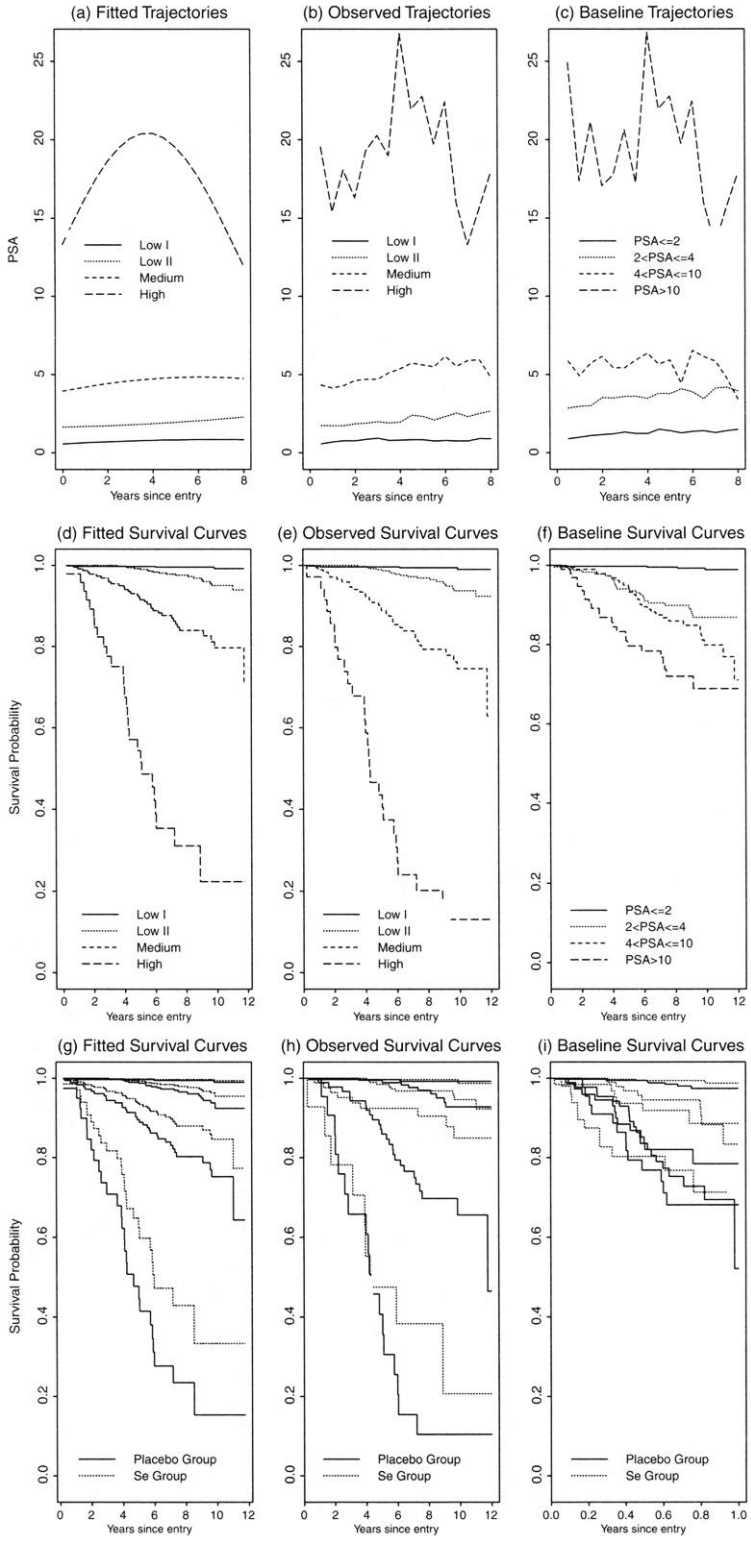

Figure 2. PSA Trajectories and Survival Curves for the Four-Class Model. The fitted values $\hat{Y}$ in (a) are calculated at 6 month intervals for a given class k using the formula $\log(1+\widehat{PSA}) = X\hat{\beta} + W\hat{\mu}_k$ with average covariate values in X and W. Fitted PSA values are transformed back to the original scale for plotting. The "observed" PSA value for class k on a 6 month interval is an average of the PSA readings available on this interval weighted by the probabilities of class k membership, $\hat{\pi}_{ik}$ (in (b)). The "baseline" PSA trajectories in (c) are obtained by dividing the subjects into four groups according to the baseline PSA values with the cutoff values shown within the panel. The fitted survival curve in (d) for class k is calculated as $(1 - prop_k(Se))\hat{S}_k^p(t) + prop_k(Se)\hat{S}_k^{Se}(t)$, where $prop_k(Se)$ is the estimated proportion of subjects in the Se treatment group for class k, and $\hat{S}_k^p$ and $\hat{S}_k^{Se}$ are the class k–specific survival probability for placebo and Se groups. The four latent classes "low I", "low II", "medium," and "high" in (g), (h), and (i) are in decreasing order of survival probability within placebo (solid lines) and Se groups (dashed lines). The observed survival curves are obtained from the S-PLUS routine survfit using, all the values of $\hat{\pi}_{ik}$ as case weight for subject i. The baseline survival curves in (c), (f), and (i) are obtained by dividing the subjects into four groups according to the baseline PSA values with the cutoff values shown.

