Research Article

Statistics in Medicine

Received 22 September 2009, Accepted 4 August 2010 Published online 20 October 2010 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4076

A Bayesian proportional hazards regression model with non-ignorably missing time-varying covariates

Patrick T. Bradshaw,[a]*† Joseph G. Ibrahim[b] and Marilie D. Gammon[a]

Missing covariate data are common in observational studies of time to an event, especially when covariates are repeatedly measured over time. Failure to account for the missing data can lead to bias or loss of efficiency, especially when the data are non-ignorably missing. Previous work has focused on the case of fixed covariates rather than those that are repeatedly measured over the follow-up period, hence, here we present a selection model that allows for proportional hazards regression with time-varying covariates when some covariates may be non-ignorably missing. We develop a fully Bayesian model and obtain posterior estimates of the parameters via the Gibbs sampler in WinBUGS. We illustrate our model with an analysis of post-diagnosis weight change and survival after breast cancer diagnosis in the Long Island Breast Cancer Study Project follow-up study. Our results indicate that post-diagnosis weight gain is associated with lower all-cause and breast cancer-specific survival among women diagnosed with new primary breast cancer. Our sensitivity analysis showed only slight differences between models with different assumptions on the missing data mechanism yet the complete-case analysis yielded markedly different results.

Keywords: proportional hazards regression; non-ignorably missing data; missing covariates; selection model

1. Introduction

Studies of survivorship are often plagued by missing covariate data, especially when assessments are made longitudinally and deal with lifestyle or behavioral characteristics that may be sensitive in nature. A formal treatment of missing data requires consideration of the process that leads to the incomplete observations, such as the taxonomy suggested by Little and Rubin [1]. Data are considered missing completely at random if the probability that data are missing is independent of both observed and unobserved data. Under this scenario the observed data essentially constitute a random sample of values from all subjects and thus a complete-case analysis, which uses data only on those subjects with no missing observations, will yield unbiased parameter estimates. If the probability that data are missing depends only upon fully observed variables then the data are referred to as missing at random (MAR). In the Bayesian framework, if the parameters that index the probability that data are missing are independent of the parameters that index the distribution of the missing variables, then in the MAR case the missing data mechanism is called ignorable [1]. The most problematic situation arises when the probability that data are missing depends upon unobserved values of the missing variable, which is what we suspect for our data. When the probability that a variable is missing depends upon its unobserved value then the data are referred to as not missing at random (NMAR) and the missing data mechanism is referred to as non-ignorable. Valid estimation under non-ignorable missingness requires simultaneously accounting for the probability that data are missing, the distribution of the values of the missing variable and the relationship between the potentially incomplete variable to the outcome of interest.

The majority of the literature on missing covariates in proportional hazards regression has focused on frequentist methods for baseline MAR covariates [2–10]. The Bayesian approach to survival analysis with covariate data that is

[a]*Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, U.S.A.*
[b]*Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, U.S.A.*
Correspondence to: Patrick T. Bradshaw, Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, U.S.A.
†*E-mail: patrickb@email.unc.edu*

MAR is described in detail by Ibrahim *et al*. [11]. Frequentist methods for non-ignorably missing covariates in survival analysis have been presented by Leong *et al*. [12] and Herring *et al*. [13]; however, both of these methods apply only to baseline (fixed) covariates, and the method by Leong and colleagues requires them to also be dichotomous. The selection model outlined by Herring *et al.* [13] specifies the joint distribution of the survival times, missing covariates and missingness indicator through a series of one-dimensional conditional distributions and uses a Monte Carlo Expectation Maximization (MCEM) algorithm for parameter estimation. Here, we propose a model that extends this approach by allowing the covariate values to vary over time, and we present a computationally easier alternative to the MCEM algorithm for parameter estimation.

The motivation for developing this model lies in our interest in identifying modifiable lifestyle factors that may be associated with survival among women with breast cancer. Factors related to energy balance, such as post-diagnosis weight gain, are of particular interest, yet the effect remains unclear. Excess adipose tissue is associated with a hormonal environment conducive to tumor promotion [14, 15] and therefore may be associated with poorer survival. This is especially concerning since weight gain after breast cancer diagnosis is common [16], and has been associated with chemotherapy treatment, menopausal status, age at diagnosis and pre-diagnosis body size. Our objective is to evaluate how post-diagnosis changes in weight over time affect survival through an analysis of data from the follow-up to the Long Island Breast Cancer Study Project (LIBCSP) [17]. An issue for this analysis is that a significant portion of the study subjects is missing data for one or more follow-up assessments of bodyweight making bias or loss of efficiency of serious concern if we limit our investigation to only those subjects with complete data. Specifically, given the stigma associated with being overweight, we suspect that those subjects with missing data on body size may tend to be heavier than those who responded, making this subset of subjects a less than representative sample of the study population. In this case, standard proportional hazards models would be inappropriate. With repeated measurements of body size at and after diagnosis, our primary covariate of interest is time varying, requiring us to develop the model we present here. To our knowledge there has been no previous work addressing inference for selection models with non-ignorably missing time-varying covariates.

The following section of the paper will outline our notation and describe the selection model in general. We then describe specific models for each of the conditional distributions: the missingness indicator given time-to-event and covariates, the time-to-event model given covariates and the distribution of the missing covariates. We then describe our estimation approach and then illustrate our model with an example of an analysis of changes in bodyweight over time and survival after breast cancer diagnosis, using data from the follow-up to the LIBCSP [17]. We conclude with a discussion of the results and the methodology.

2. The selection model

Here we outline a selection model for proportional hazards regression with time-varying covariates, which is defined by the joint distribution of the event times, missing covariates and the mechanism that describes the probability of missingness. This joint distribution is specified through a series of conditional distributions: (1) the probability that the covariate data is missing conditional on event time and (possibly unobserved) covariates, (2) the distribution of event time conditional on covariates and (3) the marginal distribution of the missing covariates dependent only on fully-observed variables. We begin by outlining the notation for the model.

Assume we have data on a sample of n-independent subjects and for subject i denote the event time by T_i and censoring time by C_i. For each of the n subjects we observe the variable $y_i = \min(T_i, C_i)$ and indicator of failure δ_i which takes on the value 1 if y_i corresponds to an occurrence of an event (i.e. $T_i \leqslant C_i$), and 0 if it represents a censored observation (i.e. $T_i > C_i$). We further assume independence between T_i and C_i. Each subject provides a series of longitudinal measurements for $(p+q)$ variables where for the kth measurement we denote the vector of p completely observed variables by $\mathbf{x}_{ik} = (x_{ik1}, \ldots, x_{ikp})$ and q variables with potentially missing values by $\mathbf{z}_{ik} = (z_{ik1}, \ldots, z_{ikq})$ measured at times v_{ik} for $k = 1, \ldots, K_i$, where $K_i \geqslant 1$. Elements of the $\mathbf{z}_{ik}$ vector are missing for only some subjects at some of the measurement points; hence, associated with each variable in $\mathbf{z}_{ik}$ is an indicator of missingness for that variable contained in the vector $\mathbf{r}_{ik} = (r_{ik1}, \ldots, r_{ikq})$, where $r_{ikl} = 1$ if z_{ikl} is missing and $r_{ikl} = 0$ otherwise, for $l = 1, \ldots, q$ and $k = 1, \ldots, K_i$. The notation $\mathbf{x}_i$, $\mathbf{z}_i$, and $\mathbf{r}_i$ will refer to matrices of size $K_i \times p$, $K_i \times q$, and $K_i \times q$ respectively, representing the set of all K_i measurements for each vector of variables for each subject i.

