Published in final edited form as:

Stat Med. 2010 September 30; 29(22): 2310–2324. doi:10.1002/sim.3977.

Latent class joint model of ovarian function suppression and DFS for premenopausal breast cancer patients

Jenny J. Zhang^{a,*†} and Molin Wang^b

^aDepartment of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Building 2, Boston, MA 02115, U.S.A.

^bHarvard School of Public Health and Dana Farber Cancer Institute, U.S.A.

Abstract

Breast cancer is the leading cancer in women of reproductive age; more than a quarter of women diagnosed with breast cancer in the US are premenopausal. A common adjuvant treatment for this patient population is chemotherapy, which has been shown to cause premature menopause and infertility with serious consequences to quality of life. Luteinizing-hormone-releasing hormone (LHRH) agonists, which induce *temporary* ovarian function suppression (OFS), has been shown to be a useful alternative to chemotherapy in the adjuvant setting for estrogen-receptor-positive breast cancer patients. LHRH agonists have the potential to preserve fertility after treatment, thus, reducing the negative effects on a patient's reproductive health. However, little is known about the association between a patient's underlying degree of OFS and disease-free survival (DFS) after receiving LHRH agonists. Specifically, we are interested in whether patients with lower underlying degrees of OFS (i.e. higher estrogen production) *after* taking LHRH agonists are at a higher risk for late breast cancer events. In this paper, we propose a latent class joint model (LCJM) to analyze a data set from International Breast Cancer Study Group (IBCSG) Trial VIII to investigate the association between OFS and DFS. Analysis of this data set is challenging due to the fact that the main outcome of interest, OFS, is unobservable and the available surrogates for this latent variable involve masked event and cured proportions. We employ a likelihood approach and the EM algorithm to obtain parameter estimates and present results from the IBCSG data analysis.

Keywords

joint modeling; latent class model; EM algorithm; masked event; cured proportion; breast cancer

1. Introduction

Breast cancer is the most common cancer and the second leading cause of cancer deaths (after lung cancer) among women in the US. About 25 per cent of all women diagnosed with breast cancer are premenopausal, making breast cancer the leading cancer in women of reproductive age [1]. Adjuvant therapy is often given to premenopausal breast cancer patients to prevent disease recurrence or progression after surgery. One common form of such adjuvant treatment is chemotherapy, which often causes premature ovarian failure leading to infertility in many young women with breast cancer [2]. In addition to its effects on fertility, the premature menopause induced by chemotherapy also causes negative side effects such as hot flashes,

osteoporosis, and sexual dysfunction, all of which contribute to poorer patient quality of life and perception of health [3]. Thus, fertility preservation and the effects of adjuvant treatments on reproductive health are playing increasingly more important roles in the treatment decision-making process of young women diagnosed with breast cancer.

Much recent attention has been focused on another form of adjuvant therapy which induces *temporary* ovarian function suppression (OFS), luteinizing-hormone-releasing hormone (LHRH) agonists. LHRH agonists greatly reduce the production of oestrogen, achieved through LHRH receptor downregulation. Once off treatment, most patients resume their ovarian function unless they entered menopause during treatment. It has been shown that LHRH agonists are a useful alternative to chemotherapy for estrogen-receptor (ER)-positive patients in this population due to the reversibility and tolerability of its effects [4]. ER-positive patients have oestrogen receptors on their breast cancer cells. When oestrogen binds to those receptors, the cancer cells are stimulated to grow and replicate, leading to disease progression or recurrence. It follows that ER-positive breast cancer patients are expected to respond to treatments that block/reduce oestrogen, which is the mechanism of LHRH agonists.

LHRH agonists give young women with breast cancer an increased chance of preserving their fertility and reproductive health after adjuvant therapy. It is unclear though whether patients with lower underlying degrees of OFS *after* taking LHRH agonists may be at a higher risk for late breast cancer events. Patients' underlying degrees of OFS after treatment will differ depending on their individual characteristics, and those with lower degrees may have a higher chance of preserving their reproductive health. However, if having lower degrees of OFS puts those patients at a higher risk for disease recurrence, then the appeal of the treatment would be greatly reduced. Thus, there is interest in the association between a premenopausal patient's underlying degree of OFS and her disease-free survival (DFS) after LHRH agonists. The availability of such information could affect how this patient population is treated in the adjuvant setting, especially those patients who still wish to become pregnant after treatment. In this paper, we present a novel approach for the analysis of data from International Breast Cancer Study Group (IBCSG) Trial VIII to investigate the extent to which OFS is associated with DFS in premenopausal breast cancer patients.

As OFS is not easily quantifiable or measurable, it is natural to consider a latent class variable to represent a patient's underlying degree of OFS. As part of the study, IBCSG Trial VIII carefully collected monthly menses data during the first 36 months and then every 6 months thereafter; details are given in [5]. As a common indicator of OFS is the cessation of menses, we propose to obtain information about latent class membership for OFS through a set of time-to-event surrogates extracted from the menses data. Specifically, the menses data contain information about the times to treatment-induced amenorrhea (TIA), menopause, and recovery of menses; TIA is defined to be a cessation of menses that is induced by adjuvant therapy. The process of TIA, however, is complicated by the fact that menopause may also occur. As both TIA and menopause result in an observed cessation of menses (the failure), the event leading to the failure is masked unless the patient recovers her menses after treatment. Although all patients are expected to experience TIA after taking LHRH agonists, we allow our method to be more general by including a cure rate for TIA, which would equal 1 if all patients experienced TIA. Furthermore, not all patients who experience TIA will recover their menses after treatment, thus, we also assume the existence of a cured proportion with respect to recovery. No cure rate is assumed for menopause since all patients are expected to experience menopause eventually. To clarify, we use the term 'cured' in reference to a subset of patients who are assumed to be unsusceptible to some event. Thus, analysis of this IBCSG data set poses several statistical challenges, including latent variables, masked event, and cured proportions.

As our main interest is in the association of OFS and DFS, we propose to jointly model the latent class variable and DFS. Joint modeling has garnered much interest in recent years. Initial work was done mostly in the area of longitudinal studies to characterize the association between the longitudinal and event time processes (e.g. [6–12]). More recently, joint models have also been applied to social science data, where the interest is in the association between a latent variable and time-to-event data (e.g. [13–17]); we refer to such models as latent class joint models. Specifically, the Cox model is extended to include a latent class variable as a predictor of time-to-event while using information about latent class membership from a latent class regression model.

In this paper, we propose a latent class joint model (LCJM) to investigate the association between a premenopausal breast cancer patient's underlying degree of OFS and DFS using the IBCSG Trial VIII data, where information on class membership is obtained indirectly through a set of menses-related, time-to-event surrogates involving masked event and cured proportions. In Section 2, we describe the IBCSG Trial VIII data in greater detail. In Sections 3 and 4, we present our LCJM and detail the likelihood formation and parameter estimation via the EM algorithm, respectively. In Section 5, we present the results from fitting the proposed model to the IBCSG Trial VIII data, and close with some discussion in Section 6.

2. IBCSG Trial VIII

In IBCSG Trial VIII [18], premenopausal, node-negative breast cancer patients were randomized to four different treatment arms and stratified according to ER status, whether radiotherapy was planned after surgery, and institution. Patients were randomized to I: no adjuvant therapy, II: goserelin for 24 months, III: six cycles of CMF (chemotherapy), and IV: six cycles of CMF followed by goserelin for 18 months, where goserelin is an LHRH agonist. The primary endpoint was DFS, defined to be the time from the date of randomization to any recurrent disease, appearance of a secondary primary cancer, or death, whichever occurred first. The no adjuvant therapy arm was discontinued less than two years into the study, and the median follow-up of the trial was 7 years. As our interest lies in the association between OFS and DFS after LHRH agonists, we will focus on the goserelin arm of IBCSG Trial VIII.

The following clinically justified assumptions are made concerning the three time-to-event surrogates used to determine latent class membership (i.e. times to TIA, menopause, and recovery). We assume that TIA cannot occur after treatment end (U) and recovery of menses can only occur after U , whereas menopause and DFS can occur any time. We further assume that observing recovery of menses would successfully unmask the event leading to cessation before U as TIA. Otherwise, the observed cessation of menses before U could be due to either TIA or menopause. To specify a more general model, we assume the possible existence of patients who are unsusceptible (i.e. cured) to TIA; as TIA can only occur before U , if no failure is observed before U , then the subject can be assumed to be cured with respect to TIA. In our IBCSG trial application, however, we expect TIA to occur in all patients taking LHRH agonists. Moreover, menopause may also occur after TIA, with or without recovery first, and recovery after menopause is not possible.