Table 2. Class-Specific Information

	Latent class k			
	Low I	*Low II*	*Medium*	*High*
Estimated class proportions; estimated and observed incidence of PCa by classes				
Class proportion $\hat{p}_k$	.47	.35	.15	.03
$\max_i\{\pi_{ik}\}$	1.00	.99	1.00	.98
Estimated PCa incidence	.01	.04	.19	.70
"Observed" PCa incidence	.01	.04	.19	.76
Average covariate values by class (standard errors in parentheses)				
Baseline variable*				
Proportion in Se group	.49 (.01)	.50 (.02)	.48 (.03)	.39 (.05)
Age (years)	62.12 (.23)	64.90 (.37)	67.92 (.39)	69.87 (.78)
PSA (ng/ml)	.57 (.01)	1.72 (.01)	4.47 (.02)	18.54 (.07)
Se (ng/ml)	115.44 (.44)	114.95 (.72)	115.00 (1.13)	114.28 (2.31)

* The averages of covariates for a given class k are calculated using the formula $\sum_{i=1}^{n} \hat{\pi}_{ik} v_i / \sum_{i=1}^{n} \hat{\pi}_{ik}$. The standard errors are approximated by the delta method.

the "low I," "low II," "medium," and "high" classes are .42, .94, 1.56, and 2.63, on the log scale (see Table 3), with an estimated within-class standard deviation of .22 ($\hat{D}_{11}^{1/2}$). Table 4 also reveals that Se supplementation significantly reduces the risk of PCa. The estimated variance, $\hat{\theta}$, of the frailty is quite small, indicating that the latent classes account for most of the variability in the onset time distributions. We also fitted a reduced model in which we set all $\lambda_k(t)$ equal to $\lambda_0(t)$ for all $k = 1, \ldots, K$. For all $0 \le t \le \tau$, the resulting deviance change from the original model is 301.96 on 156 df, indicating a significant class-specific difference in hazard estimates (see Table 3).

We now describe the latent classes themselves through covariate values. The lower part of Table 2 reveals some interesting features for covariates in each class. First, the "high" class has a remarkably low proportion in the Se supplementation group (39%), whereas the percentage of subjects in the Se group is about 49% in the study population. This means that a "high" class member is less likely to be in the Se supplementation group. Second, there is a strong monotone relation between PSA at entry and latent class. Third, subjects in the "medium" and "high" classes have higher average age at entry than those in the two "low" classes, which is in agreement with the observation that the incidence of PCa increases with age.

5.4 Performance of the Estimation Procedure

We evaluated the performance of our estimation procedure through a simulation study based on our results for the four-class solution for the NPC trials data. For each run of the simulation, we created a pseudodataset by generating PSA values (for all y_i), survival times (for all T_i), and censoring indicators from the four-class model with parameter values given by the MLEs from the NPC PSA analysis, except that "treatment indicator" was the only predictor in the frailty submodel (3). We then fit this same four-class model to the pseudodata using the EM algorithm described in Section 4. We used the true parameter values as starting points (rather than multiple starting points), and only 5 of 105 EM runs did not converge. The average values of the estimates obtained for 100 sets of pseudodata varied by less than 4% from the parameter values used to generate the pseudodata, and the sample standard errors were comparable to those resulting from the empirical information given in Section 4.3.

Table 3. Estimated Class-Specific Coefficients η and M

	Latent class			*Deviance change*	*Degrees of freedom*
	Low II	*Medium*	*High*		
Latent class submodel					
η					
Intercept	−20.51 (4.62)	−35.34 (4.64)	−50.76 (5.19)	12,181	3
Treatment indicator	−.53 (.70)	−.95 (.87)	−1.37 (1.19)	5	3
Baseline age	−.59 (.36)	−.63 (.46)	.39 (.93)	4	3
Baseline PSA	29.72 (6.80)	4.84 (6.70)	47.51 (6.78)	6770	3
Baseline Se	−.66 (.39)	−1.00 (.48)	−1.05 (.58)	6	3
Longitudinal submodel					
M matrix time trend					
Intercept	.523 (.029)	1.139 (.031)	2.211 (.067)	4008	3
Linear	−.037 (.013)	.006 (.014)	.166 (.030)	22	3
Quadratic	.006 (.002)	−.001 (.002)	−.024 (.005)	116	3

NOTE: Standard errors are in parentheses. Low I is fitted as the baseline class, and the reported coefficients are deviations from that class.