In general, estimation of a proportional hazards regression of y on $[\mathbf{x}\ \mathbf{z}]$ using complete covariate data (only where $r_{ikl} = 0$) will yield biased estimates of the regression parameters as it does not account for the distribution of the missing variables and more importantly the possibility that the reason they are missing may be related to their unobserved values. The selection model allows us to specify the joint distribution of $(\mathbf{r}_i, y_i, \mathbf{z}_i | \mathbf{x}_i)$ allowing us to account for these relationships with the goal of obtaining unbiased estimates of the regression parameters. In general, the complete data

joint distribution of $(\mathbf{r}_i, y_i, \mathbf{z}_i|\mathbf{x}_i)$ maybe expressed as a series of conditionals:

$$p(\mathbf{r}_i, y_i, \mathbf{z}_i|\mathbf{x}_i, \lambda, \beta, \alpha, \phi) = p_{\mathbf{r}}(\mathbf{r}_i|y_i, \mathbf{z}_i, \mathbf{x}_i, \phi) p_y(y_i|\mathbf{z}_i, \mathbf{x}_i, \lambda, \beta) p_{\mathbf{z}}(\mathbf{z}_i|\mathbf{x}_i, \alpha). \quad (1)$$

The parameters ϕ and α index the distribution of the missing data mechanism and the missing covariates and are nuisance parameters which are not of inferential interest in the proportional hazards model, the primary objective of undertaking the analysis. The remainder of this section describes the specification of these conditional densities and the form of the complete data likelihood in detail.

2.1. Models for the missing data mechanism

The assumption of nonignorability of the missing data process requires specification of the distribution of the probability of missingness, which is assumed to be dependent upon the unobserved value the corresponding variable would have taken if it were observed. We follow Ibrahim *et al*. [18] and Stubbendick and Ibrahim [19] by modeling the missing data mechanism $\mathbf{r}_i$ as a series of one-dimensional conditional distributions, which is effective at reducing the number of nuisance parameters while maintaining correlation between the longitudinal observations and allowing for non-monotone patterns of missingness [19, 20]. For the joint distribution of $\mathbf{r}_i$, we specify a distribution for each r_{ikl} sequentially conditioning over the other missingness indicators at measurement k, previous missingness indicators for all variables at all measurements prior to k, the corresponding vector of completely observed and possibly missing covariates, $\mathbf{x}_{ik}$ and $\mathbf{z}_{ik}$, respectively, event time y_i and vector of parameters ϕ_{kl}, with the set of all of these denoted by ϕ:

$$\begin{aligned} p_{\mathbf{r}}(\mathbf{r}_i|y_i, \mathbf{x}_i, \mathbf{z}_i, \phi) = {} & p(r_{iK_iq}|r_{iK_i1}, \ldots, r_{iK_i(q-1)}, \mathbf{r}_{i(K_i-1)}, \ldots, \mathbf{r}_{i1}, \mathbf{x}_{iK_i}, \mathbf{z}_{iK_i}, y_i, \phi_{K_iq}) \\ & \times \cdots \times p(r_{iK_i1}|\mathbf{r}_{i(K_i-1)}, \ldots, \mathbf{r}_{i1}, \mathbf{x}_{iK_i}, \mathbf{z}_{iK_i}, y_i, \phi_{K_i1}) \\ & \times \cdots \times p(r_{i(K_i-1)q}|r_{i(K_i-1)1}, \ldots, r_{i(K_i-1)(q-1)}, \mathbf{r}_{i(K_i-2)}, \ldots, \mathbf{r}_{i1}, \mathbf{x}_{i(K_i-1)}, \mathbf{z}_{i(K_i-1)}, y_i, \phi_{(K_i-1)q}) \\ & \times \cdots \times p(r_{i(K_i-1)1}|\mathbf{r}_{i(K_i-2)}, \ldots, \mathbf{r}_{i1}, \mathbf{x}_{i(K_i-2)}, \mathbf{z}_{i(K_i-2)}, y_i, \phi_{(K_i-1)1}) \\ & \times \cdots \times p(r_{i1q}|r_{i11}, \ldots, r_{i1(q-1)}, \mathbf{x}_{i1}, \mathbf{z}_{i1}, y_i, \phi_{1q}) \times \cdots \times p(r_{i11}|\mathbf{x}_{i1}, \mathbf{z}_{i1}, y_i, \phi_{11}). \end{aligned} \quad (2)$$

Sequentially conditioning on previous measurements approximates a correlation structure similar to what would be obtained using random effects models without the need to specify the random effect [19, 21]. A series of logistic regressions may be used to model these conditional distributions as each r_{ikl} is dichotomous. The contribution to the complete-data likelihood for subject i corresponding to the missing data mechanism is thus given by equation (2).

Although the specification above appears quite complicated, in practice the number of measurements K_i is likely to be small and it may be realistic to assume that only some subset of the variables in $\mathbf{z}$ are non-ignorably missing, and therefore the number of variables requiring specification of missingness models is fewer than q. Although it may be tempting to include a large number of variables and cross-products into the missing data models the analyst should strive for the most parsimonious specification possible as these models can easily become unidentifiable [13, 20, 22]. Herring *et al.* [13] and Ibrahim *et al*. [23] suggest a strategy for model selection for the missing data mechanism of these models to help avoid issues of identifiability.

2.2. Model for the time-to-event

We consider here a Cox piecewise exponential hazard model to describe the relationship between the event time and the covariates. To define the piecewise exponential model, we divide the time axis into J discrete intervals $(s_{j-1}, s_j]$ for $j=1, \ldots, J$ with $s_0=0$ and s_J greater than the maximum of the $\{y_i\}$. The measurement times for the covariate vector are assumed to fall at the boundaries of the intervals although it is possible for a measurement to span multiple intervals (e.g. if measurements on $[\mathbf{x}\,\mathbf{z}]$ are taken every 2 years but the intervals $(s_{j-1}, s_j]$ correspond to 1 year each). Thus, since the number of covariate measurements is $K_i \leqslant J$ then we define a notation so the indexes on each of the covariates match the index for the intervals of the piecewise exponential model. Then for subject i within interval j, we define $\mathbf{x}^*_{ij} = (x^*_{ij1}, x^*_{ij2}, \ldots, x^*_{ijp})'$ and $\mathbf{z}^*_{ij} = (z^*_{ij1}, z^*_{ij2}, \ldots, z^*_{ijq})'$, where $\mathbf{x}^*_{ijl}$ and $\mathbf{z}^*_{ijl}$ may represent the previous observation of the variable carried forward into the current interval, or, for continuous variables, an interpolated value between two observations. Thus $\mathbf{x}^*_{ij}$ denotes the $p \times 1$ vector of fully observed covariate values and $\mathbf{z}^*_{ij}$ denotes the $q \times 1$ vector of possibly missing covariate values corresponding to the jth interval for $j=1, \ldots, J$. We then define the piecewise exponential hazards model with the hazard function:

$$\lambda(y_i|\mathbf{x}^*_{ij}, \mathbf{z}^*_{ij}, \beta, \lambda_j) = \lambda_j \exp(\mathbf{x}^{*\prime}_{ij}\beta_1 + \mathbf{z}^{*\prime}_{ij}\beta_2) \quad \text{for } y \in (s_{j-1}, s_j]$$

with $\beta=[\beta_1,\beta_2]'$, where β_1 is the $p\times 1$ vector of coefficients on the vector of covariates $\mathbf{x}^*_{ij}$ and β_2 is the $q\times 1$ vector of coefficients on the vector of covariates $\mathbf{z}^*_{ij}$. The density for the observed failure time y_i within interval j is then:

$$p_{y,j}(y_i|\mathbf{x}^*_{ij},\mathbf{z}^*_{ij},\beta_1,\beta_2,\lambda_j)=(\lambda_j\exp(\mathbf{x}^{*\prime}_{ij}\beta_1+\mathbf{z}^{*\prime}_{ij}\beta_2))^{\delta_i}(\exp(-\Lambda_i(y_i)))^{\exp(\mathbf{x}^{*\prime}_{ij}\beta_1+\mathbf{z}^{*\prime}_{ij}\beta_2)}$$

for $y_i\in(s_{j-1},s_j]$ with cumulative hazard function:

$$\Lambda_j(y_i)=\left((y_i-s_{j-1})\lambda_j\exp(\mathbf{x}^{*\prime}_{ij}\beta_1+\mathbf{z}^{*\prime}_{ij}\beta_2)+\sum_{g=1}^{j-1}(s_g-s_{g-1})\lambda_g\exp(\mathbf{x}^{*\prime}_{ig}\beta_1+\mathbf{z}^{*\prime}_{ig}\beta_2)\right).$$

