A Monte Carlo Method for Bayesian Inference in Frailty Models

David G. Clayton

Biometrics, Vol. 47, No. 2. (Jun., 1991), pp. 467-485.

Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28199106%2947%3A2%3C467%3AAMCMFB%3E2.0.CO%3B2-M

Biometrics is currently published by International Biometric Society.

BIOMETRICS 47, 467–485
June 1991

A Monte Carlo Method for Bayesian Inference in Frailty Models

David G. Clayton*
Department of Community Health, University of Leicester, United Kingdom

SUMMARY

Many analyses in epidemiological and prognostic studies and in studies of event history data require methods that allow for unobserved covariates or "frailties." Clayton and Cuzick (1985, *Journal of the Royal Statistical Society, Series A* **148**, 82–117) proposed a generalization of the proportional hazards model that implemented such random effects, but the proof of the asymptotic properties of the method remains elusive, and practical experience suggests that the likelihoods may be markedly nonquadratic. This paper sets out a Bayesian representation of the model in the spirit of Kalbfleisch (1978, *Journal of the Royal Statistical Society, Series B* **40**, 214–221) and discusses inference using Monte Carlo methods.

1. Introduction

The last two decades have seen major advances in methods for the statistical analysis of occurrence data. This development has been of considerable value in the analysis of clinical trials, and has done much to clarify the mathematical basis of epidemiological studies and analysis. Increasingly, the methods are being applied in other settings, and recently attention has turned to the analysis of more complicated Markov and semi-Markov models for event history data. Extension of Cox's (1972) pioneering work on the proportional hazards model for survival data to more general problems of modelling counting processes has brought advances in the analysis of such data, at least in theory (Aalen et al., 1980). However, the utility of this approach in practice is limited by the assumption of independence of transitional events conditional upon covariates—an assumption that will almost always be untrue owing to an inability to measure all relevant variables. Thus, patients who are particularly at risk from one type of state transition may also tend to be prone to another, even after conditioning upon covariates.

In epidemiology, problems of related events arise in family studies of disease incidence. There may be a tendency for disease occurrence to cluster within families, either because of shared environmental exposures or because of genetic predisposition. In animal carcinogenesis experiments, litter-matched designs are often employed in an attempt to control for genetic factors. In all these examples there is a need to introduce an extra random component into the proportional hazards model. In the first and third areas described above, this is required only to achieve correct inference on fixed effects of interest, whereas in the second case the random family effect itself is of prime interest in the analysis. Such models have attracted increasing attention in recent years and are often referred to as "frailty" models. However, they pose considerable theoretical difficulties. This paper suggests a Monte Carlo approach suggested by a Bayesian casting of the problem.

**Present address:* Medical Research Council Biostatistics Unit, 5 Shaftesbury Road, Cambridge CB2 2BW, United Kingdom.

Key words: Accident proneness; Bayesian inference; Counting processes; Event history analysis; Frailty; Genetic epidemiology; Gibbs sampling; Monte Carlo methods; Multiplicative intensity model; Proportional hazards model; Stochastic substitution; Survival analysis.

Section 2 briefly reviews frequentist approaches to inference in the proportional hazard model, and introduces some notation. Section 3 sets out a Bayesian approach to the proportional hazards model. Section 4 reviews the random frailty model of Clayton and Cuzick (1985a) and Section 5 discusses their proposed method of estimation and some alternative computational algorithms. Section 6 describes the Bayesian formulation of the frailty model as a graphical model and Section 7 discusses the Monte Carlo method of inference. Finally, the method is demonstrated using data derived from an animal carcinogenesis experiment.

2. Frequentist Inference in the Proportional Hazards Model

The proportional hazards model for survival data, introduced by Cox (1972), describes the *hazard function* for an individual characterised by covariate vector $\mathbf{z}$ by the product

$$\lambda(t \mid \mathbf{z}) = \lambda_0(t)\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}), \tag{2.1}$$

where λ_0, the *baseline hazard function*, is modelled nonparametrically. Extensions to this basic model allow alternative parametric *relative risk models* for the second term of the product, which may also depend on time, either because of variation of *time-dependent covariates* or because covariates have *time-dependent effects*. Kalbfleisch and Prentice (1980) also introduced the extension in which a different baseline hazard function could be fitted in each of several *strata*.

In later work, the proportional hazards model for survival data has increasingly been seen as a special case of a *multiplicative intensity model* for counting processes which generalizes that of Aalen (1978). If $I_{\mathbf{z}}(t)$ is the intensity process for a counting process labelled by covariate vector $\mathbf{z}$, it is modelled by

$$I_{\mathbf{z}}(t) = Y(t)I_0(t)\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}), \tag{2.2}$$

the product of an observed process, $Y(t)$, a baseline intensity process to be modelled nonparametrically, $I_0(t)$, and a (usually log-linear) parametric relative risk model for the effect of covariates. For a nontechnical review of this approach, see Gill (1984).

Owing to unwillingness to make detailed assumptions concerning censoring processes, most published work has focussed upon asymptotic inference with this model. Here, the really fruitful approach has been that of *partial likelihood* as originally proposed by Cox. Although seemingly based on standard likelihood results, it is now well known that partial likelihood is misnamed and is not, in fact, a likelihood of any sort. The proof that maximum partial likelihood estimators for the parameters, $\boldsymbol{\beta}$, of the regression model have the same asymptotic properties as maximum likelihood estimators follows elegantly by applying martingale theorems in the counting process framework (Andersen and Gill, 1982).

Earlier approaches to justifying the idea of partial likelihood included that of Kalbfleisch and Prentice (1973), who pointed out that, with the assumption of *progressive Type* II *censoring* and in the absence of ties, partial likelihood is equivalent to the marginal likelihood based on a generalized rank vector for (censored) survival times.

Several authors, notably Breslow (1974) and Johansen (1983), have pointed out that (again in the absence of ties) partial likelihood may be viewed as a *profile likelihood* in which the unknown nuisance function, $\lambda_0(t)$ [or $I_0(t)$], is replaced in the total likelihood by a nonparametric maximum likelihood estimate. In the counting process formulation (2.2), for known $\boldsymbol{\beta}$ the intensities for the observed counting processes are products of the unknown process, $I_0(t)$, and observed processes. The nonparametric maximum likelihood estimate of the (integrated) unknown process is given by the Aalen estimator (Aalen, 1978). For right-censored survival data in the absence of covariates this estimator is identical to the nonparametric estimator for the integrated hazard function introduced by Nelson (1969) and (independently) by Altshuler (1970).

For Cox's model for survival data with covariates, the Aalen estimator is equivalent to that proposed by Breslow (1974) for the integrated baseline hazard function:

$$\Lambda_0(t) = \int_0^t \lambda_0(u)\, du.$$

Substitution of the expression for the maximum likelihood estimate of this function (which itself depends on β) into the total likelihood yields the partial likelihood.

To introduce some notation for later sections, let $N_i(t)$ be the process that counts observed failure of the ith study subject [so that $N_i(t)$ takes the value 0 until the observation of failure and 1 thereafter] and let $Y_i(t)$ indicate observation of the subject at t (taking the value 1 if the subject is observed and 0 otherwise). Then the intensity process for $N_i(t)$ is

$$I_i(t) = \lambda_0(t)\, Y_i(t) \exp(\beta^{\mathrm{T}} \mathbf{z}_i). \tag{2.3}$$

Under *noninformative censoring* the likelihood factorizes, with one term depending only on the censoring process and the second term:

$$\prod_i \left[\prod_{t \geqslant 0} I_i(t) \right]^{dN_i(t)} \exp\left[- \int_{t \geqslant 0} I_i(t)\, dt \right]. \tag{2.4}$$

This is essentially Poisson in form, reflecting the fact that the likelihood may be thought of as generated by independent contributions of many data "atoms," each concerned with observation of an individual over a very short interval during which the intensity may be regarded as constant (for a review of event history analysis from this standpoint, see Clayton, 1988). Introducing the further shorthand

$$R_i(t; \beta) = Y_i(t) \exp(\beta^{\mathrm{T}} \mathbf{z}_i), \tag{2.5}$$

the Aalen estimator for $\Lambda_0(t)$ is

$$\hat{\Lambda}_0(t) = \int_0^t \frac{1}{R_+(u; \beta)}\, dN_+(u) \tag{2.6}$$

(where the + subscript denotes summation over i), and substitution of (2.6) into (2.4) yields the familiar partial likelihood:

$$\prod_{t \geqslant 0} \prod_i \left[\frac{R_i(t; \beta)}{R_+(t; \beta)} \right]^{dN_i(t)}. \tag{2.7}$$

This discussion assumes no ties of observed failure or censoring times, and this is also true of the Bayesian approach described below. The occurrence of ties is awkward both theoretically and computationally and this paper will assume that observations are made with sufficient precision to preclude ties.