Table 4. Estimates for the Class-Independent Coefficients, β, σ^2, γ, and θ

Variable		Estimate	Standard error	Deviance change*
Longitudinal submodel				
$\hat{\beta}$	Intercept	.419	.023	842
	Treatment Indicator	.012	.017	1
	Baseline Age	.031	.010	11
	Time trend			
	Linear	.049	.010	172
	Quadratic	−.004	.001	167
	Baseline Se	.023	.014	2
$\hat{\sigma}^2$	Error variance	.053	4.74e−7	
Survival submodel				
$\hat{\gamma}$	Treatment Indicator	−.54	.23	5
	Baseline Age	.12	.18	0
	Baseline Se	.03	.11	0
$\hat{\theta}$	Frailty variance	3.0e−4	3.4e−4	0

*Based on a single degree of freedom for all of the covariates listed.

Table 5. Classification of the 1,182 Subjects by Baseline PSA, the Latent Class Longitudinal Only and the Joint Models as Described in Section 5.5

	Baseline PSA, ng/ml			
	<2	2–4	4–10	>10
	K = 4 Joint model			
"Low I"	560	0	0	0
"Low II"	304	106	0	3
"Medium"	1	18	83	2
"High"	0	1	11	23
	Longitudinal only model			
	Low I	*Low II*	*Medium*	*High*
	K = 4 Joint model			
"Low I"	558	2	0	0
"Low II"	88	324	1	0
"Medium"	0	27	147	0
"High"	0	0	10	24
	*Longitudinal only model**			
	I	*II*	*III*	*IV*	*V*
	*K = 5 Joint model**				
"I"	347	8	1	0	0
"II"	3	333	0	0	0
"III"	37	0	260	0	0
"IV"	0	0	31	130	0
"V"	0	0	0	9	23

* Classes are ordered by PCa risk.

5.5 Comparison With Simpler Analyses

At the suggestion of the referees, we consider some simpler analyses. First, we divided the subjects into four groups according to their baseline PSA values: <2, 2–4, 4–10, and >10 ng/ml. These intervals have clinical significance; in particular, a PSA value of 4 ng/ml is often used to trigger additional testing, whereas a value of 10 ng/ml or more is considered high. Within each of these groups, we separately calculated a longitudinal trajectory and a survival curve. These are plotted in Figure 2, with PSA in (c), survival in (f), and survival by both group and treatment in (i). It can be seen that the four PSA trajectories discovered by the latent class models track those of the four empirical groups very well, however, the four (or eight) empirical survival curves in Figure 2 (c), (f), and (i) do not separate nearly as well as those in either the fitted or "observed" curves in the other parts of the figure. Using latent classes thus provides better discrimination of the patterns of survival response. Dividing subjects into finer groups by using more of the observed covariates might be more successful at discrimination but would be subject to problematic choices of which covariates and which cutoff points to use. The latent class model avoids such complications.

For the latent class models, we classified subjects into four classes according to the maximum posterior class probabilities for subject i, $\max_{k=1,\ldots,K}(\tilde{c}_{ik})$ and compared the results to the classification derived from using the baseline PSA values with the cutoffs of 2, 4, and 10 ng/ml. Looking at the off-diagonal elements of the upper part of Table 5 reveals that the classification using only the baseline PSA information generally underestimates the risk of PCa compared with our latent class model. We also fit a latent class longitudinal data–only model with (1) and (2), not using the survival information in (N, Y). We then classified the subjects into different classes according to $\max_{k=1,\ldots,K}(\tilde{c}^y_{ik})$ resulting from this model. The middle part of Table 5 compares the four group classification according to $\max_{k=1,\ldots,K}(\tilde{c}^y_{ik})$, where only the longitudinal information is utilized with those according to $\max_{k=1,\ldots,K}(\tilde{c}_{ik})$, where both longitudinal and survival information are used. The classification using only the longitudinal PSA information seems to underestimate the risk of PCa. Also, for the longitudinal data–only model, BIC suggests $K = 5$ instead of $K = 4$, in which case the extent of underestimation is lessened. Classifying the subjects according to the survival times only resulted in no obvious pattern in PCa risk.

Another, simpler alternative to our model arises from treating the PCa diagnosis as a binary rather than a survival outcome in the submodel (3). We also have fit a $K = 4$ model using the same covariates for the binary submodel as we used in the survival portion of the model here; both models gave qualitatively similar results. In a previous article (Lin et al. 2000), we fitted a latent class model using incidence (presence/absence) of PCa diagnosis within 7 years. A binary sub model with intercept only (no covariates) was fitted; $K = 4$ was the choice using BIC. There a subset of the data comprising 1090 of the male participants with sufficient follow-up was used; PSA readings more than 7 years postrandomization were not included, and the men who were diagnosed with PCa after 7 years were treated as PCa free. However, using the latent class model with the survival outcome allows us to make more efficient use of the data and to calculate the PCa probability given PSA to any date within the entire observation period using (11); inference from the binary model is limited to a prespecified date.