We further let $\lambda=(\lambda_1,\ldots,\lambda_J)'$ denote the $J\times 1$ vector of baseline hazards λ_j and let Δ_{ij} be an indicator of if subject i died or was censored in interval j (i.e. $y_i\in(s_{j-1},s_j]$). The ith contribution to the complete data likelihood for the piecewise exponential model is then:

$$p_y(y_i|\mathbf{x}_i,\mathbf{z}_i,\beta_1,\beta_2,\lambda)=\prod_{j=1}^{J}(\lambda_j\exp(\mathbf{x}^{*\prime}_{ij}\beta_1+\mathbf{z}^{*\prime}_{ij}\beta_2))^{\Delta_{ij}\delta_i}\exp\{-\Delta_{ij}[\Lambda_j(y_i)]\exp(\mathbf{x}^{*\prime}_{ij}\beta_1+\mathbf{z}^{*\prime}_{ij}\beta_2)\} \tag{3}$$

where $\mathbf{x}^*_{ij}=\mathbf{x}_{ik}$ and $\mathbf{z}^*_{ij}=\mathbf{z}_{ik}$ with k and j such that $v_{ik}\leqslant s_{j-1}<v_{i,k+1}$. If we define $t_{ij}=\min(y_i,s_{j+1})-s_j$ if $y_i\geqslant s_j$ and 0 if $y_i<s_j$ to be the length of the overlap from the beginning of interval j to the end of the interval or until failure time y_i, then it can be shown that the likelihood function given by equation (3) is equivalent to one where $\Delta_{ij}\delta_i$ follows a Poisson distribution with mean $t_{ij}\lambda(y_i|\mathbf{x}^*_{ij},\mathbf{z}^*_{ij},\beta_1,\beta_2,\lambda_j)$.

2.3. Models for the missing covariates

For the joint distribution of the missing covariates $\mathbf{z}_i$ we again follow the strategy suggested by Lipsitz and Ibrahim [24], Ibrahim *et al.* [18], and Stubbendick and Ibrahim [19] by specifying a sequence of one-dimensional conditional distributions. We specify a model for each z_{ikl} sequentially conditioning over the other $\mathbf{z}$-variables at measurement k, all $\mathbf{z}$ variables at previous times, the corresponding vector of completely observed covariates, $\mathbf{x}_{ik}$, event time y_i, and $\alpha=(\alpha_{11},\ldots,\alpha_{K_iq})'$ where each α_{kl} is a vector of parameters indexing the distribution for each covariate l for measurement k. The joint distribution of the $\mathbf{z}$ variables for subject i is then:

$$\begin{aligned}p_{\mathbf{z}}(\mathbf{z}_i|\mathbf{x}_i,\alpha)=\;& p(z_{iK_iq}|z_{iK_i1},\ldots,z_{iK_i(q-1)},\mathbf{z}_{i(K_i-1)},\ldots,\mathbf{z}_{i1},\mathbf{x}_{iK_i},\alpha_{K_iq})\\ &\times\cdots\times p(z_{iK_i1}|\mathbf{z}_{i(K_i-1)},\ldots,\mathbf{z}_{i1},\mathbf{x}_{iK_i},\alpha_{K_i1})\\ &\times\cdots\times p(z_{i(K_i-1)q}|z_{i(K_i-1)1},\ldots,z_{i(K_i-1)(q-1)},z_{i(K_i-2)},\ldots,z_{i1},\mathbf{x}_{i(K_i-1)},\alpha_{(K_i-1)q})\\ &\times\cdots\times p(z_{i(K_i-1)1}|\mathbf{z}_{i(K_i-2)},\ldots,\mathbf{z}_{i1},\mathbf{x}_{i(K_i-2)},\alpha_{(K_i-1)1})\\ &\times\cdots\times p(z_{i1q}|z_{i11},\ldots,z_{i1(q-1)},\mathbf{x}_{i1},\alpha_{1q})\times\cdots\times p(r_{i11}|\mathbf{x}_{i1},\mathbf{z}_{i1},\alpha_{11}).\end{aligned} \tag{4}$$

Expression of the distribution of covariates this way allows considerable flexibility in the choice of distribution for each z_{ikl}, accommodating continuous and categorical variables, as well as offering a convenient way to account for intra-subject correlation without specification of a random effect. Once again, one should strive for a parsimonious specification of this joint distribution to avoid specification issues. Equation (4) then represents the ith contribution to the marginal likelihood for $\mathbf{z}$.

2.4. Estimation

Substitution of equations (2), (3) and (4) into (1) yields the complete data likelihood:

$$\begin{aligned}\ell(\beta_1,\beta_2,\lambda,\alpha,\phi)&=\prod_{i=1}^{n}p(\mathbf{r}_i,y_i,\mathbf{z}_i|\mathbf{x}_i,\beta_1,\beta_2,\lambda,\phi,\alpha)\\ &=\prod_{i=1}^{n}p_{\mathbf{r}}(\mathbf{r}_i|y_i,\mathbf{z}_i,\mathbf{x}_i,\phi)p_y(y_i|\mathbf{z}_i,\mathbf{x}_i,\beta_1,\beta_2,\lambda)p_{\mathbf{z}}(\mathbf{z}_i|\mathbf{x}_i,\alpha)\end{aligned}$$

with densities $p_{\mathbf{r}}(\cdot)$, $p_y(\cdot)$ and $p_{\mathbf{z}}(\cdot)$ defined above. Previous work with similar models has made use of the EM algorithm to obtain the parameter estimates [13]. However, here we illustrate a Fully Bayesian (FB) approach using vague priors on the parameters $\beta_1,\beta_2,\lambda,\alpha$, and ϕ which will produce estimates equivalent to the frequentist analysis using EM and

also yield variance estimates that are much easier to obtain than with the EM framework. The FB framework is also less computationally demanding than the EM framework for this model.

The joint posterior distribution of the parameters is proportional to the product of the conditional distribution of the observed data given the parameters and the joint prior distribution of the model parameters $p(\beta_1, \beta_2, \lambda, \alpha, \phi)$:

$$p(\beta_1, \beta_2, \lambda, \alpha, \phi | y, \mathbf{r}, \mathbf{x}, \mathbf{z}) \propto \left(\prod_{i=1}^{n} \int_{\mathbf{z}_i} p(y_i, \mathbf{r}_i, \mathbf{x}_i, \mathbf{z}_i | \beta_1, \beta_2, \lambda, \alpha, \phi) \, \mathrm{d}\mathbf{z}_i \right) \times p(\beta_1, \beta_2, \lambda, \alpha, \phi). \tag{5}$$

If non-informative priors are specified for $(\beta_1, \beta_2, \lambda, \alpha, \phi)$ then the posterior means and standard deviations of the parameters will be similar to maximum likelihood. We use the Gibbs Sampler [25] to sample from the posterior distribution given by (5). Although somewhat computationally intensive (but less intensive than EM), the FB approach here provides a very straightforward way to estimate parameters from a complex model, especially variance and covariance parameters.

3. Example

We apply this model to an analysis of data from the follow-up study to the LIBCSP, to evaluate whether time-varying post-diagnosis changes in body size are related to survival among women with newly diagnosed breast cancer.

3.1. Description of the LIBCSP

The details of the LIBCSP are discussed elsewhere [17] but briefly, the parent study is a population-based case-control study of breast cancer among women in Nassau and Suffolk counties on Long Island, New York conducted between August 1996 and July 1997. Cases consisted of 1508 women with newly diagnosed *in situ* or invasive breast cancer; of these, 1414 women agreed to be contacted at a later date for follow-up interviews. For those who agreed to participate in the follow-up, the case subjects or their proxy were contacted by mail approximately 5 years after initial diagnosis of breast cancer and informed consent was obtained via telephone follow-up calls. Of the 1414 women who initially agreed to participate, 316 subsequently refused or were unable to be contacted. Of the remaining 1098 subjects who agreed to the follow-up interview, only 1033 case subjects or proxies (68.5 per cent of the original 1,508 women) actually completed the interviewer-administered questionnaire [26]. The follow-up interview ascertained information similar to that gathered in the baseline questionnaire but relevant to the time period since diagnosis including treatment, reproductive history, smoking and alcohol use, and as well as body size and physical activity. Date and cause of death were ascertained for all 1508 women using the National Death Index [27] with median follow-up time of 8.8 years (range: 0.2–9.4 years).

