

Bayesian Modeling of Incidence and Progression of Disease from Cross-Sectional Data

David B. Dunson^{1,*} and Donna D. Baird²

¹ Biostatistics Branch and ² Epidemiology Branch,
MD A3-03, National Institute of Environmental Health Sciences
P.O. Box 12233, Research Triangle Park, NC 27709

* *email:* dunson1@niehs.nih.gov

SUMMARY. In the absence of longitudinal data, the current presence and severity of disease can be measured for a sample of individuals to investigate factors related to disease incidence and progression. In this article, Bayesian discrete-time stochastic models are developed for inference from cross-sectional data consisting of the age at first diagnosis, the current presence of disease, and one or more surrogates of disease severity. Semiparametric models are used for the age-specific hazards of onset and diagnosis, and a normal underlying variable approach is proposed for modeling of changes with latency time in disease severity. The model accommodates multiple surrogates of disease severity having different measurement scales and heterogeneity among individuals in disease progression. A Markov chain Monte Carlo algorithm is described for posterior computation, and the methods are applied to data from a study of uterine leiomyoma.

KEY WORDS: Current Status; Disease Markers; Factor analysis; Latency; Multiple outcomes; Multistate model; Sojourn time; Tumor size

1. Introduction

Many diseases have a long preclinical phase during which the disease progresses and eventually becomes clinically recognized. Slowing or preventing this progression may be a more effective strategy for reducing morbidity than trying to prevent onset. With improvements in laboratory and imaging technologies, it is becoming more common to screen for both clinical and preclinical disease, collecting data on occurrence and severity. Cross sectional screening data have not been considered useful for assessing preclinical disease progression, but have been used for incidence analyses. This article focuses on the problem of utilizing data on preclinical disease severity to conduct joint analyses of incidence and preclinical progression from cross sectional data. This research is motivated by the study of uterine leiomyoma (fibroids). The condition can be detected preclinically with ultrasound, and ultrasound can provide measures of severity based on the size of leiomyoma lesions.

Uterine leiomyoma are common tumors that rarely become malignant but lead to substantial morbidity; resulting in pelvic pain, infertility, pregnancy complications, excessive uterine bleeding and hysterectomy. African American women are considered at higher risk of leiomyoma than white women (ACOG, 1994). However, the factors contributing to differences in leiomyoma severity, symptoms, and age at diagnosis between black and white women (Kjerulff et al., 1996) are unknown. In earlier work, we observed higher incidence and diagnosis rates among African American women based on applying a flexible parametric model (Dunson and Baird, 2001) to current status and age at first diagnosis data for a sample of women from a large urban health plan (Baird et al., 2002). This article is motivated by the important problem of assessing black-white differences in not only leiomyoma incidence and rate of diagnosis but also rate of progression in severity (quantified by length of the uterus and size/number of fibroids).

One approach to assessing differences between groups in disease severity would be to simply compare surrogates of severity for individuals in the different groups through a simple

age-adjusted statistical test (e.g., using linear regression). However, this approach would not tell us whether an increase in severity is attributable to earlier onset of disease, more rapid progression of disease, or both. Distinguishing among these alternatives is important for understanding etiology and disease progression to develop prevention strategies.

The problem of estimating disease incidence from interval censored data has been studied in the context of animal survival/sacrifice experiments (Dinse and Lagakos, 1982; Turnbull and Mitchell, 1984), AIDS studies of incubation and diagnosis distributions (DeGruttola and Lagakos, 1989; Frydman, 1992), and chronic disease screening studies (Duffy et al., 1995; Chen et al., 2000). However, few methods are available for modeling disease progression after onset when age at onset is interval censored. Ryan and Orav (1988) incorporated data on tumor size and grade as covariates modifying the rate of natural death in a three state model for tumor onset and death, but did not model tumor progression after onset. Also in the setting of animal tumorigenicity experiments, Rai and Matthews (1995) modeled the tumor growth process using a time-homogenous birth-death model. In the setting of chronic disease screening, Chen et al. (2000) proposed a multistate time-homogenous Markov model that accounts for progression between different disease states (e.g., tumor categories) within the preclinical detectable phase of disease (PCDP). The Chen et al. model does not account for information on clinical history, for continuous measures of disease severity, or for changes with age in the transition rates.

Following Chen et al. (2000), we use a three-state stochastic model to characterize the progression of an individual through the states: (1) disease-free, (2) preclinical detectable disease, and (3) clinical disease. Semiparametric discrete-time Bayesian models are used for the hazard rates characterizing the transitions from states $1 \rightarrow 2$ and $2 \rightarrow 3$. To allow multiple ordered categorical and continuous surrogates of the severity of disease for individuals in the PCDP (i.e., state 2), we propose a general modeling framework in which the different surrogates are linked to underlying normal variables. The expectation of these

underlying variables depends on waiting time in the disease state, on covariates associated with rate of progression, and on subject-specific parameters accommodating heterogeneity among individuals.

When interest focuses on inference on the rate of progression and several surrogates are available, our underlying variable approach has advantages over previous multistate models that categorize disease severity into a single set of states (e.g., Chen et al., 2000; Craig et al., 1999). It is typically not clear how different surrogates should be combined into a single ranking of disease severity. For example, in the uterine leiomyoma application, assignments of low, medium, and high levels of leiomyoma severity based on the uterine length and size/number of tumors would be arbitrary.

Alternatively, the primary interest may be disease incidence and one may wish to include disease severity data as a covariate (e.g., modifying the rate of transition into the clinical state). If data are collected prospectively with disease severity measured at each possible time of transition into the clinical state, then it is straightforward to incorporate the severity data as covariates in the transition rate model (see, for example, Craig et al., 1999), though there may be problems with collinearity if the different severity surrogates are highly correlated. Typically, it is not practical to collect information on disease severity at every possible transition time, and severity data are collected only periodically or at a single screening exam (as in the uterine leiomyoma study). In such cases, disease severity information cannot be incorporated without modeling of the progression in the different surrogates, even when interest focuses on disease incidence. For this reason, Craig et al. (1999) chose to not incorporate glycosylated haemoglobin level, a periodically observed surrogate variable, into their model for diabetic retinopathy (refer to page 1359, paragraph 2).

We follow a Bayesian approach to inference using Markov chain Monte Carlo (MCMC) methods (Gilks, Richardson, and Spiegelhalter, 1996; Tierney, 1994). Since direct observations of the time of entry into state 2 and of the changes across time in disease severity are

not available, we simplify MCMC implementation by using a data augmentation approach (Tanner and Wong, 1987) to impute the unknown disease onset times. Bayesian stochastic modeling approaches have been used previously for interval-censored chronic disease data in order to estimate the incidence and progression of disease in the presence of intervention (Craig et al., 1999), to assess treatment effects on tumor incidence in survival/sacrifice experiments (Dunson and Dinse, 2001), and to characterize the risk of false-positive results under repeated screening tests (Gelfand and Wang, 2000).