We have computed receiver operating characteristic (ROC) curves (Fig. 3) using $P(T_i \le t | y_i^t, \breve{X}_i^t)$ in (11) with longitudinal PSA information incorporated and compared with using a single PSA measurement at randomization to predict PCa

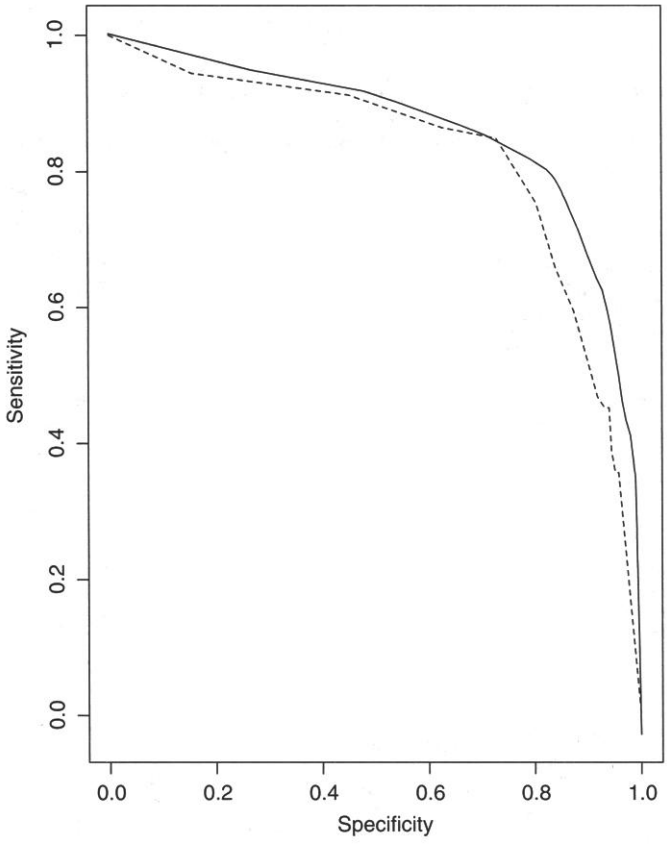

Figure 3. ROC Curves. The ROC curve based on the latent class joint model (solid line) was generated according to $P(T_i \le t|y_i^t, \breve{X}_i)$ ($t = 7$ years) given in (11) with parameter estimates from the four-class solution. The ROC curve based on a single baseline PSA measurement (dashed line) was generated by calculating sensitivity and specificity for a PSA cutoff value.

probability at 7 years. Figure 3 shows that incorporating the longitudinal information improves specificity and sensitivity for PCa diagnosis, although in our example a single PSA measurement also works fairly well. More details are available from the first author.

6. ASSESSING THE ASSUMPTIONS AND FIT

We assess the proportional hazards (PH) assumption in model (3) by using $x(t) = (x', x' \cdot \log(t))'$ and $x(t) = (x', x' \cdot t)$ in our latent class joint model. Likelihood ratio tests for the coefficients associated with $x' \cdot \log(t)$ and $x' \cdot t$ are both nonsignificant. Therefore, the PH assumption seems to be met under the conditional independence of longitudinal PSA and PCa onset. More specifically, Figure 2 (h) might suggest that a model including $x_{\text{trt}}(t) = (x_{\text{trt}}, x_{\text{trt}} \cdot \log(t))$, where x_{trt} is the Se treatment group indicator, would be more appropriate. We also fit the model with x_{trt} and $x_{\text{trt}} \cdot \log(t)$ included together with baseline age and Se level in (3) to the data; however, there is a statistically nonsignificant improvement in fit over the model without $x_{\text{trt}} \cdot \log(t)$, and qualitative results remain unchanged.

A referee suggested that including baseline PSA in the survival model may give a way to check for association between PSA and PCa over and above class membership. The $\chi^2_{df=1}$ statistics for baseline PSA are 40.7, 5.6, 3.3, 2.3, .7, and .1 for $K = 1, \ldots, 6$. So for $K \ge 3$, there is no significant residual effect of baseline PSA on time to event over and above its effect in latent classes, and the contribution of baseline PSA seems to be useful mostly in predicting class membership. We next describe a technique for assessing the CI assumption. Let $\hat{\Psi}^{(n)}$ denote the MLE of the parameters Ψ based on n subjects and let

$$\tilde{c}_{ik}^{(n)} \equiv P\big(c_{ik} = 1 | y_i, N_i, Y_i, \breve{X}_i; \hat{\Psi}^{(n)}\big) \tag{13}$$

be the estimated conditional probability that subject i is in class k. Let $C_i^{(n)} = (C_{i1}^{(n)}, \ldots, C_{iK}^{(n)})$ denote a multinomial vector with probabilistic distribution given by $\tilde{c}_i^{(n)} = \{\tilde{c}_{i1}^{(n)}, \ldots, \tilde{c}_{iK}^{(n)}\}$ in (13). Then, using an approach similar to theorem 2 of Bandeen-Roche, Miglioretti, Zeger, and Rathouz (1997), we have $\lim_{n\to\infty} P(C_{ik}^{(n)} = 1) = \pi_{ik}$ and

$$\lim_{n\to\infty} \big[y_i, N_i, Y_i | C_{ik}^{(n)} = 1, \breve{X}_i\big] = \big[y_i, N_i, Y_i | c_{ik} = 1, \breve{X}_i\big] \tag{14}$$