3. Bayesian Approaches

Several authors have discussed Bayesian inference for such models (Cornfield and Detre, 1977; Kalbfleisch and MacKay, 1978; Kalbfleisch, 1978; Kalbfleisch and Prentice, 1980). Kalbfleisch (1978) proposed a Bayesian argument that degenerates into partial likelihood theory with suitable choice of vague priors. This approach is nonparametric with respect to Λ_0 and is summarised below. In the interests of clarity, no attempt will be made to deal with the measure-theoretic difficulties arising out of the fact that Λ_0 lies in an infinite-dimensional space. Kalbfleisch dealt with this problem by assuming measurements to have been made on a fine time grid, so that the

unknown function Λ_0 is a vector with a large but finite number of elements. This also makes for some complications, and the argument below lets the time grid interval approach zero.

Kalbfleisch assumed that each component of β has the improper uniform prior on $[-\infty, +\infty]$, and that the prior for Λ_0 is a process with

$$\mathrm{E}[\Lambda_0(t)] = \Lambda^*(t), \quad \text{a known function},$$

$$\mathrm{var}[\Lambda_0(t)] = \frac{\Lambda^*(t)}{c},$$

and the increments $d\Lambda_0(t)$ are distributed as independent gamma variates with shape and scale parameters $cd\Lambda^*(t)$ and c, respectively. That is,

$$d\Lambda_0(t) \sim G(cd\Lambda^*(t), c).$$

For this process, any difference $\Lambda_0(t_i) - \Lambda_0(t_0)$ $(t_i > t_0)$ is also a gamma variate.

Kalbfleisch pointed out that the independent-increments prior for Λ_0 is not a very satisfactory representation of prior beliefs about baseline hazard functions. However, it is adequate when the main interest is in the regression parameters β, and has the virtue of being the conjugate prior for the total likelihood for censored survival data and, more generally, for counting processes. Thus, conditional upon β, the posterior for Λ_0 is another independent-increments gamma process with

$$d\Lambda_0(t) \sim G(cd\Lambda^*(t) + dN_+(t),\ c + R_+(t;\beta)), \tag{3.1}$$

so that

$$\mathrm{E}[\Lambda_0(t)] = \int_0^t \frac{c}{c + R_+(u;\beta)}\, d\Lambda^*(u) + \int_0^t \frac{1}{c + R_+(u;\beta)}\, dN_+(u) \tag{3.2}$$

and

$$\mathrm{var}[\Lambda_0(t)] = \int_0^t \frac{c}{[c + R_+(u;\beta)]^2}\, d\Lambda^*(u) + \int_0^t \frac{1}{[c + R_+(u;\beta)]^2}\, dN_+(u). \tag{3.3}$$

Note that, for this process, the differences $\Lambda_0(t_i) - \Lambda_0(t_0)$ are not, in general, gamma variates. Since $R_+(t)$ is a step function, they are distributed as sums of gamma variates with different scale factors.

The Bayes estimator of Λ_0 (given β) is the posterior expectation, (3.2), and is a "shrinkage" estimator that effects a compromise between the Aalen (maximum likelihood) estimate and the prior mean, Λ^*. When c is large, the prior belief in Λ^* is strong and the estimate approaches this function, whereas, when c is small, the prior belief is weak and the estimate approaches the Aalen estimate. It is assumed in this approach that the scientist can supply a value of c to represent the strength of prior belief.

The marginal posterior distribution for β is given by integrating the joint posterior over Λ_0 and is proportional to

$$\frac{\prod_{t\geq 0} \prod_i [R_i(t;\beta)]^{dN_i(t)}}{\prod_{t\geq 0} [c + R_+(t;\beta)]^{cd\Lambda^*(t) + dN_+(t)}}. \tag{3.4}$$

With the improper vague prior for Λ_0 defined by $c = 0$, the posterior for β is proportional to the partial likelihood. When $c \to \infty$, so that Λ_0 is known a priori, the analysis reduces to a log-linear exponential/Poisson regression analysis with likelihood as if the censoring indicators

were Poisson variates with expectations

$$\int_{t\geqslant 0} R_i(t;\beta)\, d\Lambda^*(t) = \exp(\beta^{\mathrm{T}}\mathbf{z}) \int_{t\geqslant 0} Y_i(t)\, d\Lambda^*(t).$$

There is little difficulty in principle in extending this argument to most of the extensions of the proportional hazards model. For time-dependent covariate or time-dependent effect models things become difficult because $R_+(t)$ is no longer a step function and the posterior process for Λ_0 becomes intractable except when $c = 0$. This is not surprising: it is the parametric component for the model for Λ_0 that causes the difficulty, and fully parametric models with time-dependent covariates or effects lead to difficulties for all approaches.

More recently, Gamerman and West (1987) have considered a "dynamic" model in which the regression coefficients, β, are allowed to vary with time. They divide the time scale into discrete intervals and model hazard by step functions. The priors for the step heights in the baseline hazard and for the values of β are taken as first-order autoregressive processes.

4. Frailty Models

Motivated by epidemiological studies of disease occurrence in families, litter-matched carcinogenesis experiments, and by studies of sojourn times of the same individual in different states in prognostic studies, Clayton and Cuzick (1985a) introduced a further generalization of the proportional hazards model that allowed for positive association of survival times. This model is a semiparametric generalization of other work that allowed for a random effect, or "frailty," in the hazard model (see, for example, Vaupel, Manton, and Stallard, 1979; Lancaster and Nickel, 1980). Using the language of animal experimentation, the hazard function for an animal from litter l with covariate vector $\mathbf{z}$ is

$$\lambda(t\,|\,\mathbf{z}, l) = \lambda_0(t)\xi_l \exp(\beta^{\mathrm{T}}\mathbf{z}), \tag{4.1}$$

where $\{\xi_l\}$ are random multiplicative effects, or "frailties," shared by all members of the same litter. The frailties are assumed to be i.i.d. gamma variates with mean 1 and variance γ. That is,

$$\xi_l \sim G(\gamma^{-1}, \gamma^{-1}). \tag{4.2}$$

Clayton and Cuzick also considered a "stratified" form of the model that allowed for several different baseline hazard functions. Thus, in family studies, different baseline hazard functions may be assumed for fathers and for sons.

The choice of the gamma distribution for frailties arises partly for mathematical convenience (since this distribution is conjugate to the likelihood for $\{\xi_l\}$) and partly because, in the case of bivariate survival without covariates, integration over the unknown frailty yields a class of bivariate survival time distributions with appealing properties (Clayton, 1978; Cox and Oakes, 1984). This choice has some less desirable consequences, however. The *marginal* relationship between hazard and covariates no longer follows the proportional hazards model; instead there is *convergence* of hazards, at a rate determined by γ. Clayton and Cuzick (1986) exploit this fact in a univariate semiparametric survival model. In the multivariate case with covariates this property has the consequence that information for estimation of γ comes partly from the coincidence of failure within families, and partly from marginal convergence of hazards in relation to covariates. Hougaard (1986a, 1986b) has pointed out that this is not a desirable property for the multivariate model since it renders interpretation of γ difficult, and instead showed that the assumption of a positive stable distribution of the $\{\xi_l\}$ avoids the problem, since the proportional hazards assumption for covariates then remains true marginally. While the validity of this criticism is accepted, the gamma assumption will continue to be made in this

paper since it simplifies the computations somewhat. However, there is nothing in principle to prevent the implementation of Hougaard's suggestion in the current approach. A recent review and discussion of frailty models for survival data are given by Aalen (1988).