Section 2 describes the underlying stochastic model and observed data likelihood. Section 3 presents Bayesian regression models for the transition rates, and outlines an MCMC algorithm. Section 4 illustrates the methods through application to data from the uterine leiomyoma study, and Section 5 discusses the results.

2. Modeling Disease Onset and Progression

2.1 Data Structure

Suppose that the disease onset and progression process can be characterized by the progressive three-state model shown in Figure 1. We denote the transition times into states 2 and 3 by R and S , respectively, where R is the age at onset of preclinical detectable disease and S is the age at first clinical diagnosis. Our interests focus on the age-specific disease incidence rate $\lambda(t)$ characterizing the transition from state 1 \rightarrow 2, the rate of progression of disease within state 2, and the age-specific rate of disease diagnosis $\alpha(t)$.

To obtain information about disease onset and progression, individuals are screened at a random age T , where T is assumed to be noninformative about ages R and S . We focus on the data structure described in Table 1, where $Y = 1(R \leq T)$ and $D = 1(S \leq T)$ are indicators of disease onset and clinical diagnosis, respectively, prior to screening. For individuals who are disease free at T , the ages R and S are right censored. For individuals with preclinical disease detected at T , R is left censored and S is right censored. Finally, for

individuals with a previous diagnosis of disease, R is left censored and S is known.

In addition, for individuals in the PCDP (state 2) at T , one or more surrogates of the current severity of the disease are available. In some cases, surrogate data may also be available for individuals in state 3. However, the distribution of the surrogates among previously diagnosed individuals may depend in part on medical interventions that occurred after diagnosis. Therefore, since our focus is on the onset and progression process prior to clinical intervention and we wish to avoid assumptions about the intervention process, we do not incorporate surrogate data for individuals with a clinical history.

2.2 Stochastic Model

Consider a cross-sectional study of individuals aged between t_{min} and t_J so that $T \in [t_{min}, t_J]$. Let $t_{min} < t_1 < t_2 < \dots < t_J$ denote a prespecified grid of ages which partitions $[t_{min}, t_J]$, and let $I_j = (t_{j-1}, t_j]$ for $j = 1, \dots, J$ with $t_0 = 0$. This structure ensures that I_j overlaps with the range of screening ages $[t_{min}, t_J]$ for $j = 1, \dots, J$, which is important for identifiability reasons that will be discussed later. For simplicity, we assume that the interval widths $t_j - t_{j-1}$ are approximately equal for $j = 2, \dots, J$. We characterize the onset and diagnosis process according to the transition rates:

$$\lambda_j = \Pr(R \in I_j | R > t_{j-1}) \quad \text{and} \quad \alpha_j = \Pr(S \in I_j | S > t_{j-1}, R \leq t_j), \quad (1)$$

where λ_j and α_j are the discrete hazards of onset of preclinical disease and of clinical diagnosis for individuals with disease, respectively, within I_j ($j = 1, \dots, J$).

Let Z_k denote the value for the k th measure of disease severity at T for $k = 1, \dots, q$. For example, in the uterine fibroid application described in Section 4, we let Z_1 be the length of the uterus and Z_2 be a 1-3 ranking of the size/number of tumors. Suppose that Z_k is linked to an underlying variable Z_k^* through

$$Z_k = g_k(Z_k^*; \boldsymbol{\tau}_k), \quad k = 1, \dots, q, \quad (2)$$

where $g_k(\cdot; \boldsymbol{\tau}_k)$ is a monotone function involving parameters $\boldsymbol{\tau}_k$. For continuous surrogates (e.g., length of uterus), $\boldsymbol{\tau}_k$ typically consists of intercept parameters, which define the conditional expectation of Z_k for $Z_k^* = 0$, and scale parameters, which fix the variance of Z_k relative to that for the underlying variable (e.g., $Z_1 = \exp(\tau_{11} + \tau_{12}Z_k^*)$). For an ordinal surrogate (e.g., ranking of fibroid size), $\boldsymbol{\tau}_k$ consists of threshold parameters defining the values of the underlying variable, Z_k^* , corresponding to each category of the observed Z_k (e.g., $Z_2 = 1$ if $Z_2^* < \tau_{21}$, $Z_2 = 2$ if $\tau_{21} \leq Z_2^* < \tau_{22}$, and $Z_2 = 3$ otherwise).

We assume the following model for the distribution of the underlying variable Z_k^* among individuals in the PCDP at screening conditional on the interval of disease onset and the current age:

$$(Z_k^* | R \in I_j, T \in I_l, j \leq l, S > t_l) \sim N\left(\sum_{h=1}^{l-j} \mu_{hk}, 1\right), \quad k = 1, \dots, q, \quad (3)$$

where the conditional expectation of Z_k^* is 0 when $j = l$ (i.e., screening occurs within the same interval as onset) and the expectation changes according to the number of intervals between onset and screening. For shorthand use later, let $f_{jk}(\cdot; t)$ refer to the conditional density of Z_k^* given onset in interval I_j and screening at t . We initially assume conditional independence of the elements of Z_1, \dots, Z_q given R and S , though we describe models for accommodating within-subject dependency between the different surrogates of disease severity in Section 3. Note that the variance of the underlying variable is fixed at one in expression (3), which ensures identifiability of the scale parameters for continuous surrogates and of the threshold parameters for categorical surrogates. For continuous surrogates, expression (3) can potentially be generalized to account for heteroscedastic error variances, though in our experience additional variance parameters tend to be only weakly identified by the data even under restrictions on the link function (e.g., scale = 1).

The form of expressions (2) and (3) is quite flexible in that the different surrogates of disease severity can change nonhomogeneously with time spent in the preclinical state

through a piecewise constant model. In addition, the different surrogates can have different measurement scales, expectations, error variances, and even distributional forms (e.g., by using a linear link for one surrogate and an exponential link for another). We have described a related underlying variable approach in previous work (Dunson, 2000), though this earlier approach uses a different modeling structure and does not allow the outcome variables (in this case the severity surrogates) to depend on the unknown waiting time in the disease state.

Although direct effects of disease severity on the rate of diagnosis cannot be identified from cross sectional observations of preclinical severity, our model does accommodate such effects by allowing disease severity to depend on the waiting time in the preclinical state. If a high level of severity tends to lead to symptoms that result in a clinical diagnosis, the average severity may decrease at long latency times as the more severe cases are diagnosed and removed from the preclinical state. In interpreting covariate effects on the rates of diagnosis and progression, one should keep in mind that the diagnosis rates α_j are marginal rates integrated across the density of the disease severity surrogates and that the progression rates are defined conditionally on remaining in the preclinical phase.

The model characterized by expression (1) - (3) differs from an earlier approach described by Dunson and Baird (2001) in not only the incorporation of the surrogate data (which is not considered in the earlier approach) but also in the age-specific hazard rate formulation of the three-state onset and diagnosis process. The earlier approach instead used fully parametric models for the onset and diagnosis time c.d.f.'s without direct modeling of the hazard rates. In disease progression applications, the proposed underlying variable approach has advantages over previous underlying variable models for multiple outcome data (e.g., Muthén, 1984; Arminger and Küsters, 1988; Dunson, 2000) in that the density of the underlying variables depends explicitly on the unknown waiting time in the preclinical state. In the uterine leiomyoma application and in other chronic disease settings, it is natural to

assume that surrogates of disease severity increase stochastically from the time of onset.