3. The model

Figure 1 is a schematic of our LCJM, which consists of three components: (i) latent class membership; (ii) menses-related, time-to-event surrogates for the latent class; and (iii) hazard for DFS. Components (i) and (ii) are what is usually referred to as the latent class model (LCM). We propose to use Weibull proportional hazards (PH) models for (ii) and a semiparametric PH model for (iii).

Let n and K denote the total number of subjects and latent classes, respectively, and c_{ik} be the indicator that subject i belongs to class k for $i = 1, \dots, n$ and $k = 1, \dots, K$. We adopt the generalized logit [19] to model the relationship between latent class (i.e. degree of OFS) and the $(p \times 1)$ vector of covariates, $\mathbf{W}_i$:

$$p_{ik} = P(c_{ik} = 1 | \mathbf{w}_i) = \frac{\exp(\mathbf{w}_i' \boldsymbol{\eta}_k)}{\sum_{s=1}^K \exp(\mathbf{w}_i' \boldsymbol{\eta}_s)}, \quad (1)$$

where p_{ik} is the probability that subject i belongs to class k and $\boldsymbol{\eta}_k$ is a vector of regression parameters for class k with $\boldsymbol{\eta}_K = \mathbf{0}$. Note that all vectors are column vectors and a prime, $'$, denotes the transpose.

For each class, we model the vector of menses-related, time-to-event surrogates, $\mathbf{Y} = (T_1, T_2, T_3)'$, where T_1 is time to TIA with support $[0, U]$, T_2 is time to menopause with support $[0, \infty)$, and T_3 is time to recovery of menses with support (U, ∞) . Recall that U is treatment end. Let $f_{T_m}(t_m | c_{ik} = 1)$ and $F_{T_m}(t_m | c_{ik} = 1)$ denote the conditional probability and cumulative distribution functions of T_m , $m = 1, 2, 3$, respectively. We assume that the time-to-event surrogates are mutually independent conditional on class membership, that is

$$f_{\mathbf{Y}}(\mathbf{y}_i | c_{ik} = 1) = \prod_{m=1}^3 f_{T_m}(t_{mi} | c_{ik} = 1). \quad (2)$$

Let τ_{1i} and τ_{2i} denote the cure indicators for TIA and recovery, respectively ($0 = \text{cured}$; $1 = \text{uncured}$) for subject i , and $\pi_{1ki} = P(\tau_{1i} = 1 | c_{ik} = 1)$ and $\pi_{2ki} = P(\tau_{2i} = 1 | c_{ik} = 1)$. Note that $\tau_{2i} = 1$ implies $\tau_{1i} = 1$ since recovery is undefined when the patient is cured with respect to TIA. Letting $S_{T_m}(\cdot)$ denote the survivor function for T_m , we adopt the following model specifications for T_1 and T_3 :

$$S_{T_1}(t_{1i} | c_{ik} = 1) = \pi_{1ki} S_{T_1}(t_{1i} | \tau_{1i} = 1, c_{ik} = 1) + (1 - \pi_{1ki}),$$

and

$$S_{T_3}(t_{3i} | c_{ik} = 1) = \pi_{2ki} S_{T_3}(t_{3i} | \tau_{2i} = 1, c_{ik} = 1) + (1 - \pi_{2ki}).$$

As TIA cannot occur after treatment end (U), we model the conditional distribution of time to TIA, T_1 , using a truncated Weibull PH model with hazard

$$\lambda_{T_1}(t_{1i} | t_{1i} \leq u_i, \tau_{1i} = 1, c_{ik} = 1) = \left\{ \frac{a_1 d_1 (a_1 t_{1i})^{d_1 - 1} \exp[-(a_1 t_{1i})^{d_1}]}{\exp[-(a_1 t_{1i})^{d_1}] - \exp[-(a_1 u_i)^{d_1}]} \right\} \exp(\alpha_k),$$

where the $(K \times 1)$ parameter vector $\boldsymbol{\alpha}$ contains the effect of the class variable on the hazard of TIA and (a_1, d_1) are the Weibull parameters. No truncation is needed for time to menopause, T_2 , or time to recovery, T_3 , and we model the distribution of T_2 and the conditional distribution of T_3 using Weibull PH models with respective hazards

$$\lambda_{T_2}(t_{2i} | c_{ik} = 1) a_2 d_2 (a_2 t_{2i})^{d_2 - 1} \exp(\beta_k),$$

and

$$\lambda_{T_3}(t_{3i}|\tau_{2i}=1, c_{ik}=1)=a_3 d_3 (a_3 t_{3i})^{d_3-1} \exp(\gamma_k),$$

where β and γ are the $(K \times 1)$ parameter vectors containing the effects of the class variable on the hazard of menopause and recovery of menses, respectively, and (a_2, d_2) and (a_3, d_3) are the respective Weibull parameters. For identifiability, $\alpha_K = \beta_K = \gamma_K = 0$, where K is the total number of latent classes. Note that the hazard for T_3 is conditional on being uncured with respect to recovery, and since we assume that recovery can only occur after U , $\lambda_{T_3}(t_{3i}|\tau_{2i}=1, c_{ik}=1)$ is implicitly bounded below by U .

We assume conditional independence between $\mathbf{W}_i$ and $\mathbf{Y}_i$ given class, i.e. $f_{\mathbf{Y}}(\mathbf{y}_i | c_{ik}=1, \mathbf{w}_i) = f_{\mathbf{Y}}(\mathbf{y}_i | c_{ik}=1)$, which implies that the latent classes (i.e. degree of OFS) serve as a summary of the information in the menses-related, time-to-event surrogates. This assumption is an important feature of the LCM and simplifies both the modeling and estimation procedures. One way to investigate this assumption is to add the variables in $\mathbf{W}_i$ one at a time for each of the time-to-event surrogates. More specifically, let the class-specific hazards depend on a potentially biasing covariate, ω_i , such that $\lambda_{T_m}(\cdot | c_i, \omega_i) = \lambda_{0m}(\cdot) \exp(\rho_m \omega_i + \phi_{m,c_i})$, where c_i takes one of the values $1, \dots, K$ denoting latent class, and ρ_m and ϕ_{m,c_i} are surrogate-specific parameters (i.e. for T_m , $m = 1, 2, 3$). Then, perform likelihood ratio tests (LRTs) comparing this extended model with the original LCM. If the LRTs are non-significant, then the assumption holds. A similar method was used in [15].

Figure 2 defines the five possible observed data configurations for IBCSG Trial VIII, denoted by $\phi_i = 1, \dots, 5$ for the i th subject. Let X_i be the time to cessation of menses (if a cessation is observed) and ζ_i be the indicator of the event leading to failure ($0 = \text{TIA}$; $1 = \text{menopause}$). If X_i is observed before treatment end (U) and cannot be identified as caused by either TIA or menopause (i.e. recovery is not observed), then τ_{1i} , τ_{2i} , and ζ_i are all unknown and we define $\phi_i = 1$. If X_i is observed before U and recovery is also observed, then X_i is unmasked as TIA. It follows directly that $\tau_{1i} = \tau_{2i} = 1$ and $\zeta_i = 0$, and we define ϕ_i based on whether ($\phi_i = 3$) or not ($\phi_i = 2$) menopause is observed after recovery. If X_i is not observed before U , then $\tau_{1i} = 0$ and we define ϕ_i according to whether ($\phi_i = 5$) or not ($\phi_i = 4$) menopause is observed after U . We assume that all censoring of subjects occur after U , which is the case for IBCSG Trial VIII.