From the Bayesian standpoint, the random frailty models estimate the unknown frailties under the assumption of a priori *exchangeability*. From this standpoint the frailty variance, γ, would be termed a *hyperparameter* and prior knowledge concerning its value would be encapsulated in a *hyperprior* distribution.

5. Clayton and Cuzick's EM algorithm

With the gamma distributional assumption for frailties, it is simple to integrate over the $\{\xi_l\}$ to obtain a probability of observed data given β, γ, and Λ_0. The problem then is to deal with the infinite-dimensional nuisance parameter Λ_0. The idea of partial likelihood does not carry over in a simple manner, since the integration over frailties destroys the simple way in which likelihood unfolds over time. This in turn means that the usual martingale theory is no longer applicable. Instead, Clayton and Cuzick went back to Kalbfleisch and Prentice's (1973) idea of marginal likelihood based on generalized rank vectors. With known Λ_0, the (possibly censored) survival times, $\{t_i\}$, may be transformed into scores $\{y_i = \Lambda_0(t_i)\}$ that obey a simple parametric model depending only on the parameters of interest, β and γ. Clayton and Cuzick's proposal was to maximize the marginal likelihood of ranks using an EM algorithm, as follows:

E-step:

Estimate scores (termed "bar-bar scores") which are expectations of the transformed times, $\{y_i\}$, given β, γ, and the vector carrying the information about the ranks of the uncensored survival times and the censoring pattern.

M-step:

Maximize the parametric likelihood for β and γ given the imputed "data", $\{y_i\}$.

Note that the E-step is in a sense equivalent to the estimation of Λ_0 for known β and γ. Unfortunately, however, the computation of these expectations is intractable. Clayton and Cuzick proposed two approximate scoring algorithms, one of which (their "method 2") was as follows:

E1: With the current estimate of Λ_0 (and of β and γ), calculate "imputed frailties," which are posterior expectations of $\{\xi_l\}$.

E2: Treating the frailties as known, obtain a new estimate of Λ_0 using the Aalen–Breslow estimate (2.6). This requires a slight extension of the notation (2.5) to include the frailties:

$$R_i(t;\beta,\xi) = Y_i(t)\xi_{l(i)}\exp(\beta^{\mathrm{T}}\mathbf{z}_i), \tag{5.1}$$

where $l(i)$ denotes the litter to which the ith subject belongs. Finally, use this estimate to transform the observed times into the new scores.

Steps E1 and E2 are iterated until convergence.

Gill (1985) pointed out that, with this approximate scoring method, the algorithm maximizes a *profile* likelihood obtained by nonparametric maximum likelihood estimation of Λ_0, since E2 is a maximization step in a subsidiary EM algorithm. If there are L litters, the likelihood that is maximized is an L-fold integral of a counting process likelihood of the form of (2.4) with respect to the prior distribution of the frailties $\{\xi_l\}$.

The heuristic arguments of Clayton and Cuzick failed to prove that the approximate scores are close enough to the correct scores to ensure that the estimators of β and γ have the asymptotic properties of maximum likelihood (ML) estimators. This remains unproven and is not attempted

here. It should also be noted that, in practice, likelihoods are often markedly nonquadratic with respect to gamma; indeed they are often monotonically decreasing. In such circumstances asymptotic results are unlikely to be useful. Instead, a Bayesian analysis in the spirit of Kalbfleisch (1978) is presented.

Before moving on to the Bayesian version of the model, it is interesting to note that the maximization effected in the M-step in the Clayton and Cuzick algorithm above may be achieved in a subsidiary EM algorithm, which also involves imputation of frailties. The parametric likelihood maximized in this step is that for a model in which, conditional upon the frailties ξ, the transformed (right-censored) times $\{y_i\}$ obey an exponential log-linear regression model in which the log-frailties are the additive effects of "litter." The gamma distribution of frailties is the conjugate prior for this likelihood and the probability distribution may be explicitly integrated over the entire range of ξ to yield a likelihood that depends on γ, β, and $\{y_i\}$. In Clayton and Cuzick (1985a) this likelihood is numerically maximized in each M-step. However, this likelihood itself may be regarded as a missing data likelihood in which the frailties ξ are the missing data, and this could be maximized using an EM algorithm with E and M steps as follows:

ME-step:

1. Estimate the frailties by posterior expectations given data and current values of all other parameters (this is equivalent to E1 above).

2. Estimate the posterior expectation of their sum and of the sum of their logarithms.

MM-step:

1. Compute ML estimates of β assuming all other parameters to be known. This is an exponential regression analysis.

2. Estimate the parameter, γ, of the gamma distribution of frailties by maximum likelihood using the quantities calculated in ME2 in place of the usual sufficient statistics for the parameters of the gamma distribution.

With this expansion and some rearrangement the complete algorithm may be written as a recursion with the following steps:

Step (a) (E2): Estimate Λ_0 given β, ξ.
Step (b) (MM1): Estimate β given Λ_0, ξ.
Step (c) (E1/ME1): Estimate ξ given Λ_0, β, and γ.
Step (d) (ME2 and MM2): Estimate γ given Λ_0, β.

Alternative iterative methods are possible. For example, from results given by Clayton and Cuzick (1985b), it follows that if at each stage steps (a) and (b) are iterated until convergence, the effect is to maximize the partial likelihood for the multiplicative intensity model in which the frailties are known constants. Thus steps (a) and (b) could be combined:

Step (ab):

Estimate β by its maximum partial likelihood estimate assuming ξ known, and estimate Λ_0 using the Aalen estimator.

This variant of the method might be expected to converge faster, since step (ab) maximizes the likelihood with respect to both β and Λ_0. However, each step would involve more computation.

The algorithms for computing point estimates that have been described in this section are relaxation methods involving successive substitution of current estimates into a series of estimating equations. Interval estimates are harder to find, however. In the following sections it is shown how stochastic relaxation methods, originally developed for problems in Bayesian image analysis, can be used in a Monte Carlo method for obtaining Bayes interval estimates with this model.

6. Bayesian Inference in the Clayton and Cuzick Model

The model (4.1) and (4.2) is already a hierarchical model. To specify it as a fully Bayesian hierarchical model it is necessary only to specify priors for the parameters Λ_0 and β and for the hyperparameter γ. It would also be possible to introduce hierarchical modelling for both Λ_0 and β but that is not attempted here. Since the aim is to extend the approach of Kalbfleisch (1978), the priors described in Section 3 will be adopted, together with the assumption that γ^{-1} has, a priori, a $G(\eta, \mu)$ distribution. For the usual vague improper prior for a variance, we take $\mu = \eta = 0$.

A graphical representation of the model will prove useful in the next section, and Figure 1 shows the model represented by a directed acyclic graph connecting the hyperparameters, the parameters, and the observed data. In the Bayesian formulation all these are random variables and the directed graph represents a conditional argument for parameters conditional upon hyperparameters and for data conditional upon parameters. Such a set of relationships may also be represented by a *conditional independence graph*, which is an undirected graph constructed so that if two variables, U and V, are connected only via a third variable W, then U and V are conditionally independent given W. This graph is sometimes termed the *moral graph* since its structure may be deduced from the directed graph by a process of "marrying the parents": if U and V both have directed links to W in the directed graph, then U and V must be joined in the (undirected) conditional independence graph. Applying this process to the directed graph for the frailty model, Figure 1, yields the conditional independence graph shown in Figure 2. This shows, for example, that, conditional upon the frailties ξ, the hyperparameter γ is independent of β, Λ_0, and the observed data.