2.3 Observed Data Likelihood

As a convention to account for multiple types of events within I_j , we assume that onset of preclinical disease occurs before clinical diagnosis, which occurs before the screening examination. Within this context and under expressions (1) - (2), the likelihood contribution for an individual in state 1 at T (indicated by $Y = 0, D = 0$) is

$$\prod_{j=1}^{l:T \in I_l} (1 - \lambda_j). \quad (4)$$

Individuals in state 2 at T (indicated by $Y = 1, D = 0$) with severity \mathbf{Z} contribute

$$\left[\sum_{m=1}^{l:T \in I_l} \left\{ \prod_{j=1}^{m-1} (1 - \lambda_j) \right\} \lambda_m \left\{ \prod_{j=m}^{l:T \in I_l} (1 - \alpha_j) \right\} \left\{ \prod_{k=1}^q \int 1\{Z_k = g_k(z^*; \boldsymbol{\tau}_k)\} f_{mk}(z^*; T) dz^* \right\} \right]. \quad (5)$$

Finally, individuals in state 3 at T (indicated by $Y = 1, D = 1$) reporting their first diagnosis at S ($S \leq T$) contribute

$$\left[\sum_{m=1}^{l:S \in I_l} \left\{ \prod_{j=1}^{m-1} (1 - \lambda_j) \right\} \lambda_m \left\{ \prod_{j=m}^{l:S \in I_{l+1}} (1 - \alpha_j) \right\} \right] \alpha_{\{l:S \in I_l\}} \quad (6)$$

to the observed data likelihood.

3. Bayesian Semiparametric Analysis

3.1 Component Regression Models

In this subsection, we describe Bayesian regression models for each component of the stochastic model described in Section 2. Suppose that a study involves n individuals. We assume that the discrete hazard of onset of preclinical detectable disease within interval I_j ($j = 1, \dots, J$) for individual i ($i = 1, \dots, n$) is

$$\lambda_{ij} = \Pr(R_i \in I_j | R_i > t_{j-1}) = h_1(\omega_j + \mathbf{x}_{ij}^T \boldsymbol{\beta}), \quad (7)$$

where $h_1(\omega_j)$ is the baseline discrete hazard of onset in interval I_j , \mathbf{x}_{ij} is a $p \times 1$ vector of covariates, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ is a $p \times 1$ vector of regression parameters. In addition, we

assume that the discrete hazard of diagnosis at t_j for individual i conditional on entry in the preclinical detectable phase of disease is

$$\alpha_{ij} = \Pr(S_i \in I_j \mid S_i > t_{j-1}, R_i \leq t_j) = h_2(\nu_j + \mathbf{x}_{ij}^T \boldsymbol{\psi}), \quad (8)$$

where $h_2(\nu_j)$ is the baseline hazard of diagnosis in I_j given onset by t_j , and $\boldsymbol{\psi} = (\psi_1, \dots, \psi_p)^T$ is a $p \times 1$ vector of regression parameters.

In addition to these semiparametric models for the transition rates, we specify a factor analytic model for Z_{ik} , the k th surrogate of disease severity measured for individual i conditional on presence in the PCDP at screening. This model is as specified in expressions (2) and (3) with a subscript i added to denote individual i , and with

$$\mu_{ihk} = \mathbf{u}_h^T \mathbf{v}_k + \mathbf{x}_{ih}^T \boldsymbol{\kappa}_k + \sigma \xi_i, \quad h = 1, \dots, J - 1, \quad (9)$$

where the elements of \mathbf{u}_h are orthogonal functions of h , \mathbf{v}_k are baseline parameters, \mathbf{x}_{ih} are covariates, $\boldsymbol{\kappa}_k$ is a vector of regression parameters, σ is a factor loading ($\sigma > 0$), and $\xi_i \sim N(0, 1)$ is a latent variable measuring the rate of progression of disease within the PCDP for individual i . The term $\sigma \xi_i$ accounts for heterogeneity among individuals in the rate of disease progression. Alternatively, we could have used $b_i \sim N(0, \sigma^2)$ in place of $\sigma \xi_i$. However, our factor analytic form appears to have a clear advantage in terms of rate of convergence of the MCMC algorithm in the examples we have considered. The parameter σ is identified by the magnitude of within-subject dependency in the different surrogates and one could potentially extend (9) to account for a more complex dependency structure by including additional factor loading parameters.

3.2 Model Identifiability and Prior Specification

In order to separately estimate the $\{\omega_j\}$ and $\{\nu_j\}$ parameters based on the observed data (e.g., by unconstrained maximum likelihood estimation), there must be at least one screening exam and at least one clinical diagnosis within I_j for $j = 1, \dots, J$. In addition, in order

to estimate the baseline progression process parameters $\{\boldsymbol{\nu}_k\}$ and $\{\boldsymbol{\tau}_k\}$ from the observed data, there must be at least one individual among those screened in I_j who is in the PCDP for $j = 1, \dots, J$, with one or more of the intervals having multiple such subjects. To improve identifiability of the model while avoiding restrictive parametric assumptions about the disease onset-diagnosis process, we specify an informative prior on the degree of local nonlinearity in the time-specific baseline parameters. This approach will produce smoothed estimates of the time-specific baseline parameters with the degree of smoothing dependent on investigator-specified hyperparameters.

Let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_J)^T$ refer to one of the baseline parameter vectors: $\boldsymbol{\omega} = (\omega_1, \dots, \omega_J)^T$ and $\boldsymbol{\nu} = (\nu_1, \dots, \nu_J)^T$. We place a prior on the following measure of local nonlinearity:

$$\tilde{\theta}_j = \left| \theta_j - \frac{1}{2}(\theta_{j-1} + \theta_{j+1}) \right| \quad \text{for } j = a, \dots, b,$$

where a and b place bounds on the range of times to smooth across, and this expression can easily be modified to accommodate variable interval widths. A perfectly linear change with time in the baseline parameters implies that $\tilde{\theta}_j = 0$, and the magnitude of local nonlinearity (i.e., roughness) tends to increase rapidly with $\tilde{\theta}_j$. We choose priors of the form:

$$\tilde{\theta}_j \sim \text{gamma}(c, d), \quad j = a, \dots, b, \tag{10}$$

where c and d are prior parameters chosen subjectively based on one's opinion of the degree of local nonlinearity in $\boldsymbol{\theta}$. The degree of smoothing tends to increase as $c/d \rightarrow 0$ and the prior variance c/d^2 decreases.