(see Lin 2000, chap. 6 for more details). This result states that the data classified to class k a posteriori according to (13) converge in distribution to the data generated from class k as if the class indicators $\{c_i\}$ were observable. The $[y_i, N_i, Y_i | C_{ik}^{(n)} = 1, \breve{X}_i]$ in (14) can be written as

$$\big[N_i, Y_i | C_{ik}^{(n)} = 1, y_i, \breve{X}_i\big]\big[y_i | C_{ik}^{(n)} = 1, \breve{X}_i\big].$$

If the CI assumption holds, then we would not expect to observe dependence of $\{N, Y\}$ on y given $\tilde{c}_i^{(n)}$ for large enough n.

To implement the foregoing idea, for a given K, we check whether the event process $\{N, Y\}$ depends on the longitudinal marker process $\{y\}$ within class k by fitting a frailty model treating y as a time-varying covariate,

$$\begin{aligned}\alpha_i(t | C_{ik}^{(n)} = 1, \omega_i, y_i) &\\ = \omega_i Y_i(t) \lambda_k(t) \exp\big\{\gamma_k^y \hat{\mathrm{E}}_k\{y_i(t)\} + \gamma^{\mathrm{T}} x_i(t)\big\},&\end{aligned} \tag{15}$$

where $\hat{\mathrm{E}}_k\{y_i(t)\} = X_i \hat{\beta}_k + W_i \hat{M}_k \tilde{c}_i$, with $\hat{A}_k$ denoting the MLE of A_k for a latent class joint model with K classes and γ_k^y the class-specific coefficients associated with $\hat{\mathrm{E}}_k\{y_i(t)\}$. We use $\hat{\mathrm{E}}_k\{y_i(t)\}$ because y is not observed for all event times. Because we do not directly observe $C_{ik}^{(n)} = 1$, we therefore use the estimated $\tilde{c}_{ik}^{(n)}$ as a case weight in (15) with class-specific hazard λ_k. The $\tilde{c}_{ik}^{(n)}$ can be estimated as $\tilde{c}_{ik}$ calculated in the final E step in (A.2). If the CI assumption is met, then we would expect γ_k^y to be nonsignificant. For the NPC PSA data, fitting model (15) with $\gamma_k^y = \gamma^y$ and with all γ_k^y different resulted in little difference in log-likelihood for all k. We therefore fit model (15) with $\gamma_k^y = \gamma^y$, and the likelihood $\chi^2_{df=1}$ statistics for testing $\gamma^y = 0$ are 265.91, 37.02, 3.81, .67, .76,

and .20, corresponding to $K = 1$–6 with the p values calculated as 0, 0, .05, .50, .38, and .65. It is seen that the PCa onset time depends on the longitudinal biomarker PSA within each class less and less as the number of classes, fit increases. For models with one and two classes, the CI assumption is not met. For the model with three classes, the results are marginal, whereas for models four or more classes, the CI assumption appears adequate.

To assess the fit for the latent class joint model, marginal predicted values can be obtained for the observed responses y and $\{N, Y\}$. The predicted values of y_i [fitted log(PSA + 1)] are calculated as an estimated weighted average according to $[y_i|\check{X}_i] = \sum_{k=1}^{K} \pi_{ik}\ N_{n_i}(X_i\beta + W_i\mu_k, Z_iDZ_i^{\mathrm{T}} + \sigma^2 \mathrm{I}_{n_i})$. The residuals of log(PSA + 1) versus predicted values of y for the four-class model are much smaller in absolute value than those of the one-class model (figure not shown here). The improved fit of the latent class model over the usual linear mixed model (one-class model) is because of the reallocation of some of the residual variation into the variation from the class-specific fit.

Subject i–specific martingale residuals can be calculated as

$$\hat{M}_i(\tau) = N_i(\tau) - \sum_{k=1}^{K} \hat{\pi}_{ik} \int_0^{\tau} \mathrm{d}\hat{\Lambda}_i(\tau|c_{ik} = 1),$$

where $\Lambda_i(t|c_{ik} = 1)$ is obtained through $S(t|c_{ik} = 1)$ given in (12). Plotting the smoothed martingale residuals obtained from fitting the model excluding a covariate against the covariate allows us to assess the functional form through which the covariate affects hazards (Therneau, Grambsch, and Fleming 1990). The LOWESS fitted lines in the martingale residual plots (figure not shown here) are virtually horizontal, indicating that the linear forms of both baseline age and Se level are sufficient for model (3).