The graphical representation of multivariate distributions is well known in connection with models for high-dimensional contingency tables (Wermuth and Lauritzen, 1983) and for Bayesian inference in expert systems (Lauritzen and Spiegelhalter, 1988). Graph theory is also important in the theory of Markov random fields, which are of considerable importance in statistical mechanics and, more recently, in spatial statistics and image analysis. This theory, reviewed by Geman and Geman (1984), shows that the algebraic representation of the multivariate distribution represented by a conditional independence graph follows from the graph structure; more specifically, it depends on the set of *cliques* that make up the graph, a clique being a set of nodes in which all pairs are connected. The joint distribution corresponding to a given conditional independence graph is proportional to $\exp(-E)$, where E is the *energy function* which is a sum, over all cliques (c), of *potentials* P_c, where each potential is a function depending only on the variables contained in the corresponding clique. Such a distribution is a *Gibbs distribution*.

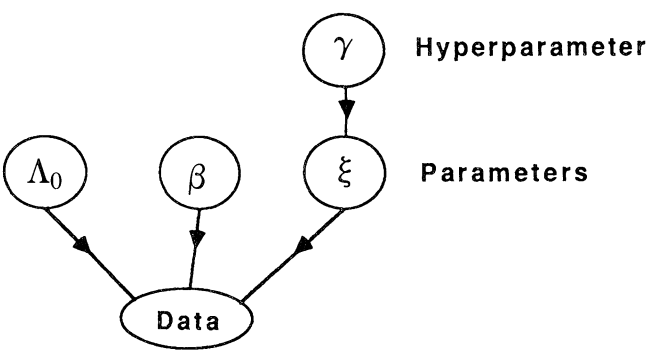

Figure 1. Directed graphical representation of the Bayesian frailty model.

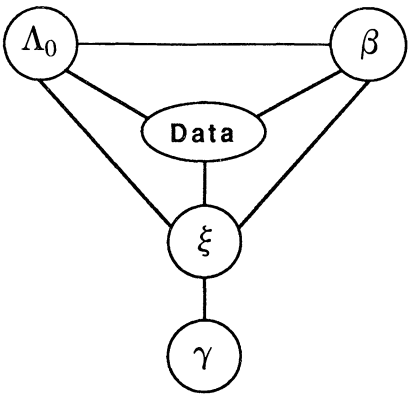

Figure 2. Conditional independence graph for the Bayesian frailty model.

The Bayesian model described above and illustrated in Figures 1 and 2 defines a Gibbs distribution with cliques $\{\Lambda_0, \beta, \xi, \text{Data}\}$ and $\{\xi, \gamma\}$. The joint probability distribution of the system is proportional to the product of the hyperprior for γ, the priors for ξ, β, and Λ_0, and the probability for the data given the multiplicative intensity model [i.e., the likelihood (2.4)]. Bayesian statistical inference requires computation of joint and marginal posterior distributions of the parameters and hyperparameters given the data. Unfortunately these are intractable, but the conditional distributions are relatively simple. In the next section it will be shown how this may be exploited in a Monte Carlo method for sampling the posterior distribution of the parameters and hyperparameters.

Denoting conditional probability density distributions by the shorthand notation $[\ldots \mid \ldots]$, the conditional distributions corresponding to the graph of Figure 2 are

$[\Lambda_0 \mid \beta, \xi, \text{Data}]$:
An independent-increments gamma process described by (3.1)–(3.3), with $R_i(t; \beta)$ redefined as in (5.1).

$[\beta \mid \xi, \Lambda_0, \text{Data}]$:
Proportional to the likelihood for the parameters of a log-linear Poisson or censored exponential regression model. For moderate sample sizes this can be closely approximated by a Gaussian distribution by matching derivatives at the mode. The calculations involved are the same as for maximum likelihood estimation using the Newton–Raphson method or the (equivalent) iteratively reweighted least squares algorithm as used in the GLIM computer program.

$[\xi \mid \Lambda_0, \beta, \gamma, \text{Data}]$:
Conditionally independent gamma variates. ξ_l is distributed as $G(D_l + \gamma^{-1}, E_l + \gamma^{-1})$, where D_l is the number of observed failures in the lth litter and

$$E_l = \sum_{i \in \text{litter } l} \int_{t \geqslant 0} Y_i(t) \exp(\beta^{\mathrm{T}} \mathbf{z}_i)\, d\Lambda_0 t.$$

$[\gamma \mid \xi]$:
$\nu = \gamma^{-1}$ is distributed with density proportional to

$$\frac{\nu^{L\nu+\eta-1} e^{-\mu\nu}}{[\Gamma(\nu)]^L} e^{\nu(\log P - S)}, \tag{6.1}$$

where L is the number of litters, S is the sum of the $\{\xi_l\}$, and P is their product. This expression is derived by multiplying the prior distribution for ν [a gamma distribution, $G(\eta, c)$] by the probability of sampling $\{\xi_l, l = 1, \ldots, L\}$ from a gamma distribution with mean 1 and variance $\gamma = \nu^{-1}$.

7. Stochastic Relaxation and Gibbs Sampling

This section describes a Monte Carlo method for generating samples from the joint posterior distribution of the parameters of the model. At first sight the method seems to be closely related to "bootstrap" methods for interval estimation, but there are important differences. Whereas in a frequentist representation of the problem the unknown parameters are regarded as fixed constants and the data values are random variables, the Bayesian analysis reverses the status of data and parameters: the data are fixed constants and the parameters are random variables. The frequentist bootstrap regenerates multiple sets of data and reanalyses each bootstrapped data set in an attempt to explore estimation errors. In contrast, a Monte Carlo Bayesian approach holds the observed data constant, and samples repeatedly from the posterior distribution of parameters given data.

Recently a number of authors have drawn attention to methods for sampling multivariate joint and marginal posterior distributions when only conditional distributions are available. The common feature of these methods is that the sampling is carried out by a stochastic process whose equilibrium distribution is that required. The idea has its roots in the Metropolis algorithm of statistical mechanics but its wider relevance seems first to have been pointed out by Hastings (1970). It has remained largely unknown except in applications in spatial statistics involving Markov random field models (Ripley, 1977). The recent recognition of its widespread applicability for Bayesian statistical inference follows the imaginative work of Geman and Geman (1984) in Bayesian image analysis. A comprehensive review of this and other Monte Carlo methods for sampling posterior densities is given by Gelfand and Smith (1990) and their application in a range of Bayesian analyses with Gaussian assumptions is illustrated by Gelfand et al. (1990).

Consider the case of two (possibly vector-valued) random variables, U and V, when we wish to draw a sample from the unavailable joint distribution $[u, v]$. However, the conditional distributions $[u \mid v]$ and $[v \mid u]$ are known. If we can sample V from the marginal $[v]$, and then sample $U \sim [u \mid v = V]$, then the pair (U, V) is a sample from $[u, v]$. Conversely, if we could sample $U \sim [u]$ we would be able to proceed to sample $V \sim [v \mid u = U]$, thus again obtaining a sample $(U, V) \sim [u, v]$. Argument by analogy with relaxation methods for the solution of linear simultaneous equations suggests a *stochastic substitution* method in which, starting from (U^0, V^0), we generate a sequence (U^i, V^i), $i = 1, 2, \ldots$ as follows:

Step i(a): Generate $U^i \sim [u \mid v = V^{i-1}]$.

Step i(b): Generate $V^i \sim [v \mid u = U^i]$.