In contrast to some alternative prior process approaches for Bayesian discrete-time survival analysis, which assume a degree of smoothness for the rates in adjacent intervals (e.g., Gamerman, 1991; Sinha, 1998), our prior focuses specifically on the degree of local nonlinearity without incorporating prior information on the overall mean and rate of change in the baseline rates, though such information can be incorporated when available through priors

for θ_1 and θ_J . Sinha (1998) instead used a prior of the form $\log(\theta_{j+1}) \sim N(\log\theta_j, \sigma^2)$, which does not restrict the overall mean but does tend to pull the overall rate of change towards a slope of zero, particularly for small σ .

We assign the parameters, $\boldsymbol{\beta}$, $\boldsymbol{\psi}$, and $\{\boldsymbol{\nu}_k, \boldsymbol{\kappa}_k\}$, normal priors with prior covariance between the elements of the different parameter vectors set to zero for simplicity. In addition, we assign the factor loading σ a normal prior truncated below by zero. The choice of the prior for the link parameters $\{\boldsymbol{\tau}_k\}$ will depend on the form of the link function.

3.3 MCMC Implementation

First, consider the case where none of the study subjects miss the screening exam. Our MCMC algorithm alternates between sampling new values for the latent data (e.g., unknown disease onset intervals, underlying severity variables) and for the population parameters defining each component of the stochastic model based on the relevant full conditional distributions. This algorithm involves both Gibbs sampling (Gelfand and Smith, 1990) and Metropolis-Hastings (MH; Hastings, 1970) steps and can be summarized as follows:

1. For subjects in states 2 or 3 sample the unknown interval of entry into state 2.
2. Sample the underlying variables $\{Z_{ik}^*\}$ and link parameters $\{\boldsymbol{\tau}_k\}$.
3. Sample the parameters: $\{\omega_j\}$, $\boldsymbol{\beta}$, $\{\nu_j\}$, and $\boldsymbol{\psi}$, characterizing the regression models (7) and (8) for the onset and diagnosis process.
4. Sample the parameters: $\{\boldsymbol{\nu}_k, \boldsymbol{\kappa}_k\}$, and σ and the latent variables: $\{\xi_i\}$, characterizing model (9) for the disease progression process.
5. Repeat steps 1-4 until apparent convergence, and estimate posterior summaries of the parameters of interest based on a large number of additional iterates.

The necessary conditional distributions are described in an Appendix. Step 1 can be easily

modified to allow for uncertainty in the disease state for subjects with no disease history ($D = 0$) who miss the screening exam.

4. Uterine Leiomyoma Application

4.1 *The Data*

We illustrate the proposed methodology through application to the uterine leiomyoma example discussed in Section 1. Data were drawn from a cross-sectional study conducted by the National Institute of Environmental Health Sciences (NIEHS) (Baird et al., 2002). Briefly, study participants were 840 African American and 524 white women aged 35-49, randomly selected from the membership list of a large urban health plan. Demographic and clinical history data were collected by telephone interview and self-administered questionnaire. For premenopausal women, pelvic ultrasound was used to determine the presence or absence of fibroids and their severity, as measured by the length of the uterus (Z_1) and an ordinal ranking of the size/number of fibroids:

$$Z_2 = \begin{pmatrix} 1 & \text{diffuse pattern or 1 tumor} < 3\text{cm} \\ 2 & 1 \text{ tumor} \geq 3\text{cm or } 2+ \text{ tumors} < 3\text{cm} \\ 3 & 2+ \text{ tumors with one or more} \geq 3\text{cm} \end{pmatrix}$$

Data for postmenopausal women consist of self report and surgical/pathology reports of hysterectomies, since postmenopausal women were not given a study ultrasound because fibroids can shrink after menopause. In the analysis we will make the simplifying assumption that the rate of occurrence of menopause is not associated with the presence or absence of preclinical leiomyoma. Under this assumption, the postmenopausal women can be factored into the likelihood through incorporation of the clinical history data for these women without biasing inference about disease onset and progression prior to menopause.

Summary statistics of the data for black and white women are provided in Table 2. Data on the age at first diagnosis and current status of fibroids were analyzed previously to assess black-white differences in cumulative incidence using a flexible parametric model fitted by maximum likelihood (Dunson and Baird, 2001). The primary goal of our current analysis is

to incorporate data on the severity of uterine leiomyoma (measured by length of the uterus and size/number of tumors) within our proposed Bayesian modeling framework to perform inference on black-white differences in rate of progression of uterine leiomyoma within the preclinical detectable phase of disease. We are also interested in obtaining new estimates of cumulative incidence under our proposed semiparametric hazards model, incorporating information on current disease severity.

4.2 The Model

We chose a discrete-time proportional hazards model for λ_{ij} , the discrete hazard of onset of preclinical detectable leiomyoma for woman i in age interval j ,

$$\log\{-\log(1 - \lambda_{ij})\} = \omega_j + x_i[1(j = 1)\beta_1 + 1(2 \leq j \leq 10)\beta_2 + 1(11 \leq j \leq 14)\beta_3], \quad (11)$$

where $x_i = 1$ for blacks and $x_i = 0$ for whites, and age is divided into $J = 14$ intervals:

$$(0, 36], (36, 37], (37, 38], (38, 39], \dots, (46, 45], (47, 48], (48, 49].$$

This model incorporates a nonparametric baseline hazard and allows differences between African Americans and whites to vary between the age intervals: $(0,36]$, $(36,45]$, and $(45,49]$. We allowed for age-ethnicity interactions, since there is evidence of nonproportional hazards in unconstrained analyses of the NIEHS data and we are interested in whether black-white differences in incidence differ in magnitude between the chosen age intervals. We avoid assuming proportionality in the diagnosis rates between blacks and whites and instead model the two sets of rates separately, since we are considering the diagnosis process as a nuisance and wish to avoid biasing estimates of incidence and progression.

To model progression of leiomyoma following preclinical onset, we first link the length of the uterus, Z_{i1} , and the ranking of the size/number of tumors, Z_{i2} , to underlying variables Z_{i1}^* and Z_{i2}^* through

$$Z_{i1} = \tau_{11} + \tau_{21}Z_{i1}^* \quad \text{and} \quad Z_{i2} = \begin{pmatrix} 1 & Z_{i2}^* \leq \tau_{12} \\ 2 & \tau_{12} < Z_{i2}^* \leq \tau_{22} \\ 3 & \tau_{22} < Z_{i2}^* \end{pmatrix} \quad (12)$$

where τ_{11} is an intercept parameter, τ_{21} is a scale parameter, and τ_{12} and τ_{22} are threshold parameters with $\tau_{22} > \tau_{12}$. The latent variable means are allowed to vary with waiting time in the disease state according to expressions (3) and (9) with

$$\mu_{ihk} = v_k + x_i \kappa_k + \sigma \xi_i \quad \text{for } k = 1, 2 \text{ and } h = 1, \dots, 13, \quad (13)$$

where v_k measures the underlying rate of change in the k th surrogate within the preclinical detectable phase for white women, κ_k measures the difference between African American and white women in this underlying rate of change, and $\sigma \xi_i$ is a factor analytic term accommodating within-woman dependency in the surrogates. Although we also considered more complex models (e.g., using a nonparametric baseline: v_{hk}), we found expression (13) to compare favorably in terms of model fit while avoiding problems with overfitting.