7. DISCUSSION

We have proposed a latent class joint model to extract meaningful subgroups with respect to the joint distribution of a longitudinal biomarker and an event process. Application to the PSA data demonstrates the flexibility of the model and the explanatory power and practicality of the methodology. Our methods merge the literature on survival outcomes, longitudinal responses, and latent class models. Our approach can be applied to other biomarker data and extends easily to multiple longitudinal biomarkers. Besides biomarkers, our latent class joint model can be applied to modeling informative dropouts in longitudinal studies, where the survival endpoint is taken to be the time to dropout and one is primarily interested in the longitudinal response that might be related to the dropout process. The latent class approach might help in identifying which patterns of measurement over time are more or less likely to be associated with the dropout.

A weakness of our latent class approach is the complexity of the model, which causes difficulties in computation and interpretation. Currently, there is no readily available software to fit the latent class joint model; however, readers are welcome to contact the first author for the S-PLUS code.

APPENDIX A: CALCULATION IN THE EM ALGORITHM

A.1 The E Step: Calculation of the Conditional Expectations

The first three terms in (8) are easily expressed, and the fourth term is given as

$$\sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}\left\{N_i(\tau)\log\omega_i + \int_0^{\tau} \log\{Y_i^{\gamma}(t)\lambda_k(t)\}\,\mathrm{d}N_i(t) - \omega_i \int_0^{\tau} Y_i^{\gamma}(t)\,\mathrm{d}\Lambda_k(t)\right\}. \quad \text{(A.1)}$$

First, we calculate the conditional expectation of class membership as

$$\tilde{c}_{ik} \equiv \mathrm{E}(c_{ik}|y_i, N_i, Y_i, \check{X}_i; \hat{\Psi}) = \frac{\pi_{ik}[y_i|c_{ik} = 1, \check{X}_i; \hat{\Psi}][N_i, Y_i|c_{ik} = 1, \check{X}_i; \hat{\Psi}]}{\sum_{k=1}^{K} \pi_{ik}[y_i|c_{ik} = 1, \check{X}_i; \hat{\Psi}][N_i, Y_i|c_{ik} = 1, \check{X}_i; \hat{\Psi}]}, \quad \text{(A.2)}$$

where $[y_i|c_{ik} = 1, \check{X}_i]$ and $[N_i, Y_i|c_{ik} = 1, \check{X}_i]$ are given in (6) and (7). The conditional expectations of b_i, $b_ic_i^{\mathrm{T}}$, $b_ib_i^{\mathrm{T}}$, and $b_i^{\mathrm{T}}Z_i^{\mathrm{T}}Z_ib_i$ in $\sum_{i=1}^{n}\{\log[y_i|b_i, c_i, \check{X}_i] + \log[b_i]\}$ are calculated as

$$\tilde{b}_i = \mathrm{E}\{\mathrm{E}(b_i|y_i, N_i, Y_i, c_i, \check{X}_i)\} = V_iZ_i^{\mathrm{T}}(y_i - X_i\beta - W_iM\tilde{c}_i)/\sigma^2, \quad \text{(A.3a)}$$

$$\widetilde{bc}_i = \mathrm{E}\{\tilde{b}_i^c c_i^{\mathrm{T}}\} = V_iZ_i^{\mathrm{T}}\{(y_i - X_i\beta)\tilde{c}_i^{\mathrm{T}} - W_iM\,\mathrm{diag}(\tilde{c}_i)\}/\sigma^2, \quad \text{(A.3b)}$$

and

$$\widetilde{bb}_i \equiv \mathrm{E}(b_i^c b_i^{c'}|y_i, N_i, Y_i, \check{X}_i) = \mathrm{E}\{V_i + \tilde{b}_i^c(\tilde{b}_i^c)^{\mathrm{T}}\} = V_i + \widetilde{b_i^c b_i^{\prime c}}. \quad \text{(A.3c)}$$

The $\widetilde{b_i^c b_i^{\prime c}}$ in the foregoing is calculated as

$$\mathrm{E}(b_ib_i'|y_i, N_i, Y_i, c_i, \check{X}_i) = \tilde{b}_i(y_i - X_i\beta)^{\mathrm{T}}Z_iV_i/\sigma^2 + V_iZ_i^{\mathrm{T}}\{W_iM\,\mathrm{diag}(\tilde{c}_i) - (y_i - X_i\beta)\tilde{c}_i^{\mathrm{T}}\} \times M^{\mathrm{T}}W_i^{\mathrm{T}}Z_iV_i/\sigma^4.$$

The conditional expectations of ω_i and $\log\omega_i$ are calculated as

$$\tilde{\omega}_i \equiv \mathrm{E}\{\omega_i|y_i, N_i(\tau), Y_i(\tau), c_{ik} = 1, \check{X}_i\} = \sum_{k=1}^{K} \tilde{c}_{ik}\frac{1 + \theta N_i(\tau)}{1 + \theta\int_0^{\tau}\lambda_k(u)Y_i^{\gamma}(u)\,\mathrm{d}u} \quad \text{(A.4a)}$$

and

$$\widetilde{\log\omega}_i = \psi\left\{\frac{1}{\theta} + N_i(\tau)\right\} - \sum_{k=1}^{K}\tilde{c}_{ik}\log\left\{\frac{1}{\theta} + \int_0^{\tau}\lambda_k(u)Y_i^{\gamma}(u)\,\mathrm{d}u\right\}. \quad \text{(A.4b)}$$