Finally, we propose a PH model conditional on latent class membership and the vector of covariates, $\mathbf{Z}_i$, with unspecified baseline hazard, $\lambda_0^*(\cdot)$, for DFS time T^* . The hazard at time t_i^* for a DFS event is modeled as:

$$\lambda_{T^*}(t_i^* | c_{ik}=1, \mathbf{z}_i) = \lambda_0^*(t_i^*) \exp(\boldsymbol{\kappa}' \mathbf{Z}_i + \nu_k), \quad (3)$$

where ν is the $(K \times 1)$ parameter vector containing the effect of the class variable on the hazard of DFS and $\boldsymbol{\kappa}$ is a vector of regression parameters corresponding to $\mathbf{z}_i$. Letting G_i denote the censoring time for DFS, the observed variables are $V_i = \min(T_i^*, G_i)$ and $\Delta_i = 1(T_i^* \leq G_i)$. Under the assumption of noninformative censoring, the distribution of (V_i, Δ_i) is proportional to

$$\{\lambda_0^*(v_i) \exp(\boldsymbol{\kappa}' \mathbf{z}_i + \nu_k)\}^{\delta_i} \exp\{-\Lambda_0^*(v_i) \exp(\boldsymbol{\kappa}' \mathbf{z}_i + \nu_k)\},$$

where $\Lambda_0^*(\cdot)$ is the cumulative baseline hazard for DFS.

For a fixed number of latent classes, K , let

$\Theta = \{\pi_1, \pi_2, \eta, a_1, d_1, \alpha, a_2, d_2, \beta, a_3, d_3, \gamma, \lambda_0^*(\cdot), \kappa, \nu\}$ denote the complete parameter vector, where $\pi_1, \pi_2, \alpha, \beta, \gamma$, and ν are all $(K \times 1)$ vectors and η is a $(K \times p)$ matrix.

4. Estimation

4.1. Parameter estimation

We employ a likelihood approach to estimate the parameters, Θ , for a fixed number of latent classes, K . As direct maximization of the observed likelihood is difficult, we employ the Expectation-Maximization (EM) algorithm [20] to estimate Θ . The complete data likelihood is

$$L_{\text{comp}}(\Theta|\mathbf{P}) = \prod_{r=1}^5 \prod_{i=1}^n p_{ik} f(P_i|\Theta, \varphi_i=r) \{\lambda_0^*(v_i) \exp(\kappa' \mathbf{z}_i + \nu_k)\}^{\delta_i} \times \exp\{-\Lambda_0^*(v_i) \exp(\kappa' \mathbf{z}_i + \nu_k)\},$$

where p_{ik} is given in (1), $P_i = \{\mathbf{Y}_i, \phi_i, \mathbf{W}_i, \mathbf{Z}_i, \tau_{1i}, \tau_{2i}, \zeta_i, \xi_i, \vartheta_i, c_i\}$, $\xi_i = 1(T_{2i} < U)$, $\vartheta_i = 1(T_{2i} > G_i)$, $\mathbf{P} = \{P_i, V_i, \Delta_i, i = 1, \dots, n\}$, and ϕ_i indicates one of the five possible observed data configurations defined in Figure 2. We assume that, given the degree of OFS, there is no additional effect of the time-to-event surrogates on DFS.

We carefully consider each data configuration, $f(P_i|\Theta, \phi_i=r)$, $r = 1, \dots, 5$, given in Appendix A. $L_{\text{comp}}(\Theta|\mathbf{P})$ can be factored into seven pieces, each containing a distinct set of parameters in Θ , such that

$L_{\text{comp}}(\Theta|\mathbf{P}) = L_{c1}(\pi_1|\mathbf{P}) L_{c2}(\pi_2|\mathbf{P}) L_{c3}(a_1, d_1, \alpha|\mathbf{P}) L_{c4}(a_2, d_2, \beta|\mathbf{P}) \times L_{c5}(a_3, d_3, \gamma|\mathbf{P}) L_{c6}(\eta|\mathbf{P}) L_{c7}(\lambda_0^*(\cdot), \kappa, \nu|\mathbf{P})$, where

$$\begin{aligned} L_{c1}(\pi_1|\mathbf{P}) &= \prod_{i:\varphi_i=1} \pi_{1k_i}^{\{\tau_{1i}\zeta_i + (1-\zeta_i)\}} (1 - \pi_{1k_i})^{(1-\tau_{1i})\zeta_i} \prod_{i:\varphi_i \in \{2,3\}} \pi_{1k_i} \prod_{i:\varphi_i \in \{4,5\}} (1 - \pi_{1k_i}), \\ L_{c2}(\pi_2|\mathbf{P}) &= \prod_{i:\varphi_i=1} \pi_{2k_i}^{\tau_{2i}(1-\zeta_i)} (1 - \pi_{2k_i})^{(1-\tau_{2i})(1-\zeta_i)} \prod_{i:\varphi_i \in \{2,3\}} \pi_{2k_i}, \\ L_{c3}(a_1, d_1, \alpha|\mathbf{P}) &= \prod_{i:\varphi_i=1} S_{T_1}(x_i | t_{1i} \leq u_i, \tau_{1i}=1, c_{ik}=1)^{\tau_{1i}\zeta_i} f_{T_1}(x_i | t_{1i} \leq u_i, \tau_{1i}=1, c_{ik}=1)^{(1-\zeta_i)} \times \prod_{i:\varphi_i \in \{2,3\}} f_{T_1}(t_{1i} | t_{1i} \leq u_i, \tau), \\ L_{c4}(a_2, d_2, \beta|\mathbf{P}) &= \prod_{i:\varphi_i=1} f_{T_2}(x_i | c_{ik}=1)^{\zeta_i} f_{T_2}(t_{2i} | c_{ik}=1)^{(1-\zeta_i)} \prod_{i:\varphi_i \in \{2,4\}} S_{T_2}(g_i | c_{ik}=1) \prod_{i:\varphi_i \in \{3,5\}} f_{T_2}(t_{2i} | c_{ik}=1), \\ L_{c5}(a_3, d_3, \gamma|\mathbf{P}) &= \prod_{i:\varphi_i=1} S_{T_3}(t_{2i} - u_i | \tau_{2i}=1, c_{ik}=1)^{\tau_{2i}(1-\zeta_i)(1-\xi_i)(1-\vartheta_i)} S_{T_3}(g_i - u_i | \tau_{2i}=1, c_{ik}=1)^{\tau_{2i}(1-\zeta_i)\vartheta_i} \times \prod_{i:\varphi_i \in \{2,3\}} f_{T_3}(t_{2i}), \\ L_{c6}(\eta|\mathbf{P}) &= \prod_{i=1}^n \left\{ \frac{\exp(\mathbf{w}_i' \eta_k)}{\sum_{s=1}^K \exp(\mathbf{w}_i' \eta_s)} \right\}, \end{aligned}$$

and $L_{c7}(\lambda_0^*(\cdot), \kappa, \nu|\mathbf{P})$

$$\begin{aligned} &= \prod_{i=1}^n \{\lambda_0^*(v_i) \exp(\kappa' \mathbf{z}_i + \nu_k)\}^{\delta_i} \exp\{-\Lambda_0^*(v_i) \exp(\kappa' \mathbf{z}_i + \nu_k)\}, \\ &= \prod_{j=1}^J \prod_{i=1}^n \{\lambda_0^*(t_{(j)}^*) \exp(\kappa' \mathbf{z}_i + \nu_k)\}^{\delta_i 1(V_i=t_{(j)}^*)} \exp\{-\lambda_0^*(t_{(j)}^*) \exp(\kappa' \mathbf{z}_i + \nu_k)\}^{1(V_i \geq t_{(j)}^*)}, \end{aligned}$$

where $t_{(1)}^* < \dots < t_{(J)}^*$ denotes the J distinct DFS times. For $L_{c7}(\cdot)$, we treat DFS time as discrete to estimate the baseline hazard, $\lambda_0^*(\cdot)$, at each event time.

The E-step of the EM algorithm computes the conditional expectation of the complete data log-likelihood, $l_{\text{comp}}(\boldsymbol{\Theta}|\mathbf{P}) = \log\{L_{\text{comp}}(\boldsymbol{\Theta}|\mathbf{P})\}$, with respect to the latent variables, $\mathbf{H} = \{\tau_{1i}, \tau_{2i}, \zeta_i, \xi_i, \vartheta_i, c_i, i=1, \dots, n\}$, given the observed data, $\mathbf{O} = \{X_i, \phi_i, V_i, \delta_i, \mathbf{W}_i, \mathbf{Z}_i, i=1, \dots, n\}$, and the latest updated parameter estimates, $\boldsymbol{\Theta}^{(L)}$. For ease of presentation, we let $\omega_{ik}^{[l, \varphi_i]}$, $l=1, \dots, 11$, denote the needed conditional expectations for ϕ_i in the E-step, which are defined and derived in Appendix B.