Based on information from exploratory analyses of the model (not using the current data) and on examination of the literature on uterus length and leiomyoma size/number, we chose relatively diffuse priors for each of the parameters. Using the approach described in subsection 3.2, we chose a gamma(1, .5) prior for the transformed age-specific baseline onset rates and gamma(1,.1) priors for the transformed ethnicity group-specific diagnosis rates. In addition, the regression parameters β_1 , β_2 and β_3 were assigned independent $N(0,2)$ priors, and the link parameters were assigned priors: $\tau_{11} \sim N(8.5, 4)$, $\tau_{12}^{-2} \sim \text{gamma}(.06, .2)$, $\tau_{21} \sim N(-1, 4)$ truncated above by τ_{22} , and $\tau_{22} \sim N(.5, 10)$ truncated below by τ_{21} . Finally, in the underlying variable model, we let $v_k \sim N(0, 2)$ truncated below by 0, $\kappa_k \sim N(0, 2)$, and $\sigma \sim N(.1, 2)$ truncated below by 0. We evaluated the sensitivity of our inferences to the prior choice by repeating our analyses under a range of reasonable alternative priors, including those with variance inflated by a factor of 5, variance decreased by a factor of 5, and with prior means chosen to vary substantially from the primary analysis.

4.3 *The Analysis*

We implemented our analyses using the MCMC algorithm described in subsection 3.3. We

generated 30,000 MCMC iterates and discarded the first 5,000 as a burn-in. We assessed convergence using a variety of diagnostic techniques, summarized by Cowles and Carlin (1996) and implemented using BOA (Smith, 2000) in S-PLUS. We found no evidence of lack of convergence or of slow mixing based on standard diagnostic tests and on examination of plots of the sampled parameters. In fact, estimates based on iterations 1001-2000 were essentially identical to our final estimates.

Posterior summaries of the parameters characterizing the rate of uterine leiomyoma growth following preclinical onset are provided in Table 3. In addition, estimated posterior densities of κ_1 and κ_2 , the parameters measuring differences between African Americans and whites in rates of growth of the uterus and of focal leiomyoma tumors, respectively, are plotted in Figure 2 for each choice of prior. Most of the mass of these posterior densities is assigned to positive values of κ_1 and κ_2 , with $\Pr(\kappa_1 > 0) = 0.98$ and $\Pr(\kappa_2 > 0) = 0.96$. In addition, letting $\bar{\kappa} = (\kappa_1 + \kappa_2)/2$, $\Pr(\bar{\kappa} > 0) > 0.99$ for the primary analysis and each of the sensitivity analyses, suggesting that African American women have a higher rate of preclinical growth of uterine leiomyoma than white women.

Both uterine length and size/number of tumors appear to increase more rapidly for African Americans than whites between the time leiomyoma become detectable by pelvic ultrasound and the time of clinical diagnosis. Although it is known that African American hysterectomy patients tend to have larger and more numerous leiomyomas than white patients (Kjerulff et al., 1996), this may simply reflect inequity in health care or higher incidence among blacks. Our result is the first finding of an increased rate of preclinical progression of leiomyoma among African Americans in a randomly selected sample of women. This finding is important from public health and clinical perspectives, and provides motivation for future longitudinal studies of causes of the black-white difference.

In addition to assessing black-white differences in leiomyoma progression, we used our proposed approach to obtain new estimates of cumulative incidence curves for African Amer-

ican and white women. The posterior means and pointwise 90% credible intervals for these curves are presented in Figure 3, along with estimates obtained using our Bayesian semi-parametric model without incorporating surrogate data on disease severity (shown as +). It appears that incorporating data on current uterine length and size/number of tumors has minimal effect on estimates of cumulative incidence, though there were modest decreases in posterior variance associated with incorporation of the surrogate data. We found that our estimates of cumulative incidence were robust to the prior, to the link functions chosen in the incidence and diagnosis rate models, and to the form of the model for progression in the surrogates.

Interestingly, we found that black-white differences in leiomyoma incidence can be attributed primarily to a higher rate among African American women younger than 36 years of age and there was no evidence of a difference in incidence between ethnic groups in other age categories. In particular, the posterior means of β_1 , β_2 and β_3 , the parameters quantifying the age group-specific differences between African Americans and whites, were 1.062, .045, and -.177, respectively. The corresponding posterior probabilities of a higher incidence rate of leiomyoma among African Americans were $\Pr(\beta_1 > 0) > .99$, $\Pr(\beta_2 > 0) = .58$, and $\Pr(\beta_3 > 0) = .32$.

The estimates shown in Figure 3 are similar to those previously reported (Dunson and Baird, 2001), but are slightly lower for the younger ages under study; a difference probably attributable to the Bayesian semiparametric hazard structure of the proposed onset-diagnosis model, which differs substantially from the earlier approach.

4.4 *Simulation Study*

We conducted a small simulation study to check our implementation of the MCMC algorithm. We simulated data using the same data structure as in the NIEHS uterine fibroid study (e.g., same number of African Americans and Whites, same screening ages, same number

of postmenopausal women), but with new values for the response variables (Y, D, Z_1, Z_2). We simulated 10 independent data sets under our proposed model, with parameter values chosen to be consistent with the prior and with no differences between African Americans and Whites in incidence, diagnosis, or progression rates. We then used our code to implement the MCMC analysis separately for each simulated data set, collecting 10,000 iterates after an initial burn-in of 5,000 iterates in each case.

Across the 10 simulated data sets, the estimated $\Pr(\bar{\kappa} > 0)$ ranged from 0.04 to 0.94, with a mean of 0.43 and a median of 0.48. The estimate of this probability for the real data (> 0.99) fell outside this range, adding to our confidence in the reported difference between African Americans and Whites. In addition, the estimated posterior means for $\tau_{11}, \tau_{21}, \nu_1, \kappa_1, \tau_{12}, \tau_{22}, \nu_2, \kappa_2$ and σ (the parameters reported in Table 3) deviated from their true values by an average of 0.00, 0.00, 0.00, 0.00, -0.03, 0.00, 0.00, -0.01, and -0.01, respectively. Therefore, the posterior distributions are reasonably centered on the target parameter values, suggesting that our MCMC implementation works.

5. Discussion

Ideally, factors related to disease onset and progression would be investigated using longitudinally collected data from a prospective cohort study. However, such studies are extremely expensive and time consuming to conduct, and have the important drawback that individuals diagnosed with disease must (for ethnical reasons) be treated; disturbing the natural disease progression process (i.e., if an effective treatment option is available). An alternative approach is to obtain a cross-sectional sample of individuals from the population of interest, and to collect information on clinical history and current presence and severity of disease for these individuals. One can then use these data to investigate factors related to disease incidence and progression, after preclinical onset and prior to clinical diagnosis, by applying the approach proposed in this article.