A.2 Score and Information for the Complete Data

The score and information for η are calculated similarly to those of Lin et al. (2000) and therefore are omitted here. The score and information for $1/\theta$ are calculated similarly to those of Nielsen, Gill, Andersen, and Sørensen (1992) and are also omitted here. Lin (2000) illustrated that the partial likelihood of γ obtained via (10c) for the frailty submodel (3) is actually the profile likelihood with the cumulative hazards profiled out, even in the presence of time-dependent covariables, recurrent events, and discontinuity in observation time. Murphy and van der Vaart (2000) gave a general result stating that the profile likelihood behaves like an ordinary likelihood. Plugging

$\{\hat{\lambda}_k(t_j), \hat{\Lambda}_k(t_j)\}$ from (9) into (A.1), we obtain, aside from constants, the following expression of the weighted partial likelihood for γ:

$$\sum_{i=1}^{n} N_i(\tau)\gamma^{\mathrm{T}} x_i(t) - \sum_{k=1}^{K}\sum_{t_j\le\tau} \tilde{c}_{(j)k} \log\left\{\sum_{i=1}^{n} \tilde{c}_{ik}\tilde{\omega}_i Y_i(t_j)\exp\left(\gamma^{\mathrm{T}} x_i(t)\right)\right\}. \quad \text{(A.5)}$$

This result allows us to use the partial likelihood estimates for γ and the weighted version of Breslow estimates for Λ, both of which are readily available from the S-PLUS routine coxph. Taking the first and second derivatives with respect to γ in the partial likelihood (20), we obtain the following score and information for γ:

$$\mathrm{S}(\gamma) = \sum_{i=1}^{n} N_i(\tau)x_i - \sum_{k=1}^{K}\sum_{t_j\le\tau} \hat{\lambda}_k(t_j)\sum_{i=1}^{n} \tilde{c}_{ik}\tilde{\omega}_i Y_i^{\gamma}(t_j)x_i$$

and

$$\mathrm{I}(\gamma) = \sum_{k=1}^{K}\sum_{t_j\le\tau} \hat{\lambda}_k(t_j)\left\{\sum_{i=1}^{n} \tilde{c}_{ik}\tilde{\omega}_i Y_i^{\gamma}(t_j)x_i^{\otimes 2} - \frac{\hat{\lambda}_k(t_j)}{\tilde{c}_{(j)k}}\left[\sum_{i=1}^{n} \tilde{c}_{ik}\tilde{\omega}_i Y_i^{\gamma}(t_j)x_i\right]^{\otimes 2}\right\},$$

where $a^{\otimes 2}$ denotes the outer product aa^{T} for a vector a.

[Received September 2000. Revised August 2001.]

REFERENCES

Anderson, G. L., and Fleming, T. R. (1995), "Model Misspecification in Proportional Hazards Regression," *Biometrika*, 82, 527–541.

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., and Rathouz, P. J. (1997), "Latent Variable Regression for Multiple Discrete Outcomes," *Journal of the American Statistical Association*, 92, 1375–1386.

Breslow, N. E. (1972), "Covariance Analysis of Censored Survival Data," *Biometrics*, 34, 216–217.

Catalona, W. J., Smith, D. S., and Ornstein, D. K. (1997), "Prostate Cancer Incidence in Men With Serum PSA Concentration of 2 to 4 ng/ml and Benign Prostatic Examination," *Journal of the American Medical Association*, 277, 1452–1455.

Chen, J., and Kalbfleisch, J. D. (1996), "Penalized Minimum-Distance Estimates in Finite Mixture Models," *Canadian Journal of Statistics*, 24, 167–175.

Clark, L. C., Combs, G. F. Jr., Turnbull, B. W., Slate, E.-H., Haiker, D. K., Chow, J., Davis, L. S., Glover, R. A., Graham, G. F., Gross, E. G., Krongrad, A., Lesher, J. L., Park, H. K., Sanders, B. B., Smith, C. L., and Taylor, J. R. (1996), "Effects of Selenium Supplementation for Cancer Prevention in Patients With Carcinoma of the Skin," *Journal of the American Medical Association*, 276, 1957–1963.

DeGruttola, V., and Tu, X. M. (1994), "Modeling Progression of CD4+ Lymphocyte Count and Its Relationship to Survival Time," *Biometrics*, 50, 1003–1014.

Faucett, C. L., and Thomas, D. C. (1996), "Simultaneously Modeling Censored Survival Data and Repeatedly Measured Covariates: A Gibbs Sampling Approach," *Statistics in Medicine*, 16, 1663–1685.