The M-step updates the estimate of $\boldsymbol{\Theta}$ by maximizing the expected complete data log-likelihood,

$$\begin{aligned} E[l_{\text{comp}}(\boldsymbol{\Theta}|\mathbf{P}, \boldsymbol{\Theta}^{(L)})] \\ = E[l_{c1}(\pi_1|\mathbf{P}, \boldsymbol{\Theta}^{(L)})] \\ + E[l_{c2}(\pi_2|\mathbf{P}, \boldsymbol{\Theta}^{(L)})] \\ + E[l_{c3}(a_1, d_1, \alpha|\mathbf{P}, \boldsymbol{\Theta}^{(L)})] + E[l_{c4}(a_2, d_2, \beta|\mathbf{P}, \boldsymbol{\Theta}^{(L)})] \\ + E[l_{c5}(a_3, d_3, \gamma|\mathbf{P}, \boldsymbol{\Theta}^{(L)})] \\ + E[l_{c6}(\eta|\mathbf{P}, \boldsymbol{\Theta}^{(L)})] \\ + E[l_{c7}(\lambda_0^*(\cdot), \boldsymbol{\kappa}, \nu|\mathbf{P}, \boldsymbol{\Theta}^{(L)})], \end{aligned}$$

with respect to $\boldsymbol{\Theta}$ for fixed $\omega_{ik}^{[l, \varphi_i]}$. Detailed formulation of $E[l_{\text{comp}}(\boldsymbol{\Theta}|\mathbf{P}, \boldsymbol{\Theta}^{(L)})]$ is given in Appendix C.

Closed-form solutions can be obtained for π_1 , π_2 , and $\lambda_0^*(\cdot)$, respectively, in the M-step as

$$\hat{\pi}_{1k} = \frac{\sum_{i:\varphi_i=1}(\omega_{ik}^{[3,1]} + \omega_{ik}^{[4,1]}) + \sum_{i:\varphi_i \in \{2,3\}} \omega_{ik}^{[1, \{2,3\}]}}{\sum_{i:\varphi_i=1}(\omega_{ik}^{[3,1]} + \omega_{ik}^{[4,1]} + \omega_{ik}^{[5,1]}) + \sum_{i:\varphi_i \neq 1} \omega_{ik}^{[1, \{2,3,4,5\}]}}}, \quad \hat{\pi}_{2k} = \frac{\sum_{i:\varphi_i=1} \omega_{ik}^{[6,1]} + \sum_{i:\varphi_i \in \{2,3\}} \omega_{ik}^{[1, \{2,3\}]}}{\sum_{i:\varphi_i=1}(\omega_{ik}^{[6,1]} + \omega_{ik}^{[7,1]}) + \sum_{i:\varphi_i \in \{2,3\}} \omega_{ik}^{[1, \{2,3\}]}}},$$

and

$$\hat{\lambda}_0^*(t^*) = \sum_{i=1}^n \frac{\delta_i 1(V_i=t^*)}{\sum_{i^\dagger \in R_j} \exp(\boldsymbol{\kappa}' \mathbf{z}_{i^\dagger}) \sum_{k=1}^K \omega_{i^\dagger k}^{[1, \{all\}]} \exp(\nu_k)},$$

where R_j is the risk set at

$$t_{(j)}^*, \omega_{ik}^{[1, \{2,3\}]} = \omega_{ik}^{[1,2]} + \omega_{ik}^{[1,3]}, \omega_{ik}^{[1, \{2,3,4,5\}]} = \omega_{ik}^{[1,2]} + \omega_{ik}^{[1,3]} + \omega_{ik}^{[1,4]} + \omega_{ik}^{[1,5]}, \text{ and}$$

$\omega_{ik}^{[1, \{all\}]} = \omega_{ik}^{[1,1]} + \omega_{ik}^{[1,2]} + \dots + \omega_{ik}^{[1,5]}$. The remaining parameters are estimated by maximizing the corresponding expected complete data log-likelihoods using a quasi-Newton method.

4.2. Variance estimation

Variance estimates for $\boldsymbol{\Theta}$ can be approximated using the following result from [21]:

$$\frac{-\partial^2 l_{\text{comp}}(\boldsymbol{\Theta}|\mathbf{O})}{\partial \boldsymbol{\Theta}^2} = - \int_{\mathbf{H}} \frac{-\partial^2 l_{\text{comp}}(\boldsymbol{\Theta}|\mathbf{P})}{\partial \boldsymbol{\Theta}^2} P(\mathbf{H}|\mathbf{O}, \boldsymbol{\Theta}) d\mathbf{H} - \text{var} \left\{ \frac{\partial l_{\text{comp}}(\boldsymbol{\Theta}|\mathbf{P})}{\partial \boldsymbol{\Theta}} \right\}, \quad (4)$$

where the integration is with respect to the missing data, $\mathbf{H}$. Let b_f denote the multiple imputations from $P(\mathbf{H}|\mathbf{O}, \boldsymbol{\Theta})$, which can be generated based on the conditional expectations in Appendix B. We can then approximate the first and second terms on the right-hand side of (4) for $F=10000$ imputations by

$$\frac{1}{F} \sum_{f=1}^F \frac{\partial^2 l_{\text{comp}}(\boldsymbol{\Theta}|\mathbf{O}, b_f)}{\partial \boldsymbol{\Theta}^2},$$

and

$$\frac{1}{F} \sum_{f=1}^F \left\{ \frac{\partial l_{\text{comp}}(\boldsymbol{\Theta}|\mathbf{O}, b_f)}{\partial \boldsymbol{\Theta}} \right\}^2 - \left\{ \frac{1}{F} \sum_{f=1}^F \frac{\partial l_{\text{comp}}(\boldsymbol{\Theta}|\mathbf{O}, b_f)}{\partial \boldsymbol{\Theta}} \right\}^2,$$

respectively [22].

5. Results from IBCSG Trial VIII

Using the algorithm described in Section 4, we fit the proposed LCJM to the goserelin arm of IBCSG Trial VIII. There are a total of 292 patients eligible for analysis; 152 have masked events before treatment end ($\phi = 1$) of which 46 had a DFS event, and 140 are unmasked as TIA ($\phi \in \{2, 3\}$) of which 33 had a DFS event. All patients experienced cessation before treatment end, thus, there are no patients in data configurations $\phi = 4$ or 5. To avoid speculations concerning the influence of DFS information on determining the latent classes (i.e. degree of OFS), we establish the LCM before incorporating the DFS data [15]. The LCM models the relationship between the time-to-event surrogates, covariates $\mathbf{W}_i$, and latent class membership through (1) and (2). Thus, we fit the LCM first to determine the classes/degrees of OFS, then fit the LCJM incorporating DFS. We consider the continuous covariate age at study entry (centered at the mean of 45 years), the binary covariate ER status (1 = positive; 0 = negative), and their interaction. Given the strong biological rationale behind the pharmacodynamics of LHRH agonists to support that goserelin is expected to induce TIA in all patients, we fix π_1 to be **1** for all models fit to the IBCSG data.

We employ the Bayesian information criterion (BIC) to determine whether the 2-class or 3-class LCM is better; the BIC imposes a relatively larger penalty on the number of parameters than other criteria, e.g. Akaike's information criterion (AIC). The BIC values for the 2-class and 3-class models, ignoring covariates, are 505.730 and 972.200, respectively, which strongly suggests that the 2-class model is better. Neither ER status nor its interaction with age was found to be significant predictors of latent class membership, which is expected since a patient's underlying degree of OFS should not be related to whether or not her tumors have hormone receptors. Also as expected, age at entry is significantly associated with degree of OFS and is thus included as a predictor in the LCM.

We also investigated the conditional independence assumption, $f_{\mathbf{Y}}(\mathbf{y}_i | c_i, \mathbf{w}_i) = f_{\mathbf{Y}}(\mathbf{y}_i | c_i)$, by adding the variable age at entry to each of the time-to-event surrogates as described in Section 3. The likelihood ratio test (LRT) statistic (p -value) on 3 degrees of freedom comparing this extended model with the original LCM fitted above is 7.09 ($p = 0.069$), which supports our

conditional independence assumption and implies that our latent class variable (OFS) serves as a summary of the time-to-event surrogates (times to TIA, menopause, and recovery).