Relevant to this analysis, the follow-up questionnaire ascertained body size (weight in pounds and height in inches) at diagnosis, 1-year post-diagnosis and at time of response to questionnaire for those subjects still living, or 1-year prior to death for interviews completed by proxy for subjects who were deceased at the follow-up but living longer than 1-year (specific timing for final follow-up measurement varied between 2 and 7 years post diagnosis with an average of 4.95 years (Table I); variation in timing of follow-up was due to logistical issues unrelated to health status or weight change). Overall refusal to participate in the follow-up interview and non-response to specific questions among people still alive at each timepoint resulted in percentages of missing data on body size of 47.6, 49.4 and 33.9 per cent at baseline, 1-year post diagnosis and final follow-up. Our concern is that heavier women may not have responded to the questionnaire in general, or to the body size questions specifically, due to self-conscious feelings or other reasons related to the amount of their weight, creating a non-ignorable mechanism for missing body size data. With the body size variables we calculate percent change in body weight between the year prior to diagnosis and kth measurement (100*(weight at measurement $k-$ weight one year before diagnosis)/weight one year before diagnosis) for $k=1$ (at baseline), 2 (at one year) and 3 (at time of interview or 1-year prior to death).

Other fixed covariates we include in our analysis (measured only once, at diagnosis) are indicators of chemotherapy regimen (yes/no), tumor size greater than 2 cm (yes/no), estrogen receptor positive tumor (ER status, yes/no) and progesterone receptor positive tumor (PR status, yes/no). Each of these covariates also exhibited a significant amount of missing data with 32.2, 31.6, 34.0 and 34.3 per cent missing, respectively. The overlap between these variables with missing values was small, and since a complete-case analysis requires all variables to be observed, the resulting completely observed data set which excluded those with missing post-diagnosis change in bodyweight or a missing value for any covariate, contained 499 subjects. Note, however, that the percentage missing for any one of these variables was moderate. Other important covariates included menopausal status, education, adult weight change and body mass index (BMI) 1-year prior to diagnosis, which were each missing for less than 2 per cent of subjects. We also include data on age at diagnosis, which is fully observed for all women. For the small amount of missing data on menopausal status, education, adult weight change and prediagnosis BMI we will exclude these subjects from the analysis, however

Table I. Descriptive statistics for the 1436 subjects from the Long Island Breast Cancer Study Project included in this analysis.

Continuous variables	Mean	Variance	Categorical variables	*n*	per cent*
Time of follow-up interview	4.95	1.65	ER-positive tumor		
Age at diagnosis	58.79	159.73	No	252	26.44
BMI 1 year before diagnosis	26.57	32.03	Yes	404	41.35
			Missing	483	
Categorical variables	*n*	Per cent*	PR-positive tumor		
Weight change at baseline			No	340	58.65
>5% loss	197	25.99	Yes	609	41.35
maintain within 5%	458	60.42	*Missing*	487	
5–10% gain	72	9.50	Tumor size > 2cm		
>10% gain	31	4.09	No	749	76.20
Missing		678	Yes	234	23.80
Weight change 1-year after diagnosis			*Missing*	453	
>5% loss	168	23.11	Menopausal Status		
maintain within 5%	400	55.02	Premenopausal	462	32.17
5–10% gain	95	13.07	Postmenopausal	974	67.83
>10% gain	64	8.80	Income (per year)		
Missing	709		<$20 000	174	12.12
Weight change at follow-up interview			$20 000–49 999	560	39.00
>5% loss	200	21.21	$50 000-89 999	427	29.74
maintain within 5%	414	43.90	>$90 000	275	19.15
5–10% gain	161	17.07	Education		
>10% gain	168	17.82	No college	688	47.91
Missing	493		Some college	342	23.82
Chemotherapy treatment			College graduate	186	12.95
No	573	58.65	Post-college	220	15.32
Yes	404	41.35			
Missing	459				

*Percent among observed data.

for the remaining 1436 subjects we will specify a selection model to account for the significant amount of missing data on follow-up body size, treatment and tumor characteristics. Out of the 1436 women included in this analysis, 292 died during the follow-up period with 156 of those deaths attributed to breast cancer.

3.2. *Selection model*

We model the time since diagnosis for subject i (denoted dur_i) as a piecewise exponential model with $J=10$ 1-year intervals, along with the indicator of death (1) or censoring (0). Percent change in bodyweight for subject i in interval j corresponding to measurement k, denoted as pcwt_{ij}^* was calculated by linear interpolation of the variable pcwt_{ik} for intervals between measurement points and was categorized into four groups using indicator functions $I_{(a,b)}(x)$ where $I_{(a,b)}(x)=1$ if $x\in(a,b)$ and 0 otherwise. The four categories represent those who lost more than 5 per cent of their pre-diagnosis body weight ($\text{pcwt}_{ij}^*<-5$), those who maintained within 5 per cent of their pre-diagnosis bodyweight ($\text{pcwt}_{ij}^*\geqslant -5$ and $\text{pcwt}_{ij}^*\leqslant 5$), those who gained between 5 and 10 per cent of their pre-diagnosis weight ($\text{pcwt}_{ij}^*>5$ and $\text{pcwt}_{ij}^*<10$) and those who gained 10 per cent or more of their prediagnosis bodyweight ($\text{pcwt}_{ij}^*\geqslant 10$), omitting the category corresponding to those maintaining weight as the referent group. We also include fixed covariates age at diagnosis (continuous, dxage_i), indicators for chemotherapy treatment (chemo_i), ER status (erstat_i), PR status (prstat_i) and tumor size >2cm (tumor_i), yielding the hazard function for our time to event model:

$$\lambda(\text{dur}_i|\mathbf{x}_{ij},\mathbf{z}_{ij},\beta_1,\beta_2,\lambda_j)=\lambda_j\exp(\beta_{11}\text{dxage}_i+\beta_{21}I_{-\infty,-5}(\text{pcwt}_{ij}^*)+\beta_{22}I_{5,10}(\text{pcwt}_{ij}^*)+\beta_{23}I_{10,\infty}(\text{pcwt}_{ij}^*)+\beta_{24}\text{chemo}_i$$
$$+\beta_{25}\text{erstat}_i+\beta_{26}\text{prstat}_i+\beta_{27}\text{tumor}_i). \quad (6)$$

for $\text{dur}_i\in(s_{j-1},s_j]$. For our analysis we assume that only percent change in weight (pcwt_{ik}) is potentially non-ignorably missing, while chemotherapy treatment (chemo_i) and the tumor characteristics tumor size >2cm (tumor_i), ER status (erstat_i) and PR status (prstat_i) are ignorably missing as we believe that their missingness is unlikely to be related to either unknown or known variables. Therefore, only one missing data mechanism need to be specified: $r_{ik}=1$ if subject i was missing body size responses at measurement k for $k=1,\ldots,K_i$, where K_i =1, 2 or 3 and $r_{ik}=0$ if the value was

present. Then, from equation (2) for $K_i = 3$ we have:

$$\begin{aligned} p_{\mathbf{r}}(\mathbf{r}_i | y_i, \mathbf{x}_i, \mathbf{z}_i, \phi) = {} & p(r_{i3} | r_{i2}, r_{i1}, \text{dxage}_i, \text{pcwt}_{i3}, \text{fu.years}_i, \phi_3) \\ & \times p(r_{i2} | r_{i1}, \text{dxage}_i, \text{pcwt}_{i2}, t_{i2}, \phi_2) \\ & \times p(r_{i1} | \text{dxage}_i, \text{pcwt}_{i1}, \phi_1) \end{aligned} \tag{7}$$

where dxage_i is age at diagnosis, fully observed. The variable fu.years contains the length of time from diagnosis to when the follow-up interview was completed, corresponding to the final weight change measurement. For subjects who did not complete the follow-up interview, this was assumed to be 5 years, or the year prior to death, whichever was smaller, which is the timing for the interviews actually completed by subject or proxy, respectively. The modification of (7) for K_i=1 or 2 is straightforward. We specify each of the conditional distributions on the right-hand side of equation (7) with a logistic regression model.