Using theorems reviewed by Gelfand and Smith, it may be shown under rather weak conditions that the sequence, which is Markovian, converges to an equilibrium distribution which is the joint distribution $[u, v]$. In the simplest case we might be interested only in the marginal distribution of U and here we note that the sequence $U^0, U^1, U^2, \ldots$ is a Markov process with transition function

$$K(U', U) = \int [U' \mid v][v \mid U]\, dv.$$

An important condition for the success of the method is that the transition function should be strictly positive. If it is ever zero then an absorbing state is reached and further sampling is

precluded. In practice there may be difficulties if there are regions in which the transition probabilities are very small, since many attempts at escape may be necessary.

In applications in Bayesian statistics, U and V may represent two sets of parameters and the required posterior distributions are conditional upon the observed data. A simple case is the three-level hierarchical model in which the distribution of data depends on a large number of parameters, U, which themselves have a distribution with hyperparameters V. For example, U may be a vector of location parameters for the distributions of observed data points and V may define the prior mean and variance of the elements of U. Sampling the joint posterior of parameters and hyperparameters by stochastic substitution is straightforward but runs into a serious difficulty in that there is an absorbing state in which all elements of U are equal and the element of V corresponding to the prior variance is zero. The algorithm cannot recover from this state and, perhaps just as seriously, may take a very long time to escape from the region of very small prior variance. In practice, therefore, the absorbing range of values for the variance hyperparameter is rather larger than the single value zero, and there is a finite probability of reaching this region. Indeed, experience suggests that the process will frequently reach this state quite quickly unless steps are taken to prevent it from doing so. A related difficulty is that there may be very strong autocorrelation and cross-correlation in the sequences, particularly for variance hyperparameters.

One solution to this problem is to adopt a prior distribution that renders impossible variances smaller than a certain limit. This prevents absorption, but the process may still become stuck in a region of small variance for a considerable period and the sequence of samples may show substantial autocorrelation. If we require a set of *independent* samples, this necessitates discarding the samples intermediate between those used to estimate the distributions. This increases the computation time, perhaps by many times. Hastings (1970) pointed out that, in order to estimate expectations of functions of the random variables by the arithmetic mean of the sampled values, the sample points are not required to be independent. However, in the presence of autocorrelation, a larger number of samples will be required for accurate estimation.

An alternative strategy is suggested by the work of Tanner and Wong (1987), who proposed a method for Monte Carlo estimation of posterior distributions that was inspired by the EM algorithm and which they termed the IP (imputation–posterior) algorithm. This turns out to be very closely related to the simple Markov method outlined above. The samples, (U^i, V^i), are generated in blocks of size B and the algorithm to generate the ith sample in each block is as follows:

Step i(a): Sample V^* at random from the values obtained in the *previous* block.

Step i(b): Sample $U^i \sim [u \mid v = V^*]$.

Step i(c): Sample $V^i \sim [v \mid u = U^i]$.

This method leads to independent samples within blocks. Tanner and Wong suggest that the block size, B, should be chosen adaptively, starting with a small value and increasing gradually to the required sample size.

Although this method is effective in reducing autocorrelation, it is costly in terms of discarded samples and slightly cumbersome to program. A convenient compromise involves using a stack to store the last B values of V, and starts the ith step by sampling at random from this stack. This *buffered* stochastic substitution algorithm is identical to the IP algorithm except that the sampled values are no longer blocked and the first step for the generation of the ith sample becomes

Step i(a): Sample V^* at random from $V^{i-1}, V^{i-2}, \ldots, V^{i-B}$.

Small values of B are ineffective in reducing autocorrelation and protecting against absorption, but large values of B slow convergence to the equilibrium distribution. It seems sensible to

increase B during sampling in the same way as was suggested by Tanner and Wong, but it is not necessary to increase B to the eventual sample size. Experience suggests that a quite modest stack size will effectively eliminate autocorrelation. Given routines to manage and randomly sample a push-down stack, addition of buffering to the simple stochastic substitution is a trivial programming exercise.

For more than two sets of variables, the basic idea of stochastic substitution may be extended in various ways according to the availability of conditional distributions. The simplest and most attractive is the *Gibbs sampler* algorithm of Geman and Geman (1984), which generates random samples from the Gibbs distributions which were defined in Section 6. The Gibbs sampler algorithm visits each node of a conditional independence graph and generates a value from the conditional distribution of the corresponding random variable given the current values of all its neighbours. Geman and Geman showed that, under similar conditions to those for simple stochastic substitution, the algorithm converges to a sample from the joint distribution. This follows regardless of the order in which nodes are visited, provided each node is visited sufficiently frequently; indeed the algorithm may be implemented by parallel processing. More usually it is implemented sequentially by visiting the nodes in a repeated predetermined sequence.

Use of the simple Gibbs sampler for the graph of Figure 2 runs the same risks described above for stochastic substitution: γ is a hyperparameter describing the variance of the $\{\xi_l\}$ and $\gamma = 0$ defines an absorbing state. However, the Gibbs sampler performs extremely satisfactorily in the upper section of the graph—i.e., for sampling from $[\Lambda_0, \boldsymbol{\beta}, \boldsymbol{\xi} \mid \gamma, \text{Data}]$. The sampling algorithm proposed uses a mixture of Gibbs sampling and buffered stochastic substitution; the hyperparameter γ takes the role of V in the description of stochastic substitution above, while U is the set $(\Lambda_0, \boldsymbol{\beta}, \boldsymbol{\xi})$. Sampling of U is by G steps of the Gibbs sampler. The complete algorithm is as follows:

Step i(a): Draw γ^* at random from $\gamma^{i-1}, \ldots, \gamma^{i-B}$.

Step i(b): Repeat G times:
Assuming $\gamma = \gamma^*$, apply the Gibbs sampler to the top portion of the graph of Figure 1, corresponding to the parameters Λ_0, $\boldsymbol{\beta}$, and $\boldsymbol{\xi}$.

Step i(c): Generate γ^i given the current values of $\{\xi_l\}$.

Initially both B and G can be small, but should be increased on attainment of equilibrium and before drawing samples for analysis.

The distributions that must be sampled at each step of the algorithm are those listed in Section 6. They are, for the most part, standard distributions (Gaussian, exponential, and gamma) and methods for sampling them are described in the comprehensive review of Ripley (1987).

If the integrated baseline hazard function, Λ_0, is not a main focus of interest, it is only necessary to generate its value at specific time points corresponding to the jumps in the $Y_+(t)$ and $N_+(t)$, and this series of values is a cumulative sum of gamma variates. In the case of prior ignorance concerning Λ_0 (represented by $c = 0$), the sampled function is a step function with jumps occurring at the jump points in $N_+(t)$. The size of each jump is a unit exponential variate divided by the sum of relative risks for all individuals "at risk" at that time. This is a stochastic version of the Aalen estimator.

Sampling from (6.1) is more difficult, but needs to be done less frequently. In the work described here, a rejection method was used, after first enclosing the curve in a histogram. Alternatively, the "ratio of uniforms" method (Ripley, 1987) could be used.

The sampling algorithm is closely analogous to the EM algorithm described in Section 5, but is stochastic rather than deterministic in nature.

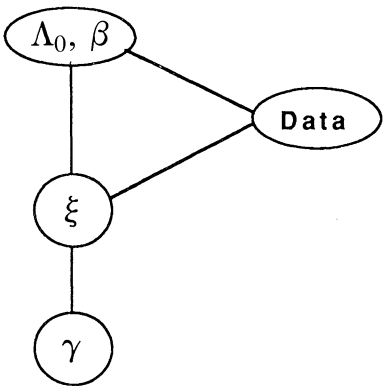

Figure 3. Simplified conditional independence graph.