Given the above results, all LCJMs considered will include age at entry as a predictor of class membership in the latent class part. However, different predictors will be considered for the DFS part of the model (i.e. OFS, age, ER, and age-ER interaction). As our main interest is in the association between OFS and DFS, we keep OFS in the model regardless of its significance and perform variable selection on the remaining covariates. The LRT statistic on two degrees of freedom for the joint model with DFS predictors OFS, age, ER, and age-ER interaction vs OFS and ER only is 1.367 ($p = 0.459$). Thus, the final joint model includes only OFS and ER status in the DFS part. Table I gives the parameter estimates for the final joint model. Note that the estimates of π_2 and the latent class regression parameters for the joint model are very similar to those obtained from the LCM, i.e. not including DFS (results not shown).

Figure 3 shows the estimated survival curves for TIA, menopause, and recovery of menses by class. Note the different time-scales, where times to TIA and menopause are both from study entry whereas time to recover is from treatment end, which is 2 years for the goserelin arm. From Figure 3 and the parameter estimates in the latent class membership model in Table I, we can identify the defining characteristics of patients belonging to class 1 and class 2. Class 1 patients are older and have a shorter time to menopause than class 2 patients. Both classes have similar time to TIA, which is expected given the pharmacodynamics of LHRH agonists. Moreover, almost 80 per cent (i.e. $(1 - \pi_{21}) \times 100$) of patients in class 1 never recover their menses (i.e. they are cured with respect to recovery or, in other words, they entered menopause before recovery) compared to about 16 per cent (i.e. $(1 - \pi_{22}) \times 100$) in class 2. Conditional on being uncured/susceptible to recovery, class 1 patients take a longer time to recover their menses than class 2. These defining characteristics lead us to label the underlying degree of OFS of class 1 patients as 'high' and of class 2 patients as 'low'.

Given these latent class labels, the parameter estimate for latent class in the DFS part of the LCJM implies a non-significant 8 per cent decrease in the risk of DFS for patients with a 'high' degree of OFS compared to those with a 'low' degree of OFS. This result implies that lower degrees of OFS, or higher circulating levels of estrogen, do *not* significantly increase a patient's risk of late breast cancer events after LHRH agonists. Moreover, the parameter estimate for ER suggests that having an ER-positive tumor decreases the patient's hazard of DFS by 25 per cent compared to having an ER-negative tumor, holding class constant ($p = 0.064$). As mentioned in Section 1, since goserelin is an LHRH agonist, we expect better results for ER-positive patients. Similar trends were reported in the IBCSG Trial VIII clinical paper [17]. These results further support that LHRH agonists may be a useful alternative to chemotherapy for premenopausal, ER-positive breast cancer patients, especially those who are concerned about their reproductive health and fertility after treatment.

The first row of Figure 4 shows the estimated DFS curves by class, holding ER status constant, whereas the second row shows the estimated DFS curves by ER status, holding class constant.

Futhermore, we can estimate the proportion of patients in class k by $\hat{p}_k = (1/n) \sum_{i=1}^n p_{ik}$. About 40 per cent of patients are in the 'high' OFS class and 60 per cent are in the 'low' OFS class. Each class also has some unambiguously assigned patients since the maximum value of $\hat{p}_{ik}$ for both classes is very close to one. Similar to [15], the proportional hazards assumption can be checked graphically by simulating the latent class variables from the conditional distribution given the observed variables and the final estimates in Table I. In Figure 5, we plot the logarithm of the estimated integrated hazard for the two classes averaged over 50 simulations. We see that there is no evidence against the proportional hazards assumption between classes since the distance between the two curves appears to be independent of time.

6. Discussion

Joint modeling has become more widely used due to its ability to use data more efficiently, resulting in less biased estimates. In this paper, we proposed a latent class joint model to assess the association between a premenopausal breast cancer patient's underlying degree of OFS and DFS after adjuvant treatment with an LHRH agonist. Our joint model obtains information about latent class membership indirectly through a set of menses-related, time-to-event surrogates and accommodates the statistical challenges of masked event and cured proportions present in those surrogates.

As pointed out by a referee, the computational stability of the estimates in our LCJM is a very important issue. In practice, local convergence of an algorithm is usually sufficient when the reliability of the results is checked by starting the algorithm from several different starting values. In our LCJM analysis, we tried three potential starting values for each parameter; we randomly sampled 30 sets of starting values from the 3^{17} possible combinations of the 17 parameters in our model, and then obtained estimates for each sampled set of starting values. The potential starting values for each parameter chosen included both positive and negative values and had relatively broad numeric ranges as well. The reliability of the estimates were checked for each sampled set of starting values. Out of 30 sets of starting values simulated for the final LCJM, only two sets gave results that were slightly inconsistent computationally, but not in terms of parameter interpretations; zero sets gave inconsistent answers to whether OFS is associated with DFS. In addition, given the complexity of our model, we recommend using starting values based on some estimates from simpler models (e.g. the LCM) and also to utilize any available biological knowledge/rationale.

The proposed method is dependent on the presence of a reasonable number of unmasked failures. In the IBCSG study, a large amount of missing/masked data could make the likelihood maximization difficult. About 50 per cent of the cessations are unmasked in the goserelin arm and the resulting estimates were quite stable. In the case where there are too few unmasked cessations, a simpler model could be fit where TIA and menopause are not distinguished, i.e. times to TIA and menopause could be collapsed into the more general 'time to cessation' outcome. However, this would reduce the amount of information used to determine latent class membership unless other surrogate measurements for degree of OFS are available. For IBCSG Trial VIII, whose primary objective was to study how premenopausal, node-negative breast cancer patients respond to combination therapy (arm IV), no other surrogate measures were recorded, which is a limitation of the data set for this analysis.

Acknowledgments

The authors thank the patients, physicians, nurses, and data managers who participated in the International Breast Cancer Study Group (IBCSG) Trial VIII. They further acknowledge support from the United States National Cancer Institute (CA-75362) and the United States National Institute of Health Cancer Training Grant (Jing J. Zhang). They express their gratitude to Richard Gelber, Robert Gray, and Ann Partridge for their assistance throughout. They also thank the editor and referee for their insightful comments and suggestions.

Appendix A: Formulations of likelihood contribution for each data configuration: $f(\pi_i|\Theta, \varphi_i=r)$, $r = 1, \dots, 5$ (Section 4.1)

$$\begin{aligned} f(P_i|\Theta, \varphi_i=1) &= [f_{T_2}(x_i|c_{ik}=1)(1 - \pi_{1k_i})^{(1-\tau_{1i})} \pi_{1k_i}^{\tau_{1i}} S_{T_1}(x_i|t_{1i} \leq u_i, \tau_{1i}=1, c_{ik}=1)^{\tau_{1i}}]^{\zeta_i} \times [f_{T_1}(x_i|t_{1i} \leq u_i, \tau_{1i}=1, c_{ik}=1) \pi_{1k_i} (1 \\ f(P_i|\Theta, \varphi_i=2) &= f_{T_1}(t_{1i}|t_{1i} \leq u_i, \tau_{1i}=1, c_{ik}) \pi_{1k_i} f_{T_3}(t_{3i}|\tau_{2i}=1, c_{ik}) \pi_{2k_i} S_{T_2}(g_i|c_{ik}=1), \\ f(P_i|\Theta, \varphi_i=3) &= f_{T_1}(t_{1i}|t_{1i} \leq u_i, \tau_{1i}=1, c_{ik}) \pi_{1k_i} f_{T_3}(t_{3i}|\tau_{2i}=1, c_{ik}=1) \pi_{2k_i} f_{T_2}(t_{2i}|c_{ik}=1), \\ f(P_i|\Theta, \varphi_i=4) &= (1 - \pi_{1k_i}) S_{T_2}(g_i|c_{ik}=1), \end{aligned}$$

and

$$f(P_i|\Theta, \varphi_i=5) = (1 - \pi_{1k_i}) f_{T_2}(t_{2i}|c_{ik}=1).$$

Appendix B: Definition and derivation of conditional expectations in the E-step (Section 4.1)

For notational simplicity, we let $S_{T_1}(\cdot|c_i) = S_{T_1}(\cdot|t_{1i} \leq u_i, \tau_{1i}=1, c_{ik}=1)$, $f_{T_1}(\cdot|c_i) = f_{T_1}(\cdot|t_{1i} \leq u_i, \tau_{1i}=1, c_{ik}=1)$, $S_{T_3}(\cdot|c_i) = S_{T_3}(\cdot|t_{2i}=1, c_{ik}=1)$, $f_{T_3}(\cdot|c_i) = f_{T_3}(\cdot|t_{2i}=1, c_{ik}=1)$, and $S_{T^*}(t_i^*|c_i, \mathbf{z}_i) = S_{T^*}(t_i^*|c_{ik}=1, \mathbf{z}_i)$. Recall that $\omega_{ik}^{[l, \varphi_i]}$ denotes the conditional expectations for ϕ_i , $l = 1, \dots, 11$. Expectations containing integrals are evaluated using numerical integration.