Using equation (4) we express the joint distribution of the missing time-varying covariates percent change in weight (pcwt_{ik}) for $k = 1, \ldots, 3$, and fixed (baseline) covariates chemotherapy treatment (chemo_i), tumor size >2 cm (tumor_i), ER status (erstat_i), PR status (prstat_i), as functions of menopausal status at diagnosis (menpstat_i), BMI in the year prior to diagnosis (bmiref_i), income reported at diagnosis (income_i), years of education completed (education_i) and age at diagnosis (dxage_i) as:

$$\begin{aligned} p_{\mathbf{z}}(\mathbf{z}_i | y_i, \mathbf{x}_i, \alpha, \tau) = {} & p(\text{pcwt}_{i3} | \text{pcwt}_{i2}, \text{pcwt}_{i1}, \text{dxage}_i, \text{chemo}_i, \text{menpstat}_i, \text{bmiref}_i, \alpha_7, \tau_3) \\ & \times p(\text{pcwt}_{i2} | \text{pcwt}_{i1}, \text{dxage}_i, \text{chemo}_i, \text{menpstat}_i, \text{bmiref}_i, \alpha_6, \tau_2) \\ & \times p(\text{pcwt}_{i1} | \text{dxage}_i, \text{chemo}_i, \text{menpstat}_i, \text{bmiref}_i, \alpha_5, \tau_1) \\ & \times p(\text{chemo}_i | \text{dxage}_i, \text{income}_i, \text{education}_i, \alpha_4) \times p(\text{erstat}_i | \text{dxage}_i, \alpha_3) \\ & \times p(prstat_i | \text{dxage}_i, \alpha_2) \times p(\text{tumor}_i | \text{dxage}_i, \text{income}_i, \text{education}_i, \alpha_1). \end{aligned} \tag{8}$$

We model the conditional distributions of percent change in weight as linear regression models while the dichotomous treatment and tumor characteristic variables are modeled using logistic regression models. Note that we also explicitly include the precision parameters $\tau = (\tau_1, \tau_2, \tau_3)$ (the reciprocal of the variances) for the continuous variables. The choice of covariates for the models for percent change in bodyweight was determined based on consensus of previous studies on postdiagnosis weight change among breast cancer patients [16, 28]. The models for treatment variables were selected in the interest of parsimony and to represent those variables we believe to be associated with access to care. An alternative option for the models for percent change in weight would be to generate a single four-level ordinal categorical variable for pcwt_{ik} with the corresponding indicator variables in the piecewise exponential model, and ordinal logistic regression models for the categorical variable in the joint distribution given by equation (8). However, given the inherently continuous nature of the underlying variable the method presented here yields an equivalent and more intuitive specification.

We selected noninformative priors for the unknown parameters in the model. For the slope parameters for the regression models ($\beta_1, \beta_2, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7, \phi_1, \phi_2, \phi_3$), we specified independent normal distributions with zero mean and precision of 10^{-6} and for the baseline hazards (λ) and precision parameters for the linear regression models (τ_1, τ_2, τ_3) we specified gamma distributions with shape and inverse scale parameters of 0.01. Estimation was performed using the Gibbs sampler in WinBUGS 1.4 [29] run for 100 000 iterations with an additional 100 000 burn-in. Convergence was established through the use of Geweke's Z-statistic [30] calculated with the CODA package in R [31] as well as through visual inspection of trace plots (see online supplementary material). To evaluate the robustness of our model to assumptions on the missing data mechanism we also estimated this model assuming that the longitudinal body size variable was MAR, therefore omitting the specification for the missing data mechanism, equation (7). For comparison, we also performed a complete-case analysis by eliminating those subjects who were missing observations for one or more covariates, as well as the models for the missing data mechanism (7) and the covariate distributions (8), which is analogous to the usual approach in most commercial software packages. Each set of models was estimated for both all-cause and breast-cancer specific mortality. In order to assess the sensitivity of the model to changes in the parameterization of the prior distributions, we additionally estimated models for all-cause mortality with precision on the prior for the β parameters of 10^{-3} and 10^{-1}.

3.3. Results

In Table II we report the parameters from the piecewise exponential model under the various missing data assumptions for all-cause mortality. Post-diagnosis weight loss appears to be associated with poor survival across all missing data assumptions as this is likely indicative of women with more advanced disease or otherwise less than robust health. Greater

Table II. Coefficient estimates (posterior log-hazard ratios) and 95 per cent credible intervals from piecewise exponential proportional hazards model for all-cause mortality in the Long Island Breast Cancer Study Project under different missing data assumptions.

	Posterior estimate of β (95 per cent credible interval)				
Variable	Complete-Case* ($n=499$)	MAR† ($n=1436$)	NMAR 1‡ ($n=1436$)	NMAR 2§ ($n=1436$)	NMAR 3¶ ($n=1436$)
Change in bodyweight					
>5 per cent loss	1.50 (0.93, 2.07)	1.65 (1.25, 2.07)	1.67 (1.27, 2.08)	1.66 (1.27, 2.08)	1.65 (1.26, 2.06)
5–10 per cent gain	−0.12 (−1.14, 0.79)	0.12 (−0.64, 0.80)	0.07 (−0.67, 0.77)	0.08 (−0.68, 0.77)	0.07 (−0.69, 0.77)
>10 per cent gain	0.56 (−0.28, 1.38)	1.22 (0.60, 1.81)	1.00 (0.34, 1.63)	1.01 (0.33, 1.63)	0.99 (0.33, 1.62)
Chemotherapy treatment	0.58 (0.02, 1.15)	0.67 (0.31, 1.02)	0.67 (0.32, 1.03)	0.68 (0.33, 1.03)	0.68 (0.33, 1.03)
ER positive tumor	−0.66 (−1.28, −0.04)	−0.51 (−0.86, −0.16)	−0.52 (−0.87, −0.17)	−0.52 (−0.87, −0.16)	−0.52 (−0.86, −0.16)
PR positive tumor	−0.30 (−0.89, 0.31)	−0.29 (−0.61, 0.05)	−0.29 (−0.62, 0.04)	−0.29 (−0.62, 0.04)	−0.29 (−0.61, 0.04)
Tumor size > 2 cm	0.63 (0.13, 1.13)	0.66 (0.36, 0.95)	0.67 (0.37, 0.96)	0.67 (0.36, 0.96)	0.66 (0.37, 0.96)
Age at diagnosis	0.04 (0.02, 0.07)	0.04 (0.03, 0.06)	0.04 (0.03, 0.06)	0.04 (0.03, 0.06)	0.04 (0.03, 0.05)

*Excludes subjects with missing data on one or more covariates.
†Specifies model for distribution of missing covariates.
‡Specifies model for distribution of missing covariates and missing data indicator for change in bodyweight. Precision 10^{-6} on beta.
§Specifies model for distribution of missing covariates and missing data indicator for change in bodyweight. Precision 10^{-3} on beta.
¶Specifies model for distribution of missing covariates and missing data indicator for change in bodyweight. Precision 10^{-1} on beta.

Table III. Posterior estimates of logistic regression coefficients and 95 per cent credible intervals for models for the indicator of missing weight change from selection model under NMAR assumption for all-cause mortality in the Long Island Breast Cancer Study Project.

	Posterior estimate of ϕ (95 per cent credible interval)		
Variable	At diagnosis	1-year after diagnosis	At follow-up interview
Constant	−1.78 (−2.37, −1.23)	−4.09 (−5.56, −2.68)	0.85 (−0.76, 2.39)
Percent weight change*	−0.07 (−0.14, −0.01)	−0.01 (−0.05, 0.03)	0.00 (−0.03, 0.02)
Age	0.02 (0.01, 0.03)	0.02 (0.00, 0.05)	0.00 (−0.02, 0.01)
Missing weight change at diagnosis	—	6.77 (6.15, 7.46)	6.06 (4.46, 7.74)
Missing weight change 1-year after diagnosis	—	—	3.48 (2.15, 4.96)
Time since diagnosis	—	—	−1.77 (−2.12, −1.43)

*At corresponding time point.

post diagnosis weight gain is also positively associated with all-cause mortality for all model assumptions. However, the findings in the complete-case analysis are notably attenuated compared with those that account for missing data. The magnitudes of the associations with post-diagnosis weight change are similar between the MAR and NMAR models, although the estimates of effect for either category of weight gain appear attenuated in the NMAR model compared with the MAR model. In addition, the credible interval for larger weight gain (>10%) excludes the null effect in both missing data models, yet includes it in the complete-case analysis. In both missing data models, compared with women who maintain their pre-diagnosis weight, moderate weight gain is associated with a modest increase in risk of death (MAR posterior log-hazard ratio (lnHR): 0.12, 95 per cent credible interval: −0.64, 0.80; NMAR posterior lnHR: 0.07, 95 per cent credible interval: −0.67, 0.77) while larger gain is associated with a much greater risk of death in both models (MAR posterior lnHR: 1.22, 95 per cent credible interval: 0.60, 1.81; NMAR posterior lnHR: 1.00, 95 per cent credible interval: 0.34, 1.63). Estimates of effect for the covariates appear nearly identical between the two models that account for the missing data, and the same direction and similar magnitude as the estimates from the complete-case analysis. Varying the precision on the prior distribution for the β coefficients to 10^{-3} and 10^{-1} had essentially no effect on the posterior estimates or credible intervals (models NMAR 2 and NMAR 3, respectively, Table II).