As we have demonstrated through application to data from a National Institute of Environmental Health Sciences study of uterine leiomyoma, this approach can lead to important insights that would not have been possible using earlier methods. In particular, we found that the rate of growth of uterine leiomyoma within the preclinical stage of disease is higher for African Americans than whites; a result that has important public health, clinical and biological implications. This difference would not have been apparent from analyses that compared the current severity of leiomyoma between African American and white women without adjusting for differences in incidence between these groups.

Due to identifiability concerns, we have not assessed the effect of disease severity on the rate of diagnosis, and we have focused on the rates of progression conditional on remaining in the preclinical phase of disease. Our model does not require the diagnosis rates to be independent of disease severity. However, if a high level of severity tends to lead to symptoms that result in clinical diagnosis, then inferences about differences between groups in progression may be conservative due to the conditional interpretation of the progression rates. Under this scenario, individuals with high rates of progression would be diagnosed more quickly and hence less likely to be in the preclinical state at screening than individuals with low rates of progression. Thus, if anything, we may have under estimated the magnitude of the difference between African Americans and whites.

Our approach has several appealing characteristics. First, by using semiparametric regression models for the age-specific transition rates characterizing the disease onset and diagnosis process, we avoid restrictive parametric assumptions while allowing the incorporation of general covariate effects. A prior distribution on the degree of local nonlinearity in the baseline rates facilitates smoothing, and allows models to be fit that are not estimable by unrestricted maximum likelihood procedures that use nonparametric step functions for the baseline distributions. We have found the resulting estimates to be robust to the choice of the prior parameters. Our approach incorporates disease severity information through a

flexible underlying variable model that allows multiple discrete and continuous surrogates of the current severity of disease. This underlying variable approach allows covariate effects distinct to the different surrogates, and accommodates heterogeneity among women in disease progression through a factor analytic term.

REFERENCES

- ACOG (1994). ACOG technical bulletin. *International Journal of Gynecology and Obstetrics* **46**, 73-82.
- Arminger, G. and Küsters, U. (1988). Latent trait models with indicators of mixed measurement level. In *Latent Trait and Latent Class Models*. (eds. R. Langeheine and J. Rost), 51-73. New York: Plenum.
- Baird, D. D., Dunson, D. B., Hill, M. C., Cousins, D., and Schectman, J. M. (2001). High cumulative incidence of uterine leiomyoma in African American and white women. *American Journal of Obstetrics and Gynecology*, submitted.
- Chen, T. H. H., Kuo, H. S., Yen, M. F., Lai, M. S., Tabar, L., and Duffy, S. W. (2000). Estimation of sojourn time in chronic disease screening without data on interval cases. *Biometrics* **56**, 167-172.
- Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association* **91**, 883-904.
- Craig, B. A., Fryback, D. G., Klein, R., and Klein, B. E. K. (1999). A Bayesian approach to modelling the natural history of a chronic condition from observations with intervention. *Statistics in Medicine* **18**, 1355-1371.
- De Gruttola, V. G. and Lagakos, S. W. (1989). Analysis of doubly-censored survival data, with application to AIDS. *Biometrics* **45**, 1-11.

- Dinse, G. E., and Lagakos, S. W. (1982). Nonparametric estimation of lifetime and disease onset distributions from incomplete observations. *Biometrics* **38**, 921-932.
- Duffy, S. W., Chen, H. H., Tabar, L., and Day, N. E. (1995). Estimation of mean sojourn time in breast cancer screening using a Markov chain model of both entry to and exit from the preclinical detectable phase. *Statistics in Medicine* **14**, 1531-1543.
- Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society B* **63**, 355-366.
- Dunson, D. B. and Baird, D. D. (2001). A flexible parametric model for combining current status and age at first diagnosis data. *Biometrics* **57**, 396-403.
- Dunson, D. B. and Dinse, G. E. (2001). Bayesian incidence analysis of animal tumorigenicity data. *Applied Statistics* **50**, 125-141.
- Frydman, H. (1992). A nonparametric estimation procedure for a periodically observed three state Markov process, with application to AIDS. *Journal of the Royal Statistical Society, Series B* **54**, 853-866.
- Gamerman, D. (1991). Dynamic Bayesian models for survival data. *Applied Statistics* **40**, 63-79.
- Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398-409.
- Gelfand, A. E. and Wang, F. (2000). Modelling the cumulative risk for a false-positive under repeated screening events. *Statistics in Medicine* **19**, 1865-1879.
- Gilks, W. R., Richardson, S. and Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

- Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* **41**, 337-348.
- Gilks, W. R., Wang, C. C., Yvonnet, B. and Coursaget, P. (1993). Random-effects models for longitudinal data using Gibbs sampling. *Biometrics* **49**, 441-453.
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97-109.
- Kjerulff, K.H., Langenberg, P., Seidman, J.D., Stolley, P.D., and Guzinski, G.M. (1996). Uterine leiomyomas: Racial differences in severity, symptoms and age at diagnosis. *The Journal of Reproductive Medicine* **41**, 483-490.
- Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika*, **49**, 115-132.
- Rai and Matthews (1995). The analysis of incomplete data using stochastic covariates. *Canadian Journal of Statistics* **23**, 29-42.
- Ryan and Orav (1988). On the use of covariates for rodent bioassay and screening experiments. *Biometrika* **75**, 631-637.
- Sinha, D. (1998). Posterior likelihood methods for multivariate survival data. *Biometrics* **54**, 1463-1474.
- Smith, B. J. (2000). *Bayesian Output Analysis Program (BOA) Version 0.5.0 User Manual*, Department of Biostatistics, The University of Iowa College of Public Health.
- Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82**, 528-550.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics* **22**, 1701-1762.

Turnbull, B. W. and Mitchell, T. J. (1984). Nonparametric-estimation of the distribution of time to onset for specific diseases in survival sacrifice experiments. *Biometrics* **40**, 41-50.

APPENDIX

Conditional Distributions for MCMC Algorithm

Step 1: Sample the interval of entry into state 2 for each individual with $R_i \leq T_i$ from its full conditional posterior distribution. If individual i is in state 3 at T_i , then

$$\Pr(R_i \in I_j | S_i \leq T_i, -) = \frac{1(j \leq S_i) \left\{ \prod_{h=1}^{j-1} (1 - \lambda_{ih}) \right\} \lambda_{ij} \left\{ \prod_{h=j}^{l: S_i \in I_{l+1}} (1 - \alpha_{ih}) \right\}}{\sum_{m=1}^{l: S_i \in I_l} \left\{ \prod_{v=1}^{m-1} (1 - \lambda_{iv}) \right\} \lambda_{im} \left\{ \prod_{v=m}^{l: S_i \in I_{l+1}} (1 - \alpha_{iv}) \right\}},$$

for $j = 1, \dots, J$. If individual i is in state 2 at T_i , then

$$\begin{aligned} \Pr(R_i \in I_j | R_i \leq T_i, S_i > T_i, Z_{ik}^*, k = 1, \dots, q, -) \\ = \frac{1(j \leq T_i) \left\{ \prod_{h=1}^{j-1} (1 - \lambda_{ih}) \right\} \lambda_{ij} \left\{ \prod_{h=j}^{l: T_i \in I_l} (1 - \alpha_{ih}) \right\} \left\{ \prod_{k=1}^q f_{ijk}(Z_{ik}^*; T_i) \right\}}{\sum_{m=1}^{l: T_i \in I_l} \left\{ \prod_{v=1}^m (1 - \lambda_{iv}) \right\} \lambda_{im} \left\{ \prod_{v=m}^{l: T_i \in I_l} (1 - \alpha_{iv}) \right\} \left\{ \prod_{k=1}^q f_{imk}(Z_{ik}^*; T_i) \right\}}, \end{aligned}$$

where $f_{ijk}(z^*; t)$ denotes the conditional density of the k th underlying variable.