$$\begin{aligned} \omega_{ik}^{[1,1]} &= P(C_{ik}=1|x_i, T^* > t_i^*, \mathbf{w}_i, \mathbf{z}_i) = \frac{A_{k_i}}{\sum_{k=1}^K A_{k_i}}, \text{ where } A_{k_i} = P(C_{ik}=1|\mathbf{w}_i) S_{T^*}(t_i^*|c_i, \mathbf{z}_i) [(1 - \pi_{1k_i}) f_{T_2}(x_i|c_i) + \pi_{1k_i} \pi_{2k_i} S_{T_1}(x_i|c_i) \\ \omega_{ik}^{[1,2]} &= P(C_{ik}=1|\tau_{2i}=1, \zeta_i=0, T_1=x_i, T_2>g_i, T_3=t_{3i}, T^* > t_i^*, \mathbf{w}_i, \mathbf{z}_i) = \frac{P(C_{ik}=1|\mathbf{w}_i) S_{T^*}(t_i^*|c_i, \mathbf{z}_i) \pi_{1k_i} f_{T_1}(x_i|c_i) S_{T_2}(g_i|c_i)}{\sum_{k=1}^K P(C_{ik}=1|\mathbf{w}_i) S_{T^*}(t_i^*|c_i, \mathbf{z}_i) \pi_{1k_i} f_{T_1}(x_i|c_i) S_{T_2}(g_i|c_i)} \\ \omega_{ik}^{[1,3]} &= P(C_{ik}=1|\tau_{2i}=1, \zeta_i=0, T_1=x_i, T_2=t_{2i}, T_3=t_{3i}, T^* > t_i^*, \mathbf{w}_i, \mathbf{z}_i) = \frac{P(C_{ik}=1|\mathbf{w}_i) S_{T^*}(t_i^*|c_i, \mathbf{z}_i) \pi_{1k_i} f_{T_1}(x_i|c_i) f_{T_2}(t_{2i}|c_i)}{\sum_{k=1}^K P(C_{ik}=1|\mathbf{w}_i) S_{T^*}(t_i^*|c_i, \mathbf{z}_i) \pi_{1k_i} f_{T_1}(x_i|c_i) f_{T_2}(t_{2i}|c_i)} \\ \omega_{ik}^{[1,4]} &= P(C_{ik}=1|\tau_{1i}=0, T_2>g_i, T^* > t_i^*, \mathbf{w}_i, \mathbf{z}_i) = \frac{P(C_{ik}=1|\mathbf{w}_i) S_{T^*}(t_i^*|c_i, \mathbf{z}_i) (1 - \pi_{1k_i}) S_{T_2}(g_i|c_i)}{\sum_{k=1}^K P(C_{ik}=1|\mathbf{w}_i) S_{T^*}(t_i^*|c_i, \mathbf{z}_i) (1 - \pi_{1k_i}) S_{T_2}(g_i|c_i)}, \\ \omega_{ik}^{[1,5]} &= P(C_{ik}=1|\tau_{1i}=0, T_2=t_{2i}, T^* > t_i^*, \mathbf{w}_i, \mathbf{z}_i) = \frac{P(C_{ik}=1|\mathbf{w}_i) S_{T^*}(t_i^*|c_i, \mathbf{z}_i) (1 - \pi_{1k_i}) f_{T_2}(t_{2i}|c_i)}{\sum_{k=1}^K P(C_{ik}=1|\mathbf{w}_i) S_{T^*}(t_i^*|c_i, \mathbf{z}_i) (1 - \pi_{1k_i}) f_{T_2}(t_{2i}|c_i)}, \\ \omega_{ik}^{[2,1]} &= P(C_{ik}=1|\mathbf{O}, \Theta^{(L)}) E(\zeta_i|c_{ik}=1, \mathbf{O}, \Theta^{(L)}) = P(C_{ik}=1, \zeta_i=1|x_i, T^* > t_i^*, \mathbf{w}_i, \mathbf{z}_i) = \omega_{ik}^{[4,1]} + \omega_{ik}^{[5,1]}, \end{aligned}$$

where

$$\begin{aligned}
\omega_{ik}^{[4,1]} &= P(C_{ik} \\
&= 1 | \mathbf{O}, \boldsymbol{\Theta}^{(L)}) E(\tau_{1i} \zeta_i | c_{ik} = 1, \mathbf{O}, \boldsymbol{\Theta}^{(L)}) \\
&= P(C_{ik} \\
&= 1, \tau_{1i} \\
&= 1, \zeta_i \\
&= 1 | x_i, T^* > t_i^*, \mathbf{w}_i, \mathbf{z}_i) \\
&= \frac{P(C_{ik} = 1 | \mathbf{w}_i) S_{T^*}(t_i^* | c_i, \mathbf{z}_i) \pi_{1k_i} S_{T_1}(x_i | c_i) f_{T_2}(x_i | c_i)}{\sum_{k=1}^K A_{k_i}},
\end{aligned}$$

and

$$\begin{aligned}
\omega_{ik}^{[5,1]} &= P(C_{ik} = 1 | \mathbf{O}, \boldsymbol{\Theta}^{(L)}) E((1 - \tau_{1i}) \zeta_i | c_{ik} = 1, \mathbf{O}, \boldsymbol{\Theta}^{(L)}) = P(C_{ik} = 1, \tau_{1i} = 0, \zeta_i = 1 | x_i, T^* > t_i^*, \mathbf{w}_i, \mathbf{z}_i) = \frac{P(C_{ik} = 1 | \mathbf{w}_i) S_{T^*}(t_i^* | c_i, \mathbf{z}_i)}{\sum_{k=1}^K A_{k_i}}, \\
\omega_{ik}^{[3,1]} &= P(C_{ik} = 1 | \mathbf{O}, \boldsymbol{\Theta}^{(L)}) E((1 - \zeta_i) | c_{ik} = 1, \mathbf{O}, \boldsymbol{\Theta}^{(L)}) = P(C_{ik} = 1, \zeta_i = 0 | x_i, T^* > t_i^*, \mathbf{w}_i, \mathbf{z}_i) = \frac{B_i}{\sum_{k=1}^K A_{k_i}},
\end{aligned}$$

where

$$\begin{aligned}
B_i &= P(C_{ik} = 1 | \mathbf{w}_i) S_{T^*}(t_i^* | c_i, \mathbf{z}_i) \{ \pi_{1k_i} (1 - \pi_{2k_i}) f_{T_1}(x_i | c_i) S_{T_2}(g_i | c_i) + \pi_{1k_i} \pi_{2k_i} f_{T_1}(x_i | c_i) \times \left[\{ F_{T_2}(u_i | c_i) - F_{T_2}(x_i | c_i) \} + \int_{u_i}^{g_i} f_{T_2}(t_i) \right. \\
\omega_{ik}^{[6,1]} &= P(C_{ik} = 1 | \mathbf{O}, \boldsymbol{\Theta}^{(L)}) E(\tau_{2i} (1 - \zeta_i) | c_{ik} = 1, \mathbf{O}, \boldsymbol{\Theta}^{(L)}) = P(C_{ik} = 1, \tau_{2i} = 1, \zeta_i = 0 | x_i, T^* > t_i^*, \mathbf{w}_i, \mathbf{z}_i) = \frac{D_i}{\sum_{k=1}^K A_{k_i}},
\end{aligned}$$