An alternative Monte Carlo method parallels the modified method for point estimation introduced at the end of Section 5. It exploits the analytical result that the posterior $[\beta \mid \xi, \text{Data}]$, integrated over Λ_0, is proportional to (3.4), again with $R_i(t; \beta)$ redefined as in (5.1). This allows us to directly sample $[\Lambda_0, \beta \mid \xi, \text{Data}]$ by first sampling $[\beta \mid \xi, \text{Data}]$ and then $[\Lambda_0 \mid \beta, \xi, \text{Data}]$. Often it will be appropriate to accept an asymptotic Gaussian approximation to this distribution. When $c = 0$, this corresponds to sampling from a Gaussian distribution approximately proportional to the partial likelihood. This method corresponds to Gibbs sampling in the simplified conditional independence graph of Figure 3, and the repeated step (b) of the algorithm is

Step i(b′):

Sample from $[\Lambda_0, \beta \mid \xi, \text{Data}]$ by first sampling β from $[\beta \mid \xi, \text{Data}]$, a distribution proportional to (or approximating) (3.4), and then sampling Λ_0 from $[\Lambda_0 \mid \beta, \xi, \text{Data}]$ as before. Finally, sample ξ from $[\xi \mid \Lambda_0, \beta, \gamma, \text{Data}]$.

As in the case of the deterministic algorithm for the point estimate, this modification speeds the convergence, so that less repetition of the step is necessary, but at the expense of extra computation at each step.

The choice of starting value for the algorithm is not critical. It will often be thought desirable to compute the posterior mode as a point estimate and, if so, this provides a convenient start. Computation of the posterior mode is most conveniently achieved using one of the EM algorithms described in Section 5.

8. An Example

In this section the method is illustrated using the animal carcinogenesis data described by Mantel and Ciminera (1979) (Table 1). The experiment was litter-matched and used three rats from each of 50 litters, one rat being treated with putative carcinogen and the other two serving as controls. The time to tumour occurrence or censoring is recorded to the nearest week. The data set out in Table 1 have been slightly modified from those published by Mantel and Ciminera by randomly breaking tied uncensored failure times. Following the usual convention, ties of censored and uncensored times are treated as if failures precede censoring. In this example, there is a single covariate, z, with value 1 or 0 indicating exposure or nonexposure to carcinogen.

As a reference, the conventional "Cox" analysis was carried out using the SAS PHGLM procedure. This yielded a maximum partial likelihood estimate of β of .907, and an estimate of the asymptotic variance of .1008.

Table 1
Survival times (weeks) of one exposed (E) *and two control rats* (C1, C2) *in each of* 50 *litters*

Litter	E	C1	C2	Litter	E	C1	C2
1	101.0*	49.0	104.0*	26	89.0*	104.0*	104.0*
2	104.0*	102.0*	104.0*	27	78.0*	104.0*	104.0*
3	104.0*	104.0*	104.0*	28	104.0*	81.0	64.0
4	77.0*	97.0*	79.0*	29	86.0	55.0	94.0*
5	89.0*	104.0*	104.0*	30	34.0	104.0*	54.0
6	88.0	96.0	104.0*	31	76.0*	87.0*	74.0*
7	104.0	94.0*	77.0	32	102.8	73.0	83.9
8	95.9	104.0*	104.0*	33	101.9	104.0*	80.0*
9	82.0*	77.0*	104.0*	34	79.9	104.0*	73.0*
10	70.0	104.0*	77.0*	35	45.0	79.0*	104.0*
11	88.9	91.0*	90.0*	36	94.0	104.0*	104.0*
12	91.0*	70.0*	92.0*	37	104.0*	104.0*	104.0*
13	39.0	45.0*	50.0	38	104.0*	101.0	94.0*
14	102.9	69.0*	91.0*	39	76.0*	84.0	78.0
15	93.0*	104.0*	103.0*	40	80.0	80.9	76.0*
16	85.0*	72.0*	104.0*	41	72.0	95.0*	104.0*
17	104.0*	63.0*	104.0*	42	72.9	104.0*	66.0
18	104.0*	104.0*	74.0*	43	92.0	104.0*	102.0
19	81.0*	104.0*	69.0*	44	104.0*	98.0*	73.0*
20	67.0	104.0*	68.0	45	55.0*	104.0*	104.0*
21	104.0*	104.0*	104.0*	46	49.0*	83.0*	77.0*
22	104.0*	104.0*	104.0*	47	89.0	104.0*	104.0*
23	104.0*	83.0*	40.0	48	88.0*	79.0*	99.0*
24	87.0*	104.0*	104.0*	49	103.0	91.0*	104.0*
25	104.0*	104.0*	104.0*	50	104.0*	104.0*	79.0

*Right-censored times.

From Mantel and Ciminera (1979), with some modification: ties of uncensored failure times have been broken at random.

For the Bayesian analyses, vague priors were adopted throughout. The posterior mode of (β, γ) was calculated using the EM algorithm of Section 5, and found to be (.919, .502). Figure 4 shows 2,000 pairs, (β, γ), sampled from the joint posterior distribution using the method described in Section 6. Starting from the posterior mode, 500 cycles of the recursion were carried out before sampling began, during which the buffer size, B, was gradually increased from 1 to 100. Thereafter the buffer size remained at 100. The repetition, G, of the Gibbs sampling loop was 10 throughout. Figures 5 and 6 show the estimated marginal posterior distributions of β and γ, respectively, and Table 2 summarises these estimates. It may be seen that the inference concerning β depends only weakly on γ.

The choice of $G = 5$ was sufficient to exclude serious serial dependency in the sampled values of β; $G = 10$ was used for safety. Table 3 shows an analysis of runs above and below the median. The mean run length was $2{,}000/1{,}009 = 1.98$, which does not differ significantly from the expected value (2).

9. Discussion

There are many ways in which the proposed approach could be extended. First, it is an extremely simple matter to extend the model by introducing further random effect, or frailty, terms, a possibility that has hitherto been excluded owing to the intractability of the resultant

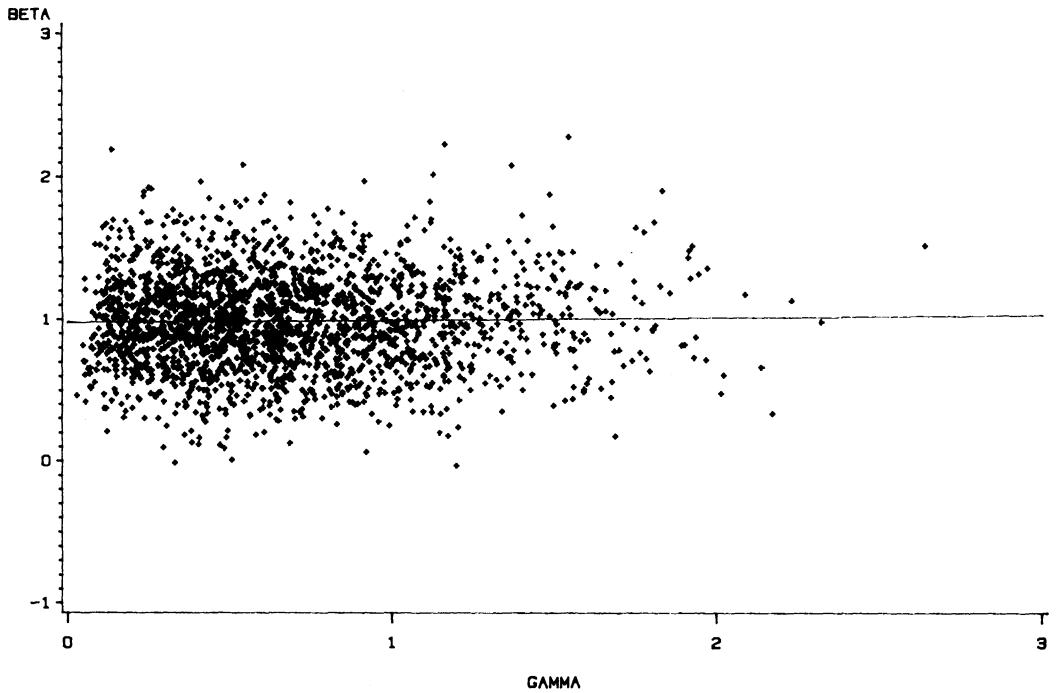

Figure 4. 2,000 simulations of (β, γ) from the joint marginal posterior distribution.

integrals. Since the conditional posterior distributions remain simple, there is no limitation to a single random effect with the proposed approach—the conditional independence graph of Figure 2 simply acquires new nodes and the sampling algorithm must be suitably extended.