Step 2: Sample the link parameters $\{\tau_k\}$ and underlying variables $\{Z_{ik}^*\}$ for individuals in state 2 at screening by (i) sampling a candidate value for τ_k and accepting or rejecting using an MH step with acceptance probability dependent on the candidate and prior densities and on the conditional likelihood ratio of $\{Z_{ik}^*\}$ given the candidate versus current values of τ_k (this likelihood ratio follows a simple form conditional on the onset interval imputed in Step 1); and (ii) for continuous surrogates let $Z_{ik}^* = g^{-1}(Z_{ik}; \tau_k)$ and for ordinal surrogates sample Z_{ik}^* from its conditional density given Z_{ik} and τ_k .

Step 3: Sample the parameters $\{\omega_j\}$, β , $\{\nu_j\}$, and ψ . Conditional on risk indicator variables that can be calculated based on the imputed interval of entry into state 2, the complete data likelihoods for regression models (7) and (8) follow simple Bernoulli forms. Thus, the parameters in these models can be sampled using existing Gibbs sampling methods developed for generalized linear models (e.g., Gilks and Wild, 1992) with the addition of MH steps to sample the baseline parameters under the priors described in Section 3.

Step 4: Sample the parameters: $\{\boldsymbol{v}_k, \boldsymbol{\kappa}_k\}$ and σ , and the latent variables $\{\xi_i\}$. Conditional on the imputed number of intervals spent in state 2, the model for the underlying severity variables follows a simple normal mixed model form. Thus, previous methods (e.g., Gilks et al., 1993) can be used to sample the parameters and latent variables in expression (9).

Table 1

Data structure for cross-sectional chronic disease screening studies

Current Status	State	Y	D	Onset Time	Diagnosis Time	Severity Surrogates
Disease Free	1	0	0	(T, ∞)	(T, ∞)	—
Preclinical Disease	2	1	0	$(0, T]$	(T, ∞)	Z_1, \dots, Z_q
Previously Diagnosed	3	1	1	$(0, S]$	S	— ^a

^a If available, surrogate may depend on intervention

T = age at screening

S = age at diagnosis

Table 2*Summary of uterine leiomyoma data for African American and white women*

Race	Current Status	State	Number	Surrogate Data ^a	
				Uterus length(cm)	Tumor rank ^b
Black	Leiomyoma-Free	1	130	—	—
	Preclinical leiomyoma	2	185	9.07 (2.17)	1.81 (.76)
	Leiomyoma history	3	420	—	—
	No history, missing exam	?	105	—	—
	All black		840		
White	Leiomyoma-Free	1	190	—	—
	Preclinical leiomyoma	2	140	8.62 (1.48)	1.49 (.65)
	Leiomyoma history	3	125	—	—
	No history, missing exam	?	69	—	—
	All white		524		

^a Mean (sd) among women with leiomyoma detected at screening.^b Ordinal measure of size/number of tumors was averaged.

Table 3

Posterior summaries of the parameters characterizing growth in uterine leiomyoma following preclinical onset (results from primary analysis)

Surrogate	Parameter	Posterior Summaries			
		Mean	Median	SD	90% Credible Interval
Uterine Length (Z_1)	τ_{11}	8.20	8.19	0.18	(7.91, 8.50)
	τ_{21}	1.47	1.46	0.08	(1.33, 1.61)
	ν_1	0.09	0.09	0.04	(0.03, 0.16)
	κ_1	0.07	0.07	0.04	(0.01, 0.14)
Tumor Size/ Number (Z_2)	τ_{12}	0.15	0.15	0.26	(-0.27, 0.59)
	τ_{22}	1.40	1.39	0.33	(0.88, 1.97)
	ν_2	0.03	0.02	0.03	(0.00, 0.10)
	κ_2	0.06	0.06	0.04	(0.00, 0.14)
Shared Parameter	σ	0.12	0.12	0.02	(0.09, 0.16)

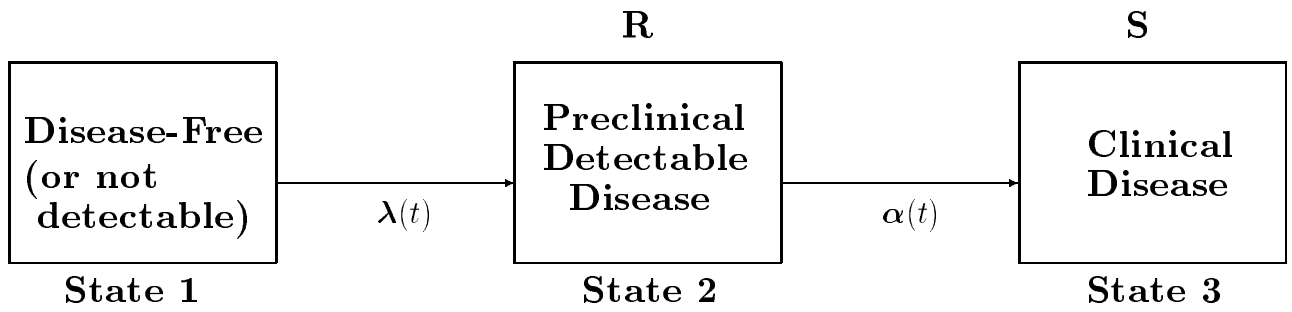


Figure 1. Progressive three-state model of the onset and diagnosis process.