Posterior estimates of the parameters from the models explaining the probability of missing weight change at each assessment (equation (7)) and the distribution of the missing covariates (equation (8)) from NMAR model 1 are presented in Tables III and IV, respectively. The probability of missing weight change at any of the three assessment points was only modestly associated with the value of percent weight change, with the credible intervals for assessments at 1-year post-diagnosis and time of follow-up interview including the null effect (Table III). Only age at diagnosis and weight change at diagnosis were consistently associated with post-diagnosis change in bodyweight over time (Table IV). Similarly, no clear relationships emerged in the models for treatment or tumor characteristics other than age at diagnosis.

Table IV. Posterior estimates of regression coefficients and 95 per cent credible intervals from models for covariates with missing data: percent change in body weight (linear regression), chemotherapy treatment, tumor size, ER and PR status (logistic regressions) from selection model under NMAR assumption for all-cause mortality in the Long Island Breast Cancer Study Project.

		Posterior estimate of α (95 per cent credible interval)		
		Percent change in bodyweight		
Variable		At diagnosis	1-year after diagnosis	At follow-up interview
Constant		15.00 (10.79, 19.17)	7.79 (3.77, 11.87)	15.13 (10.24, 19.94)
Age		−0.13 (−0.20, −0.05)	−0.10 (−0.16, −0.03)	−0.17 (−0.25, −0.09)
Chemotherapy treatment		−1.04 (−2.23, 0.13)	−0.04 (−1.13, 1.05)	−0.82 (−2.11, 0.48)
Postmenopausal at diagnosis		1.47 (−0.12, 3.06)	−0.87 (−2.39, 0.64)	−0.05 (−1.84, 1.78)
BMI 1-year before diagnosis		−0.43 (−0.53, −0.33)	−0.03 (−0.13, 0.07)	−0.11 (−0.22, 0.01)
Percent weight change at diagnosis*		—	0.77 (0.70, 0.84)	0.63 (0.54, 0.72)
Percent weight change 1-year after diagnosis*		—	—	0.12 (0.00, 0.23)
Precision (τ)		0.017 (0.015, 0.020)	0.021 (0.019, 0.023)	0.013 (0.012, 0.014)
Variable	Chemotherapy treatment	Tumor size >2cm	ER positive tumor	PR positive tumor
Constant	4.23 (3.16, 5.39)	0.60 (−0.40, 1.60)	−0.47 (−1.12, 0.20)	0.79 (0.17, 1.40)
Age	−0.07 (−0.09, −0.06)	−0.02 (−0.04, −0.01)	0.03 (0.01, 0.04)	0.00 (−0.01, 0.01)
Income				
$ 20 000–49 999	−0.21 (−0.76, 0.35)	−0.43 (−0.94, 0.09)	—	—
$ 50 000–89 999	−0.35 (−0.95, 0.24)	−0.65 (−1.22, −0.08)	—	—
⩾$90 000	−0.23 (−0.88, 0.41)	−0.41 (−1.03, 0.22)	—	—
Education				
Some college	−0.15 (−0.50, 0.21)	−0.04 (−0.42, 0.34)	—	—
College graduate	−0.26 (−0.73, 0.20)	0.03 (−0.47, 0.52)	—	—
Post-college	−0.13 (−0.56, 0.30)	0.02 (−0.47, 0.50)	—	—

*Coded as continuous variable.

Table V. Coefficient estimates (posterior log-hazard ratios) and 95 per cent credible intervals from piecewise exponential proportional hazards model for breast cancer mortality in the Long Island Breast Cancer Study Project under different missing data assumptions.

	Posterior estimate of β (95 per cent credible interval)		
Variable	Complete-Case* ($n=499$)	MAR† ($n=1461$)	NMAR‡ ($n=1461$)
Change in bodyweight			
>5% loss	2.15 (1.30, 3.09)	1.98 (1.39, 2.61)	1.98 (1.40, 2.59)
5–10% gain	0.00 (−1.57, 1.37)	−0.13 (−1.40, 0.96)	−0.17 (−1.44, 0.90)
>10% gain	0.50 (−0.77, 1.72)	1.25 (0.39, 2.06)	1.03 (0.12, 1.87)
Chemotherapy treatment	0.85 (0.05, 1.69)	1.09 (0.56, 1.62)	1.10 (0.57, 1.64)
ER positive tumor	−0.20 (−1.10, 0.72)	−0.38 (−0.85, 0.09)	−0.39 (−0.86, 0.09)
PR positive tumor	−0.50 (−1.34, 0.37)	−0.41 (−0.85, 0.05)	−0.41 (−0.87, 0.04)
Tumor size >2cm	1.01 (0.32, 1.68)	1.06 (0.66, 1.45)	1.06 (0.67, 1.46)
Age at diagnosis	0.01 (−0.03, 0.04)	0.02 (0.00, 0.03)	0.01 (0.00, 0.03)

*Excludes subjects with missing data on one or more covariates.

†Specifies model for distribution of missing covariates.

‡Specifies model for distribution of missing covariates and missing data indicator for change in bodyweight. Precision 10^{-6} on beta.

Table V shows the results for breast cancer-related deaths for the proportional hazard regressions. As observed in the models for all-cause mortality, post-diagnosis weight loss is associated with poorer survival across all missing data assumptions, whereas the effect of larger weight gain is intensified in the models that account for missing data. The effect for moderate weight gain appears modestly protective in both the MAR and NMAR models; however, the corresponding credible intervals contain the null effect.

4. Discussion

We have presented a model for the analysis of time-to-event data with time-varying covariates when data on some covariates maybe missing and have proposed an easy to implement solution strategy. We employed this model in an

analysis of the association of longitudinal changes in bodyweight and survival after diagnosis with breast cancer in a large, population-based case-control study where we were concerned that data on post-diagnosis changes in body size may be non-ignorably missing. Our findings from the analysis suggest that weight gain after diagnosis is associated with greater mortality, both from any cause and specifically for death due to breast cancer with greater weight gain associated with a larger effect. Through sensitivity analysis we found that the point estimates for the association with post-diagnosis weight change from the complete-case analysis were somewhat attenuated compared with the models that account for the missing data. Additionally, the credible intervals were more narrow for the missing data models, resulting from the greater statistical efficiency afforded by accounting for the missing data. Models assuming an ignorable and non-ignorable missing data mechanism yielded equivalent conclusions although some differences in parameter estimates were observed.

In the model that accounted for the missing data mechanism, we failed to observe significant associations in the models for the probability that weight change was missing, as well as the models for the missing covariates. We would like to emphasize, however, that the inferential value of these ancillary models is limited, since the objective of including them was to explain the distributions that influence the relationships in the proportional hazards model in the most parsimonious manner. A more detailed analysis to evaluate reasons that data are missing should be undertaken in a framework that does not include the proportional hazards model (e.g. only considering the joint distribution of $\mathbf{r}$ and $\mathbf{z}$), which would allow for a richer specification of the $\mathbf{r}$ model with a lower risk of overparameterization.