Greenlee, R. T., Murray, T., Bolden, S., and Wingo, P. A. (2000), "Cancer Statistics, 2000," *CA—A Cancer Journal for Clinicians*, 50, 7–33.

Henderson, R., Diggle, P., and Dobson, A. (2000), "Joint Modelling of Longitudinal Measurements and Event Time Data," *Biostatistics*, 1, 465–480.

Leroux, B. G. (1992), "Consistent Estimation of a Mixture Distributions," *The Annals of Statistics*, 20, 1350–1360.

Lin, H. Q. (2000), "A Finite Mixture Model for Joint Longitudinal Biomarker and Survival Outcome Responses," Ph.D. thesis, Cornell University, Ithaca, New York.

Lin, H. Q., McCulloch, C. E., Turnbull, B. W., Slate, E. H., and Clark, L. C. (2000), "A Latent Class Mixed Model for Analyzing Biomarker Trajectories in Longitudinal Data With Irregularly Scheduled Observations," *Statistics in Medicine*, 19, 1303–1318.

McLachlan, G. J., and Krishnan, T. (1997), *The EM Algorithm and Extensions*, New York: Wiley.

Morrell, C., Pearson, J. D., Carter, H. B., and Brant, L. J. (1995), "Estimating Unknown Transition Times Using a Piecewise Nonlinear Mixed-Effects Model in Men With Prostate Cancer," *Journal of the American Statistical Association*, 90, 45–53.

Murphy, S. A., and van der Vaart, A. W. (2000), "On Profile Likelihood" (with discussion), *Journal of the American Statistical Association*, 95, 449–485.

Muthén, B., and Shedden, K. (1999), "Finite Mixture Modeling With Mixture Outcome Using the EM Algorithm," *Biometrics*, 55, 463–469.

Nielsen, G. G., Gill, R. D., Andersen, P. K., and Sørensen, T. I. A. (1992), "A Counting Process Approach to Maximum Likelihood Estimation in Frailty Models," *Scandinavian Journal of Statistics*, 19, 25–44.

Padilla-Nash, H., Heselmeyer-Haddad, K., Wangsa, N., Zhang, H., Ghadimi, B., Macville, M., Augustus, M., Schrock, E., Hilgenfeld, E., and Ried, T. (2001), "Jumping Translocations are Common in Solid Tumor Cell Lines and Result in Recurrent Fusions of Whole Chromosome Arms," *Genes, Chromosones and Cancer*, 30, 349–363.

Parner, E. (1998), "Asymptotic Theory for the Correlated Gamma—Frailty Model," *The Annals of Statistics*, 26, 183–214.

Pawitan, Y., and Self, S. (1993), "Modeling Disease Marker Processes in AIDS," *Journal of the American Statistical Association*, 88, 719–726.

Pearson, J. D., Morrell, C. H., Landis, P. K., Carter, H. B., and Brant, L. J. (1994), "Mixed-Effects Regression Models for Studying the Natural History of Prostate Disease," *Statistics in Medicine*, 13, 587–601.

Rosen, O., and Tanner, M. (1999), "Mixtures of Proportional Hazards Regression Models," *Statistics in Medicine*, 18, 1119–1131.

Schluchter, M. D. (1992), "Methods for the Analysis of Informatively Censored Longitudinal Data," *Statistics in Medicine*, 11, 1861–1870.

Schwartz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Slate, E. H., and Clark, L. C. (1999), "Using PSA to Detect Prostate Cancer Onset: An Application of Bayesian Retrospective and Prospective Changepoint Identification," in *Case Studies in Bayesian Statistics IV*, eds. C. Gatsonis, B. Carlin, A. Carriquiry, A. Gelman, R. Kass, I. Verdinelli, and M. West, New York: Springer-Verlag, pp. 511–534.

Slate, E. H., and Turnbull, B. W. (2000), "Statistical Models for Longitudinal Biomarkers of Disease Onset," *Statistics in Medicine*, 19, 617–637.

Therneau, T. M., Grambsch, P. M., and Fleming, T. R. (1990), "Martingale Based Residuals for Survival Models," *Biometrika*, 77, 147–160.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1998), *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley.

Tsiatis, A. A., DeGruttola, V., and Wulfsohn, M. S. (1995), "Modeling the Relationship of Survival to Longitudinal Data Measured With Error—Applications to Survival and CD4 Counts in Patients With AIDS," *Journal of the American Statistical Association*, 90, 27–37.

Whittemore, A. S., Lele, C., Friedman, G. D., Stamey, T., Vogelman, J. H., and Orentreich, N. (1995), "Prostate-Specific Antigen as Predictor of Prostate Cancer in Black Men and White Men," *Journal of the National Cancer Institute*, 87, 354–359.

Wulfsohn, M. S., and Tsiatis, A. A. (1997), "A Joint Model for Survival and Longitudinal Data Measured With Error," *Biometrics*, 53, 330–339.