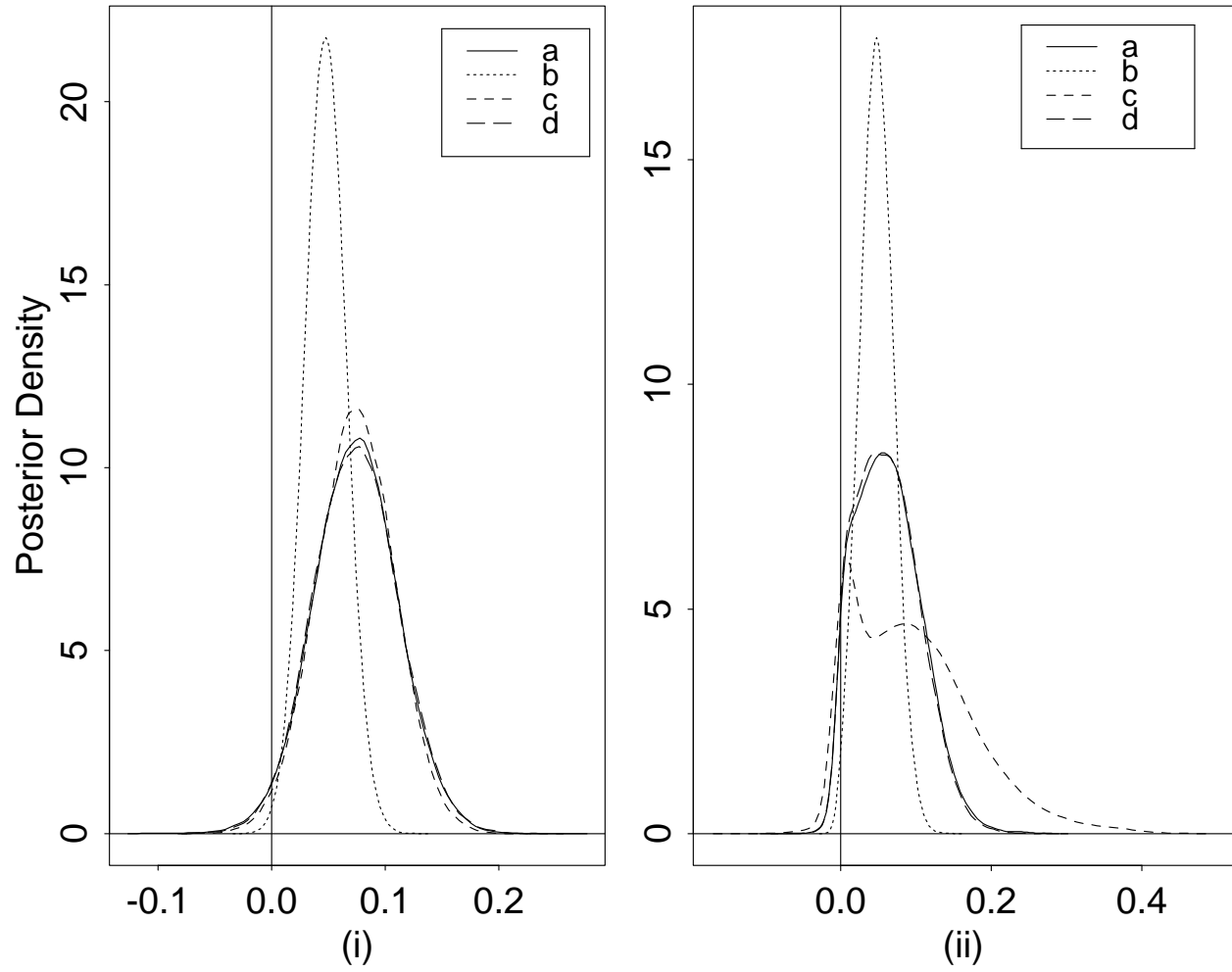


Figure 2. Estimated posterior densities for the parameters (i) κ_1 and (ii) κ_2 , which measure differences between African Americans and Whites in the rates of growth of the uterus and of focal leiomyoma tumors, respectively (a=main analysis, b=prior variance inflated, c=prior variance decreased, d=different prior mean)

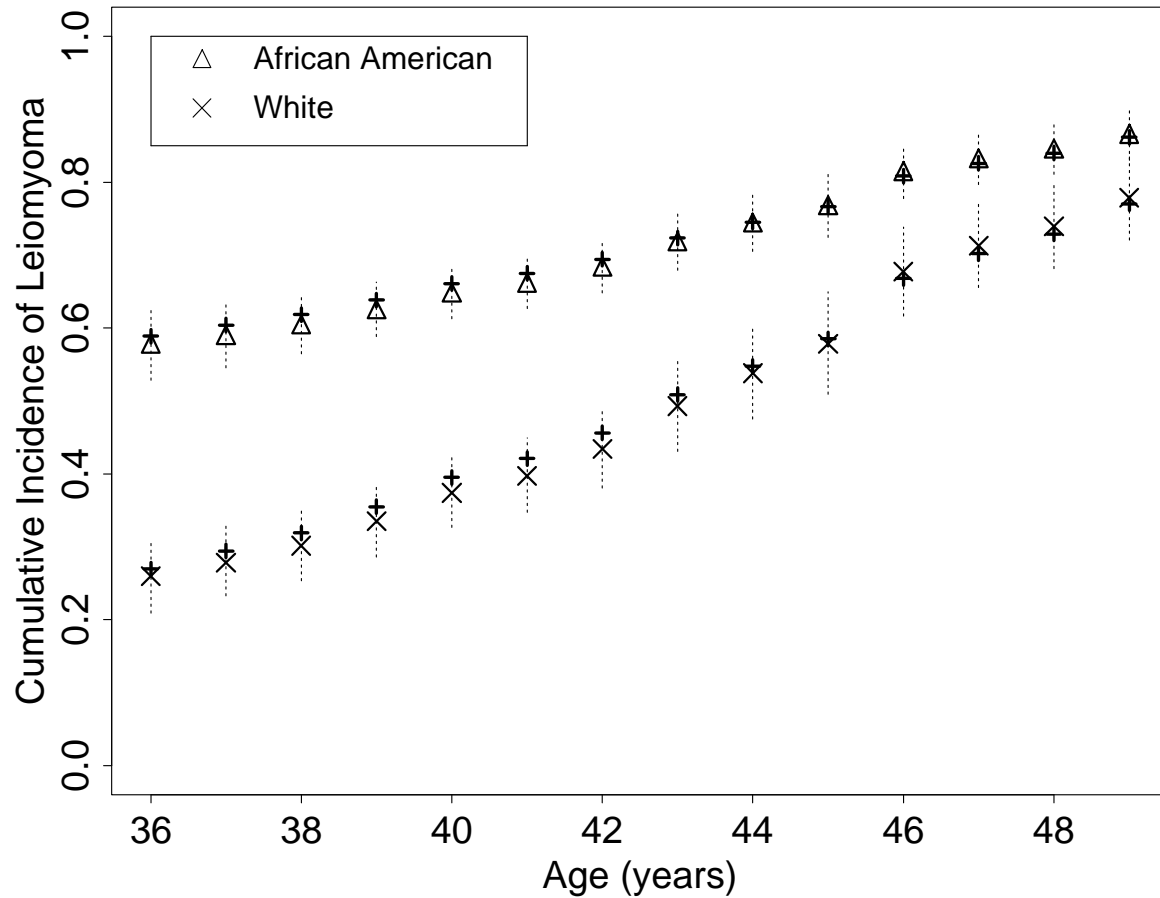


Figure 3. Estimated posterior means and 90% credible intervals for the age-specific cumulative incidence of uterine fibroids among premenopausal African American and white women. Estimates not incorporating disease severity data are denoted by +.