where

$$\begin{aligned}
D_i &= P(C_{ik} = 1 | \mathbf{w}_i) S_{T^*}(t_i^* | c_i, \mathbf{z}_i) \pi_{1k_i} \pi_{2k_i} f_{T_1}(x_i | c_i) \times \left[\{ F_{T_2}(u_i | c_i) - F_{T_2}(x_i | c_i) \} + \int_{u_i}^{g_i} f_{T_2}(t_2 | c_i) S_{T_3}(t_2 - u_i | c_i) dt_2 + S_{T_2}(g_i | c_i) S_{T_3}(g_i | c_i) \right] \\
\omega_{ik}^{[7,1]} &= P(C_{ik} = 1 | \mathbf{O}, \boldsymbol{\Theta}^{(L)}) E((1 - \tau_{2i})(1 - \zeta_i) | c_{ik} = 1, \mathbf{O}, \boldsymbol{\Theta}^{(L)}) = P(C_{ik} = 1, \tau_{2i} = 0, \zeta_i = 0 | x_i, T^* > t_i^*, \mathbf{w}_i, \mathbf{z}_i) = \frac{P(C_{ik} = 1 | \mathbf{w}_i) S_{T_1}(x_i | c_i)}{\sum_{k=1}^K A_{k_i}}, \\
\omega_{ik}^{[8,1]} &= P(C_{ik} = 1 | \mathbf{O}, \boldsymbol{\Theta}^{(L)}) E(\tau_{2i} (1 - \zeta_i) \vartheta_i | c_{ik} = 1, \mathbf{O}, \boldsymbol{\Theta}^{(L)}) = P(C_{ik} = 1, \tau_{2i} = 1, \zeta_i = 0, \vartheta_i = 1 | x_i, T^* > t_i^*, \mathbf{w}_i, \mathbf{z}_i) = \frac{P(C_{ik} = 1 | \mathbf{w}_i) S_{T_1}(x_i | c_i) f_{T_2}(x_i | c_i)}{\sum_{k=1}^K A_{k_i}}, \\
\omega_{ik}^{[9,1]} &= P(C_{ik} = 1 | \mathbf{O}, \boldsymbol{\Theta}^{(L)}) E(\tau_{2i} (1 - \zeta_i) (1 - \xi_i) (1 - \vartheta_i) | c_{ik} = 1, \mathbf{O}, \boldsymbol{\Theta}^{(L)}) = P(c_i, \tau_{2i} = 1, \zeta_i = 0, \xi_i = 0, \vartheta_i = 0 | x_i, T^* > t_i^*, \mathbf{w}_i, \mathbf{z}_i) = \\
\omega_{ik}^{[10,1]} &= P(C_{ik} = 1 | \mathbf{O}, \boldsymbol{\Theta}^{(L)}) E((1 - \zeta_i) \log f_{T_2}(t_2 | c_i) | c_{ik} = 1, \mathbf{O}, \boldsymbol{\Theta}^{(L)}) = P(C_{ik} = 1, \zeta_i = 0 | \mathbf{O}, \boldsymbol{\Theta}^{(L)}) E(\log f_{T_2}(t_2 | c_i) | c_{ik} = 1, \zeta_i = 0)
\end{aligned}$$

where

$$M_{1i} = P(C_{ik} = 1 | \mathbf{w}_i) S_{T^*}(t_i^* | c_i, \mathbf{z}_i) \pi_{1k_i} f_{T_1}(t_{1i} | c_i) f_{T_2}(t_2 | c_i), \quad M_{2i} = M_{1i} \{ (1 - \pi_{2k_i}) + \pi_{2k_i} S_{T_3}(t_2 - u_i | c_i) \},$$

and

$$M_{3i} = M_{1i} \{ (1 - \pi_{2k_i}) + \pi_{2k_i} S_{T_3}(g_i - u_i | c_i) \}.$$

It is straightforward to show that

$$\begin{aligned} & \int_{x_i}^{u_i} \log f_{T_2}(t_2 | c_i) M_{1i} dt_2 \\ & + \int_{u_i}^{g_i} \log f_{T_2}(t_2 | c_i) M_{2i} dt_2 \\ & + \int_{g_i}^{\infty} \log f_{T_2}(t_2 | c_i) M_{3i} dt_2 \\ & = P(C_{ik} \\ & = 1 | \mathbf{w}_i) S_{T^*}(t_i^* | c_i, \mathbf{z}_i) \pi_{1k_i} f_{T_1}(t_{1i} | c_i) \left[\int_{x_i}^{u_i} \log f_{T_2}(t_2 | c_i) f_{T_2}(t_2 | c_i) dt_2 + (1 - \pi_{2k_i}) \int_{u_i}^{g_i} \log f_{T_2}(t_2 | c_i) f_{T_2}(t_2 | c_i) dt_2 + \pi_{2k_i} \int_{u_i}^{g_i} \log f_{T_2}(t_2 | c_i) f_{T_2}(t_2 | c_i) dt_2 \right] \end{aligned}$$

and

$$\begin{aligned} & \int_{x_i}^{u_i} M_{1i} dt_2 + \int_{u_i}^{g_i} M_{2i} dt_2 + \int_{g_i}^{\infty} M_{3i} dt_2 = P(C_{ik} = 1 | \mathbf{w}_i) S_{T^*}(t_i^* | c_i, \mathbf{z}_i) \pi_{1k_i} f_{T_1}(t_{1i} | c_i) \left[\{F_{T_2}(u_i | c_i) - F_{T_2}(x_i | c_i)\} + (1 - \pi_{2k_i}) \{F_{T_2}(g_i | c_i) - F_{T_2}(u_i | c_i)\} \right] \\ & \omega_{ik}^{[11,1]} = P(C_{ik} = 1 | \mathbf{O}, \boldsymbol{\Theta}^{(L)}) E(\tau_{2i}(1 - \zeta_i)(1 - \xi_i)(1 - \vartheta_i) \log \{S_{T_3}(t_{2i} | c_i)\} | c_{ik} = 1, \mathbf{O}, \boldsymbol{\Theta}^{(L)}) = P(C_{ik} = 1, \tau_{2i} = 1, \zeta_i = 0, \xi_i = 0, \vartheta_i = 0) \end{aligned}$$

Appendix C: Detailed formulation of $E[l_c(\boldsymbol{\Theta} | \mathbf{P}, \boldsymbol{\Theta}^{(L)})]$ in the M-step (Section 4.1)

$$\begin{aligned} & E[l_c(\boldsymbol{\Theta} | \mathbf{P}, \boldsymbol{\Theta}^{(L)})] \\ & = E[l_{c1}(\pi_1 | \mathbf{P}, \boldsymbol{\Theta}^{(L)})] \\ & + E[l_{c2}(\pi_2 | \mathbf{P}, \boldsymbol{\Theta}^{(L)})] \\ & + E[l_{c3}(a_1, d_1, \alpha | \mathbf{P}, \boldsymbol{\Theta}^{(L)})] + E[l_{c4}(a_2, d_2, \beta | \mathbf{P}, \boldsymbol{\Theta}^{(L)})] \\ & + E[l_{c5}(a_3, d_3, \gamma | \mathbf{P}, \boldsymbol{\Theta}^{(L)})] \\ & + E[l_{c6}(\eta | \mathbf{P}, \boldsymbol{\Theta}^{(L)})] \\ & + E[l_{c7}(\lambda_0^*(\cdot), \boldsymbol{\kappa}, \nu | \mathbf{P}, \boldsymbol{\Theta}^{(L)})] \end{aligned} \quad , \text{ where}$$