Second, other models could be adopted for the prior and posterior distributions of frailties. As a purely empirical model, Hougaard's proposal of the positive stable distributions has much to recommend it, and it is possible that it could be implemented using this approach. However, the rationale of Hougaard's model is such that it has convenient properties *marginally*—that is, integrating over frailties; it does not have particularly convenient properties *conditionally*, and this makes Gibbs sampling and related methods less attractive than in the gamma frailty model.

If, on the other hand, a more explicitly biological model suggests itself, the flexibility of the method is such that this could be accommodated. The most important class of models is genetic models and frailty models with a genetic basis have been discussed by Mack, Langholz, and Thomas (1989), who also draw attention to the potential for Monte Carlo methods. A brief indication of the generalization of the proposed Monte Carlo method to such models follows.

Consider a model in which inheritance of disease proneness is mediated by a single gene. There is a single locus and let us assume two possible alleles, a and A, so that the genotype is either aa, aA, or AA. Assuming no selective mating, the prior probabilities of these possibilities are, under Hardy–Weinberg equilibrium, $(1 - q)^2$, $2q(1 - q)$, and q^2, respectively, where q is the gene frequency of allele A. Conditional upon the genotype, the frailty may be assumed to take the value 1, ϕ_1, or ϕ_2. The "inherited relative risk" parameters ϕ_1 and ϕ_2 determine the penetrance and the mode of inheritance. With this model, the analysis of epidemiological studies of disease incidence in related individuals could be approached by Gibbs sampling, since it is straightforward to generate the genotype (and hence frailty) of each individual from the posterior distribution of genotype given (i) the event history of the

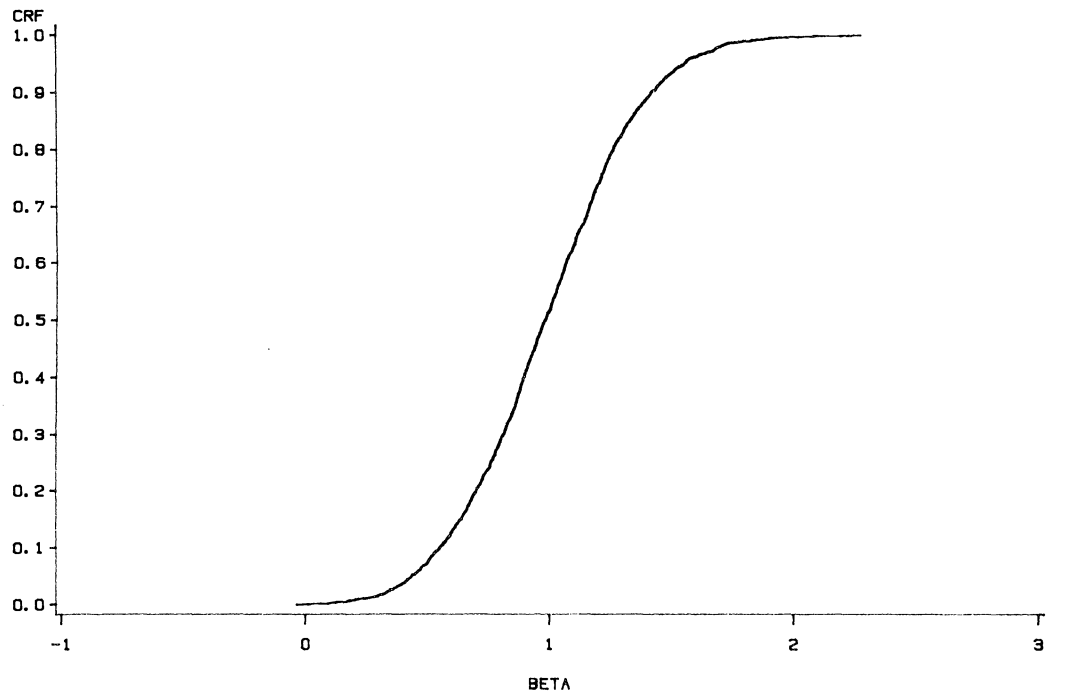

Figure 5. The marginal posterior distribution of β.

individual, (ii) the genotypes of related individuals, and (iii) the gene frequency, q, and relative risks, ϕ_1 and ϕ_2. It would also be possible to incorporate genetic marker data using linkage models. However, a difficulty arises in that genetic models may give rise to absorbing states or groups of states so that some extension of the basic algorithm will be necessary. This is an important area for further wwork.

The use of Gibbs sampling for some problems in descriptive epidemiology has been described by Clayton (technical report, University of Leicester, 1988). This work includes the analysis of cancer incidence maps using spatial priors for area-specific relative risks—an application of the Bayesian image analysis methods that motivated the work of Geman and Geman on Gibbs sampling. The cancer mapping problem is closely related to that reported here, as can be seen if geographical areas are regarded as rather large litters!

A closely related problem is that of allowing for exposure measurement error in proportional hazards models. This problem can be treated in a very similar manner, with the unknown true exposures playing a role similar to the frailties in the present model.

An attractive feature of Monte Carlo Bayes methods is the ease with which the suggestions of Rubin (1984) for *model monitoring* may be implemented, as follows. First, decide upon some diagnostic statistic, T say, and compute its value for the observed data—T_{obs}. Then, for each sample from the posterior distribution of parameters and hyperparameters, generate a simulated data set in accord with the model assumptions and the current parameter values and recalculate the statistic, T_i^*, say. The adequacy of the model may be assessed by evaluating the ranking of T_{obs} amongst the $\{T_i^*\}$. For example, the adequacy of the gamma frailty model (or of a particular genetic model for inheritance of frailty) might be assessed by examining the observed pattern of clustering of disease cases in families with that demonstrated by repeatedly simulated data sets.

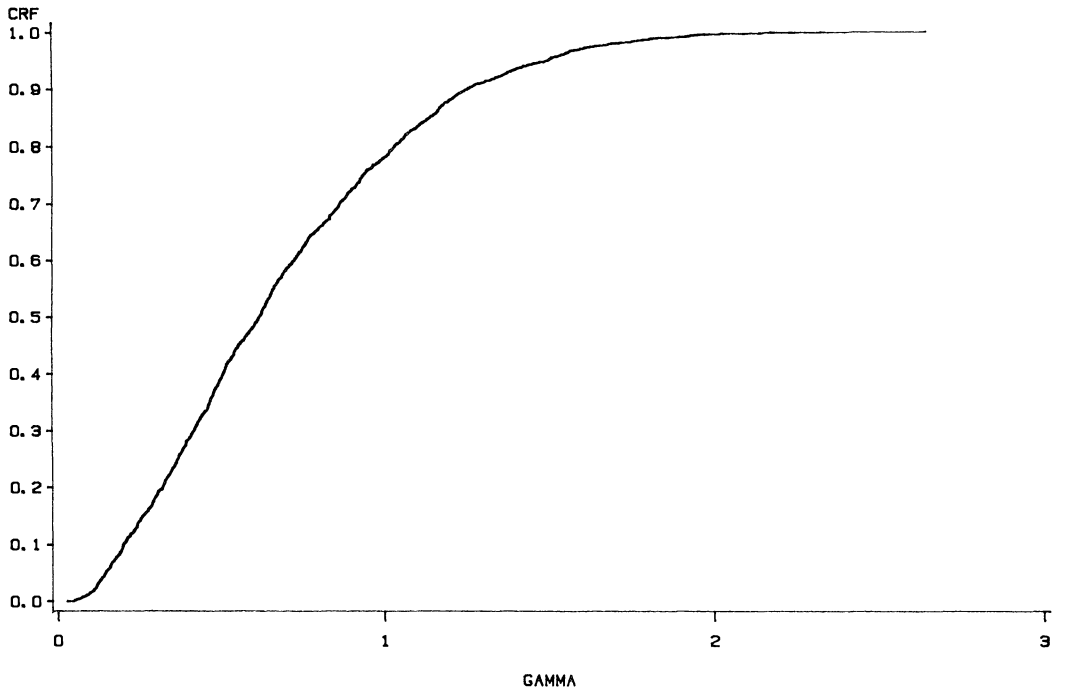

Figure 6. The marginal posterior distribution of γ.