This analysis is the first to our knowledge to examine the association between post-diagnosis weight change and survival utilizing multiple assessments on a population-based cohort of breast cancer survivors from date of diagnosis. The method we proposed here allowed us to fully utilize the available data and examine the effects of various assumptions regarding the missingness mechanism. Although our study has these unique strengths, there are some potential limitations. The study subjects in the LIBCSP are predominately older Caucasian women with higher socioeconomic status; hence, these findings may not be generalizable to the population of all breast cancer survivors. The use of proxy interviews for deceased subjects maybe problematic, however, these comprised less than 8 per cent of our sample, and previous work suggests that data from proxy and self-report interviews are highly similar [32].

Models for missing data can be useful analytic tools yet there are no substitutes for complete data. The assumptions on the missing data mechanism are untestable and in some cases selection models similar to the one we illustrate have shown to be quite sensitive to misspecification [13]. When employing selection models careful consideration must be given to the form of the model for the missing data mechanism—the desire for a thorough and accurate specification must be balanced with parsimony as convergence can become problematic for models with many parameters. Our model likely benefited from the inclusion of variables BMI before diagnosis and menopausal status at diagnosis in the model for missing weight change ($\mathbf{z}$). Those variables, through the distribution of $\mathbf{z}$ implicitly informed the portion of the model accounting for the NMAR nature of the data ($p(\mathbf{r}|\mathbf{z})$), and likely improved model fit and convergence. Although computationally intensive, the Bayesian approach to parameter estimation that we employed is easy to implement and accessible to analysts with a wide variety of computational ability. The ideal situation would be to completely observe data on all subjects, however this is unlikely to ever happen in reality, especially in longitudinal population-based studies. For cases where there is concern about the potential for covariate data to be not MAR, techniques such as the one we propose here offer a practical means of analyzing such incomplete data sets.

Appendix

Sample WinBUGS code for the NMAR selection model. The parameters `alpha`, `phi`, `beta`, `tau` and `lambda` correspond to the model specification in the text, $\alpha, \phi, \beta, \tau,$ and λ, respectively. Input variables, also described in the text are `death`, `dur`, `pcwt.fu`, `chemo`, `erstat`, `prstat`, `tsize_cat`, `dxage`, `income`, `educ`, `bmiref` and `menpstat`. The variable `fu.years` is a vector containing the times of each weight change in years since diagnosis for the corresponding vector of percent weight change variables, `pcwt.fu`. Additional scalar inputs include the sample size `N`, number of intervals for the piecewise exponential model `J`, and `eps`, a scalar equal to 10^{-7} which is used with the step function to perform logical evaluations of strict inequalities.

```
model select;
{
 for (k in 1:J+1) { a[k] <- 10*(k-1)/J; } # Partition time axis, ten evenly spaced intervals
   for (i in 1:N) {
     for (k in 1:J) {
       # Indicator if event-time in interval k
       d[i,k]<- death[i]*step(dur[i]-a[k]+eps)*step(a[k+1] - dur[i]);
```

```
  # length of overlap of dur[i] with interval k
  delta[i,k] <- (min(dur[i], a[k+1]) - a[k])*step(dur[i]-a[k]);

  # Assign exposure to correct interval:
  # Linear interpolation for percent weight change between follow-up measures.
  # Assumes that the subjects reach a constant weight after their final assessment
  pcwt[i,k] <- step(fu.years[i,2]-a[k])*(pcwt.fu[i,1] +
    (a[k]-fu.years[i,1])*(pcwt.fu[i,2] -
      pcwt.fu[i,1])/(fu.years[i,2]-fu.years[i,1])) +
    (step(a[k]-fu.years[i,2]-eps)*step(fu.years[i,3]-a[k]))*
      (pcwt.fu[i,2] + (a[k]-fu.years[i,2])*(pcwt.fu[i,3] -
        pcwt.fu[i,2])/(fu.years[i,3]-fu.years[i,2])) +
    step(a[k] - fu.years[i,3] - eps)*pcwt.fu[i,3];

  # Assign pcwt to categories
  # > 5% loss in bodyweight
    pcwt0[i,k] <- step(-5 - pcwt[i,k]+eps);
  # change in bodyweight <= 5% (gain or loss) [Maintenance, REF]
    pcwt1[i,k] <- step(pcwt[i,k] + 5)*step(5 - pcwt[i,k]);
  # change in bodyweight > than 5% and <=10%
    pcwt2[i,k] <- step(pcwt[i,k] - 5 + eps)*step(10 - pcwt[i,k]);
  # change in bodyweight > 10%
    pcwt3[i,k] <- step(pcwt[i,k] - 10 + eps);

  # Model for time to event conditional upon observed and unobserved variables
  theta[i,k] <- lambda[k]*exp(beta[1]*pcwt0[i,k] + beta[2]*pcwt2[i,k] +
    beta[3]*pcwt3[i,k] + beta[4]*chemo[i] + beta[5]*erstat[i] +
    beta[6]*prstat[i] + beta[7]*tsize_cat[i] + beta[8]*dxage[i]) ;

  # define the likelihood
  d[i,k] ~ dpois(mu[i,k]);
  mu[i,k] <- delta[i,k]*theta[i,k];
}


# Models for missing fixed (baseline) covariates
logit(p.chemo[i]) <- alpha4[1] + alpha4[2]*dxage[i] +
  alpha4[3]*equals(income[i],2) + alpha4[4]*equals(income[i],3) +
  alpha4[5]*equals(income[i],4) + alpha4[6]*equals(educ[i],2) +
  alpha4[7]*equals(educ[i],3) + alpha4[8]*equals(educ[i],4) ;
chemo[i] ~ dbin(p.chemo[i],1);

logit(p.erstat[i]) <- alpha3[1] + alpha3[2]*dxage[i];
erstat[i] ~ dbin(p.erstat[i],1);

logit(p.prstat[i]) <- alpha2[1] + alpha2[2]*dxage[i] ;
prstat[i] ~ dbin(p.prstat[i],1);

logit(p.tumor2[i]) <- alpha1[1] + alpha1[2]*dxage[i] +
  alpha1[3]*equals(income[i],2) + alpha1[4]*equals(income[i],3) +
  alpha1[5]*equals(income[i],4) + alpha1[6]*equals(educ[i],2) +
  alpha1[7]*equals(educ[i],3) + alpha1[8]*equals(educ[i],4) ;
tsize_cat[i] ~ dbin(p.tumor2[i],1);

# Models for missing time-varying covariates
# Model for weight change at 3rd f/u
mu.pcwt[i,3] <- alpha7[1] + alpha7[2]*dxage[i] + alpha7[3]*chemo[i] +
```

```
    alpha7[4]*postmenp[i] + alpha7[5]*bmiref[i] + alpha7[6]*pcwt.fu[i,2] +
    alpha7[7]*pcwt.fu[i,1];
  pcwt.fu[i,3] ~ dnorm(mu.pcwt[i,3],tau3);

  # Model for weight change at 2nd f/u (1-year post-diagnosis)
  mu.pcwt[i,2] <- alpha6[1] + alpha6[2]*dxage[i] + alpha6[3]*chemo[i] +
    alpha6[4]*postmenp[i] + alpha6[5]*bmiref[i] + alpha6[6]*pcwt.fu[i,1];
  pcwt.fu[i,2] ~ dnorm(mu.pcwt[i,2],tau2);

  # Model for weight change at 1st f/u (at diagnosis)
  mu.pcwt[i,1] <- alpha5[1] + alpha5[2]*dxage[i] + alpha5[3]*chemo[i] +
    alpha5[4]*postmenp[i] + alpha5[5]*bmiref[i];
  pcwt.fu[i,1] ~ dnorm(mu.pcwt[i,1],tau1);

  # Models for missing time-varying covariates
  # Missingness model for third f/u
  logit(p.r.fu[i,3]) <- phi3[1] + phi3[2]*pcwt.fu[i,3] + phi3[3]*dxage[i] +
    phi3[4]*fu.years[i,3] + phi3[5]*r.fu[i,1] + phi3[6]*r.fu[i,2];
  r.fu[i,3] ~ dbin(p.r.fu[i,3],1);

  # Missingness model for second f/u
  logit(p.r.fu[i,2]) <- phi2[1] + phi2[2]*pcwt.fu[i,2] + phi2[3]*dxage[i] +
    phi2[4]*r.fu[i,1];
  r.fu[i,2] ~ dbin(p.r.fu[i,2],1);

  # Missingness model for first f/u
  logit(p.r.fu[i,1]) <- phi1[1] + phi1[2]*pcwt.fu[i,1] + phi1[3]*dxage[i] ;
  r.fu[i,1] ~ dbin(p.r.fu[i,1],1);
  }

  # PRIORS ON PARAMETERS
  # Parameters for survival model
  for (k in 1:J) { lambda[k] ~ dgamma(0.01, 0.01); }
  for (l in 1:8) { beta[l] ~ dnorm(0, 0.000001); }

  # Regression parameters for pcwt model
  for (l in 1:7) { alpha7[l] ~ dnorm(0, 0.000001); }
  for (l in 1:6) { alpha6[l] ~ dnorm(0, 0.000001); }
  for (l in 1:5) { alpha5[l] ~ dnorm(0, 0.000001); }

  # Variance parameters for missing data models
  tau3 ~ dgamma(0.01, 0.01);
  tau2 ~ dgamma(0.01, 0.01);
  tau1 ~ dgamma(0.01, 0.01);

  # Regression parameters for missingness model
  for (l in 1:6) { phi3[l] ~ dnorm(0, 0.000001); }
  for (l in 1:4) { phi2[l] ~ dnorm(0, 0.000001); }
  for (l in 1:3) { phi1[l] ~ dnorm(0, 0.000001); }

  # Regression parameters for treatment and tumor characteristic models
  for (l in 1:8) { alpha1[l] ~ dnorm(0, 0.000001);}
  for (l in 1:2) { alpha2[l] ~ dnorm(0,0.000001);}
  for (l in 1:2) { alpha3[l] ~ dnorm(0,0.000001);}
  for (l in 1:8) { alpha4[l] ~ dnorm(0, 0.000001);}
}
```