$$\begin{aligned}
E[l_{c1}(\pi_1|\mathbf{P}, \boldsymbol{\Theta}^{(L)})] &= \sum_{i:\varphi_i=1} \left\{ \sum_{k=1}^K (\omega_{ik}^{[3,1]} + \omega_{ik}^{[4,1]}) \log(\pi_{1k}) + \sum_{k=1}^K \omega_{ik}^{[5,1]} \log(1 - \pi_{1k}) \right\} + \sum_{i:\varphi_i \in \{2,3\}} \sum_{k=1}^K \omega_{ik}^{[1,\{2,3\}]} \log(\pi_{1k}) + \sum_{i:\varphi_i \in \{4,5\}} \sum_{k=1}^K \omega_{ik}^{[1,\{2,3\}]} \log(1 - \pi_{1k}), \\
E[l_{c2}(\pi_2|\mathbf{P}, \boldsymbol{\Theta}^{(L)})] &= \sum_{i:\varphi_i=1} \left\{ \sum_{k=1}^K \omega_{ik}^{[6,1]} \log(\pi_{2k}) + \sum_{k=1}^K \omega_{ik}^{[7,1]} \log(1 - \pi_{2k}) \right\} + \sum_{i:\varphi_i \in \{2,3\}} \sum_{k=1}^K \omega_{ik}^{[1,\{2,3\}]} \log(\pi_{2k}), \\
E[l_{c3}(a_1, d_1, \alpha|\mathbf{P}, \boldsymbol{\Theta}^{(L)})] &= \sum_{i:\varphi_i=1} \left\{ \sum_{k=1}^K \omega_{ik}^{[4,1]} \log\{S_{T_1}(x_i|c_i)\} + \sum_{k=1}^K \omega_{ik}^{[3,1]} \log\{f_{T_1}(x_i|c_i)\} \right\} + \sum_{i:\varphi_i \in \{2,3\}} \sum_{k=1}^K \omega_{ik}^{[1,\{2,3\}]} \log\{f_{T_1}(t_{1i}|c_i)\}, \\
E[l_{c4}(a_2, d_2, \beta|\mathbf{P}, \boldsymbol{\Theta}^{(L)})] &= \sum_{i:\varphi_i=1} \left\{ \sum_{k=1}^K \omega_{ik}^{[2,1]} \log\{f_{T_2}(x_i|c_i)\} + \sum_{k=1}^K \omega_{ik}^{[10,1]} \log\{S_{T_2}(g_i|c_i)\} \right\} + \sum_{i:\varphi_i \in \{2,4\}} \sum_{k=1}^K \omega_{ik}^{[1,\{2,4\}]} \log\{S_{T_2}(g_i|c_i)\} + \sum_{i:\varphi_i \in \{3,5\}} \sum_{k=1}^K \omega_{ik}^{[1,\{2,4\}]} \log\{f_{T_2}(t_{2i}|c_i)\}, \\
E[l_{c5}(a_3, d_3, \gamma|\mathbf{P}, \boldsymbol{\Theta}^{(L)})] &= \sum_{i:\varphi_i=1} \left\{ \sum_{k=1}^K \omega_{ik}^{[11,1]} + \sum_{k=1}^K \omega_{ik}^{[8,1]} \log\{S_{T_3}(g_i - u_i|c_i)\} \right\} + \sum_{i:\varphi_i \in \{2,3\}} \sum_{k=1}^K \omega_{ik}^{[1,\{2,3\}]} \log\{f_{T_3}(t_{3i}|c_i)\}, \\
E[l_{c6}(\eta|\mathbf{P}, \boldsymbol{\Theta}^{(L)})] &= \sum_{i=1}^n \left[\sum_{k=1}^K \omega_{ik}^{[1,\{all\}]} \mathbf{w}'_i \eta_k - \log \left\{ \sum_{k=1}^K \exp(\mathbf{w}'_i \eta_k) \right\} \right],
\end{aligned}$$

$$\text{and } E[l_{c7}(\lambda_0^*(.), \boldsymbol{\kappa}, \nu|\mathbf{P}, \boldsymbol{\Theta}^{(L)})]$$

$$= \sum_{j=1}^J \sum_{i=1}^n \delta_i 1(V_i = t_{(j)}^*) \left\{ \log \lambda_0^*(t_{(j)}^*) + \boldsymbol{\kappa}' \mathbf{z}_i + \sum_{k=1}^K \omega_{ik}^{[1,\{all\}]} \nu_k \right\} - \lambda_0^*(t_{(j)}^*) \sum_{i' \in R_j} \exp(\boldsymbol{\kappa}' \mathbf{z}_{i'}) \sum_{k=1}^K \omega_{i'k}^{[1,\{all\}]} \exp(\nu_k),$$

where R_j is the risk set at time $t_{(j)}^*$, $\omega_{ik}^{[1,\{g,h\}]} = \omega_{ik}^{[1,f]} + \omega_{ik}^{[1,g]}$ for $g, h \in \{1, 2, 3, 4, 5\}$, and $\omega_{ik}^{[1,\{all\}]} = \omega_{ik}^{[1,1]} + \omega_{ik}^{[1,2]} + \dots + \omega_{ik}^{[1,5]}$.

References

- Goodwin PJ, Ennis M, Pritchard KI, Trudeau M, Hood N. Risk of menopause during the first year after breast cancer diagnosis. *Journal of Clinical Oncology*. 1999; 17:2365–2370. [PubMed: 10561298]
- Bines J, Oleske DM, Cobleigh MA. Ovarian function in premenopausal women treated with adjuvant chemotherapy for breast cancer. *Journal of Clinical Oncology*. 1996; 14:1718–1729. [PubMed: 8622093]
- Ganz PA, Greendale GA, Petersen L, Kahn B, Bower JE. Breast cancer in young women: reproductive and late health effects of treatment. *Journal of Clinical Oncology*. 2003; 21:4184–4193. [PubMed: 14615446]
- Wilcken N, Stockler M. Ovarian suppression for early breast cancer. *The Lancet*. 2007; 369:1668–1670.
- Szwarc SE, Bonetti M. Modelling menstrual status during and after adjuvant treatment for breast cancer. *Statistics in Medicine*. 2006; 25:3534–3547. [PubMed: 16345025]
- Tsiatis AA, DeGruttola V, Wulfsohn MS. Modeling the relationship of survival to longitudinal data measured with error: application to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*. 1995; 90:27–37.
- Wulfsohn MS, Tsiatis AA. A joint model for survival and longitudinal data measured with error. *Biometrics*. 1997; 53:330–339. [PubMed: 9147598]
- Tsiatis AA, Davidian M. A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*. 2001; 88:447–458.

9. Song X, Davidian M, Tsiatis AA. A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics*. 2002; 58:742–753. [PubMed: 12495128]
10. Law NJ, Taylor JMG, Sandler H. The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure. *Biostatistics*. 2002; 3:547–563. [PubMed: 12933597]
11. Tsiatis AA, Davidian M. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*. 2004; 14:809–834.
12. Yu M, Law NJ, Taylor JMG, Sandler H. Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica*. 2004; 14:835–862.
13. Huang W, Zeger SL, Anthony JC, Garrett ES. Latent variable model for joint analysis of multiple repeated measures and bivariate event times. *Journal of the American Statistical Association*. 2001; 96:906–914.
14. Lin H, Turnbull BW, McCulloch CE, Slate EH. Latent class models for joint analysis of longitudinal biomarker and event process data: application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association*. 2002; 97:53–65.
15. Larsen K. Joint analysis of time-to-event and multiple binary indicators of latent classes. *Biometrics*. 2004; 60:85–92. [PubMed: 15032777]
16. Larsen K. The Cox proportional hazards model with a continuous latent variable measured by multiple binary indicators. *Biometrics*. 2005; 61:1049–1055. [PubMed: 16401278]
17. Proust-Lima C, Joly P, Dartigues JF, Jacqmin-Gadda H. Joint modeling of multivariate longitudinal outcomes and a time-to-event: a nonlinear latent class approach. *Computational Statistics and Data Analysis*. 2009; 53:1142–1154.
18. Castiglione-Gertsch M, O'Neill A, Price KN, Goldhirsch A, Coates AS, Colleoni M, Nasi ML, Bonetti M, Gelber RD. on behalf of the International Breast Cancer Study Group. Adjuvant chemotherapy followed by goserelin versus either modality alone for premenopausal lymph node-negative breast cancer: a randomized trial. *Journal of the National Cancer Institute*. 2003; 95:1833–1846. [PubMed: 14679153]
19. Agresti, A. *Analysis of Categorical Data*. New York: Wiley; 1984.
20. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*. 1977; 39:1–22.
21. Louis TA. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*. 1982; 44:226–233.
22. Tanner, MA. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. 3rd edn. New York: Springer; 1996.

Figure 1.
LCJM schematic.

Figure 2.
Possible observed data configurations for IBCSG Trial VIII.

Figure 3.
Estimated survival curves for TIA, menopause, and recovery of menses.

Figure 4.
Estimated DFS curves by class and ER status.

Figure 5.
Proportional hazards assumption.

Table I

LCJM estimates.

	Class 1	Class 2
Proportion of patients in each class (p_k)	0.393	0.607
Probability of experiencing TIA (π_1)	1 (fixed*)	1 (fixed*)
Probability of experiencing recovery (π_2)	0.217 (0.067)	0.839 (0.033)
<i>Latent class membership (class 1 vs class 2):</i>		
	<u>Estimate (p-value)</u>	
Intercept (η_{11})	-1.554 (0.009)	
Mean-centered age (η_{12})	0.662 (<0.001)	
DFS		
	<u>Estimate (p-value)</u>	<u>Hazard ratio</u>
ER-status (κ_1)	-0.298 (0.064)	0.732
class 1 vs class 2 (v_1)	-0.086 (0.700)	0.918

* As LHRH agonists are expected to induce TIA in all patients, we fix π_1 to be 1.