Table 2
Summary of the distribution of 2,000 *samples of* (β, γ) *from the joint posterior distribution*

	β	γ
Mean	.986	.685
Variance	.114	.170
Quantiles:		
.05	.436	.147
.25	.759	.366
.50	.981	.618
.75	1.211	.935
.95	1.548	1.494

Model monitoring by repeated generation of data parallels the bootstrap very closely and, for censored survival data analyses, similar problems are encountered. In particular, it seems necessary to model the censoring process in order to generate the "bootstrap" data sets. For this reason, this idea has not been pursued here and it awaits further work.

While the proposed method is computationally intensive, this fact has decreasing relevance. As originally described by Geman and Geman, the Gibbs sampling algorithm is essentially a parallel one. For problems involving parallel processing, very powerful computing equipment is now available at very modest prices. For example, Transputer arrays are very suitable for implementing methods such as these. On the other hand, the time necessary to implement such a method is quite short, since a program to simulate a rather complex model may be built using only simple building bricks.

Table 3
Runs above and below the median in 2,000 *consecutive samples of* β

Run length	Above median	Below median	Above + Below Obs	Above + Below (Exp)[a]
1	246	240	486	504.5
2	135	147	282	252.2
3	62	54	116	126.1
4	30	38	68	63.1
5	19	13	32	31.5
6	7	8	15	15.8
7	3	2	5	7.9
8	0	1	1	3.9
9	1	2	3	2.0
10	1	0	1	2.0
Total	504	505	1,009	(1,001)

[a]The expected frequencies of runs of specified lengths are those expected conditional upon observing 1,009 runs in total. The expected total number of runs is that expected in a sequence of 2,000 observations.

Acknowledgement

I am grateful to Professor Duncan Thomas, Department of Preventive Medicine, University of Southern California, for helpful comments and discussions, in particular for drawing my attention to the potential application of the method in genetic epidemiology.

Résumé

Des méthodes permettant la prise en compte de covariables non observées ou "fragilités" sont nécessaires pour analyser des études épidémiologiques, pronostiques, ou étudiant la survenue d'événements au cours du temps. Clayton et Cuzick (1985, *Journal of the Royal Statistical Society, Series A* **148**, 82–117) ont proposé une généralisation du modèle à risques proportionnels qui prend en compte de tels effets aléatoires, mais la preuve de propriétés asymptotiques de la méthode semble difficile. La pratique de la méthode suggère que les vraisemblances peuvent être nettement non-quadratique. Le présent article établit une représentation bayesienne du modèle selon Kalbfleisch (1978, *Journal of the Royal Statistical Society, Series B* **40**, 214–221) et discute les résultats avec des méthodes de simulation.

References

Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics* **6**, 701–726.

Aalen, O. O. (1988). Heterogeneity in survival analysis. *Statistics in Medicine* **7**, 1121–1137.

Aalen, O. O., Borgan, O., Keiding, N., and Thormann, J. (1980). Interaction between life history events. Nonparametric analysis for prospective and retrospective data in the presence of censoring. *Scandinavian Journal of Statistics* **7**, 161–171.

Altshuler, B. (1970). Theory for the measurement of competing risks in animal experiments. *Mathematical Biosciences* **6**, 1–11.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large-sample study. *Annals of Statistics* **10**, 1100–1120.

Breslow, N. E. (1974). Covariance analysis of censored survival data. *Biometrics* **30**, 89–99.

Clayton, D. G. (1978). A model for association in bivariate life-tables and its application in epidemiological studies of chronic disease incidence. *Biometrika* **65**, 141–151.

Clayton, D. G. (1988). The analysis of event history data: A review of progress and outstanding problems. *Statistics in Medicine* **7**, 819–841.

Clayton, D. G. and Cuzick, J. (1985a). Multivariate generalizations of the proportional hazards model (with Discussion). *Journal of the Royal Statistical Society, Series A* **148**, 82–117.

Clayton, D. G. and Cuzick, J. (1985b). The EM algorithm for Cox's regression model using GLIM. *Applied Statistics* **34**, 148–156.

Clayton, D. G. and Cuzick, J. (1986). The semiparametric Pareto model for regression analysis of survival times. In *Papers on Semiparametric Models at the ISI Centenary Session, Amsterdam*, R. D. Gill and M. N. Voors (eds). Amsterdam: Centre for Mathematics and Computer Science.

Cornfield, J. and Detre, K. (1977). Bayesian life table analysis. *Journal of the Royal Statistical Society, Series B* **39**, 264–296.

Cox, D. R. (1972). Regression models and life tables (with Discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. London: Chapman and Hall.

Gammerman, D. and West, M. (1987). An application of dynamic survival models in unemployment studies. *The Statistician* **36**, 269–274.

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., Hills, S. E., Racine-Poon, A., and Smith A. F. M. (1990). Illustration of Bayesian inference in normal data models. *Journal of the American Statistical Association* **85**, 972–985.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-6**, 721–741.

Gill, R. D. (1984). Understanding Cox's regression model: A martingale approach. *Journal of the American Statistical Association* **79**, 441–447.

Gill, R. D. (1985). Contribution to the discussion of the paper by Clayton and Cuzick. *Journal of the Royal Statistical Society, Series A* **148**, **000–000**.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Hougaard, P. (1986a). Survival models for heterogeneous populations derived from stable distributions. *Biometrika* **73**, 387–396.

Hougaard, P. (1986b). A class of multivariate failure time distributions. *Biometrika* **73**, 671–678.

Johansen, S. (1983). An extension of Cox's regression model. *International Statistical Review* **51**, 165–174.

Kalbfleisch, J. D. (1978). Nonparametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society, Series B* **40**, 214–221.

Kalbfleisch, J. D. and MacKay, R. J. (1978). Remarks on a paper by Cornfield and Detre. *Journal of the Royal Statistical Society, Series B* **40**, 175–177.

Kalbfleisch, J. D. and Prentice, R. L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika* **60**, 267–278.

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

Lancaster, T. and Nickell, S. (1980). The analysis of reemployment probabilities for the unemployed (with Discussion). *Journal of the Royal Statistical Society, Series A* **143**, 141–165.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B* **50**, 157–224.

Mack, W., Langholz, B., and Thomas, D. C. (1989). Survival models for familial aggregation of cancer. *Environmental Health Perspectives* **00, 000–000**.

Mantel, N. and Ciminera, J. L. (1979). Mantel–Haenszel analysis of litter-matched time-to-response data with modifications for recovery of interlitter information. *Cancer Research* **37**, 3863–3868.

Nelson, W. (1969). Hazard plotting for incomplete failure data. *Journal of Quality Technology* **1**, 27–52.

Ripley, B. D. (1977). Algorithm 137. Simulating spatial patterns: Dependent samples from a multivariate density. *Applied Statistics* **28**, 109–112.

Ripley, B. D. (1987). *Stochastic Simulation*, New York: Wiley.

Rubin, D. B. (1984). Bayesian justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics* **12**, 1151–1172.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with Discussion). *Journal of the American Statistical Association* **82**, 528–550.

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16**, 439–454.

Wermuth, N. and Lauritzen, S. L. (1983). Graphical and recursive models for contingency tables. *Biometrika* **70**, 537–552.

Received May 1989; *revised January* 1990; *accepted March* 1990.