Acknowledgements

The authors wish to thank the editor, associate editor and reviewers for helpful comments and suggestions, which have led to an improvement of this article. This work was supported in part by grants from the National Institutes of Health. Dr Bradshaw's work was supported by grants #T32CA009330 and #T32HL007055. Dr Ibrahim's research was partially supported by grants #GM 70335 and #CA 74015. Dr Gammon's work supported in part by grants #UO1CA/ES66572 and #P30ES10126.

References

1. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. Wiley: New York, 2002.
2. Chen HY, Little RJ. Proportional hazards regression with missing covariates. *Journal of the American Statistical Association* 1999; **94**:896–908.
3. Herring AH, Ibrahim JG. Likelihood-based methods for missing covariates in the Cox proportional hazards model. *Journal of the American Statistical Association* 2001; **96**:292–302.
4. Lin DY, Ying Z. Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association* 1993; **88**: 1341–1349.
5. Lipsitz SR, Ibrahim JG. Estimating equations with incomplete categorical covariates in the Cox model. *Biometrics* 1998; **54**:1002–1013.
6. Martinussen T. Cox regression with incomplete covariate measurements using the EM-algorithm. *Scandinavian Journal of Statistics* 1999; **26**:479–491.
7. Paik MC. Multiple imputation for the Cox proportional hazards model with missing covariates. *Lifetime Data Analysis* 1997; **3**:289–298.
8. Paik MC, Tsai W-Y. On using the Cox proportional hazards model with missing covariates. *Biometrika* 1997; **84**:579–593.
9. Reilly M, Pepe MS. A mean score method for missing and auxiliary covariate data in regression models. *Biometrika* 1995; **82**:299–314.
10. Zhou H, Pepe MS. Auxiliary covariate data in failure time regression. *Biometrika* 1995; **82**:139–149.
11. Ibrahim JG, Chen MH, Sinha D. *Bayesian Survival Analysis*. Springer: New York, 2001.
12. Leong T, Lipsitz SR, Ibrahim JG. Incomplete covariates in the Cox model with applications to biological marker data. *Applied Statistics* 2001; **50**:467–484.
13. Herring AH, Ibrahim JG, Lipsitz SR. Non-ignorable missing covariate data in survival analysis: a case-study of an International Breast Cancer Study Group trial. *Applied Statistics* 2004; **53**:293–310.
14. Kaaks R, McTiernan A. Obesity and sex hormones. In *Cancer Prevention and Management Through Exercise and Weight Control*, McTiernan A (ed.). CRC Press: Boca Raton, 2005; 289–300.
15. Blackburn G, Waltman B. Obesity and insulin resistance. In *Cancer Prevention and Management Through Exercise and Weight Control*, McTiernan A (ed.). CRC Press: Boca Raton, 2005; 301–316.
16. Goodwin PJ. Energy balance and cancer prognosis, breast cancer. In *Cancer Prevention and Management Through Exercise and Weight Control*, McTiernan A (ed.). CRC Press: Boca Raton, 2005; 405–435.
17. Gammon MD, Neugut AI, Santella RM, Teitelbaum SL, Britton JA, Terry MB, Eng SM, Wolff MS, Stellman SD, Kabat GC, Levin B, Bradlow HL, Hatch M, Beyea J, Camann D, Trent M, Senie RT, Garbowski GC, Maffeo C, Montalvan P, Berkowitz GS, Kemeny M, Citron M, Schnabe F, Schuss A, Hajdu S, Vincguerra V, Collman GW, Obrams GI. The Long Island Breast Cancer Study Project: description of a multi-institutional collaboration to identify environmental risk factors for breast cancer. *Breast Cancer Research and Treatment* 2002; **74**:235–254.
18. Ibrahim JG, Lipsitz SR, Chen MH. Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society B* 1999; **61**:173–190.
19. Stubbendick AL, Ibrahim JG. Maximum likelihood methods for nonignorable missing responses and covariates in random effects models. *Biometrics* 2003; **59**:1140–1150.
20. Ibrahim JG, Chen MH, Lipsitz SR. Missing responses in generalized linear mixed models when the missing data mechanism is nonignorable. *Biometrika* 2001; **88**:551–564.
21. Lipsitz SR, Ibrahim JG, Fitzmaurice GM. Likelihood methods for incomplete longitudinal binary responses with incomplete categorical covariates. *Biometrics* 1999; **55**:214–223.
22. Huang L, Chen MH, Ibrahim JG. Bayesian analysis for generalized linear models with missing covariates. *Biometrics* 2005; **61**(3):767–780. DOI: 10.1198/016214508000001057.
23. Ibrahim JG, Zhu H, Tang N. Model selection criteria for missing-data problems using the EM algorithm. *Journal of the American Statistical Association* 2008; **103**(484):1648–1658. DOI: 10.1198/016214508000001057.
24. Lipsitz SR, Ibrahim JG. A conditional model for incomplete covariates in parametric regression models. *Biometrika* 1996; **83**:125–134.
25. Casella G, George EI. Explaining the Gibbs Sampler. *The American Statistician* 1992; **46**:167–174.
26. Fink BN, Gaudet MM, Britton JA, Abrahamson PE, Teitelbaum SL, Jacobson J, Bell P, Thomas JA, Kabat GC, Neugut AI, Gammon MD. Fruits, vegetables, micronutrient intake in relation to breast cancer survival. *Breast Cancer Research and Treatment* 2006; **98**:199–208.
27. Cowper DC, Kubal JD, Maynard C, Hynes DM. A primer and comparative review of major US mortality databases. *Annals of Epidemiology* 2002; **12**:462–468.
28. Demark-Wahnefried W, Rimer BK, Winer EP. Weight gain in women diagnosed with breast cancer. *Journal of the American Dietetic Association* 1997; **97**:519–526; 529; quiz 527–528.
29. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS—a Bayesian modelling framework: concepts, structure and extensibility. *Statistics and Computing* 2000; **10**:325–337.
30. Geweke J. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics 4*, Bernardo JM, Berger JO, Dawid AP, Smith AFM (eds). Clarendon Press: Oxford, U.K., 1992.
31. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, 2004. ISBN 3-900051-07-0. Available from: http://www.R-project.org.
32. Campbell PT, Sloan M, Kreiger N. Utility of proxy versus index respondent information in a population-based case-control study of rapidly fatal cancers. *Annals of Epidemiology* 2007; **17**:253–257.