Biometrics 71, 90–101
March 2015

DOI: 10.1111/biom.12252

A Joint Model for Multistate Disease Processes and Random Informative Observation Times, with Applications to Electronic Medical Records Data

Jane M. Lange,[1] Rebecca A. Hubbard,[1,2,] Lurdes Y. T. Inoue,[1] and Vladimir N. Minin[3,*]**

[1]Department of Bioststatistics, University of Washington, Seattle, Washington, U.S.A.
[2]Biostatistics Unit, Group Health Research Institute, Seattle, Washington, U.S.A.
[3]Departments of Statistics and Biology, University of Washington, Seattle, Washington, U.S.A.
*email: vminin@u.washington.edu
**email: rhubb@u.washington.edu

Summary. Multistate models are used to characterize individuals' natural histories through diseases with discrete states. Observational data resources based on electronic medical records pose new opportunities for studying such diseases. However, these data consist of observations of the process at discrete sampling times, which may either be pre-scheduled and non-informative, or symptom-driven and informative about an individual's underlying disease status. We have developed a novel joint observation and disease transition model for this setting. The disease process is modeled according to a latent continuous-time Markov chain; and the observation process, according to a Markov-modulated Poisson process with observation rates that depend on the individual's underlying disease status. The disease process is observed at a combination of informative and non-informative sampling times, with possible misclassification error. We demonstrate that the model is computationally tractable and devise an expectation-maximization algorithm for parameter estimation. Using simulated data, we show how estimates from our joint observation and disease transition model lead to less biased and more precise estimates of the disease rate parameters. We apply the model to a study of secondary breast cancer events, utilizing mammography and biopsy records from a sample of women with a history of primary breast cancer.

Key words: Disease process; Electronic medical records; Informative observations; Markov-modulated Poisson process; Multistate model; Panel data.

1. Introduction

Multistate modeling is a statistical tool that allows medical researchers to characterize the evolution of disease natural histories through discrete states, including progressive diseases (like HIV (Longini and Clark, 1989)) and episodic diseases with reversible transitions (like asthma (Saint-Pierre et al., 2003)). Many methods exist for modeling disease processes with known transition times and trajectories (Andersen and Keiding, 2002; Meira-Machodo et al., 2009). However, recent interest in mining large databases of electronic medical records (Dean et al., 2009) poses new statistical and computational challenges. In such data, patients' disease statuses are recorded only at clinic visits, and exact transition times are unknown. Our goal is to develop a multistate disease modeling framework that accommodates the complexities of observational data from electronic medical records. Features of this type of data include panel observation of disease trajectories, duration-dependent hazard functions, misclassified disease observations, and random visit times that may depend on the disease trajectory.

There are many options for modeling discretely observed multistate processes when visit times are non-informative. The simplest, most tractable models for panel data are time-homogeneous continuous-time Markov chains (CTMCs) (Kalbfleisch and Lawless, 1985). However, CTMCs are limited by an assumption of constant hazard functions that is frequently unrealistic. More flexible models used for panel data include inhomogeneous CTMCs (Kay, 1986; Hubbard, Inoue, and Fann, 2008; Titman, 2011) that allow hazard functions to vary with respect to time since the process origin. Although these models expand the functionality of CTMCs, for many diseases, hazard functions vary with disease state sojourn duration, not just external time. In these cases, semi-Markov models are appealing, yet estimation for such models proves less tractable in the presence of reversible transitions (Chen and Tien, 2004; Kang and Lagakos, 2007). Recent research has suggested advantages of using latent CTMCs in the discrete observation setting (Titman and Sharples, 2010; Lange and Minin, 2013). These models have the backbone of standard CTMCs, retaining their tractability; but multiple latent states map to each disease state, yielding duration-dependent sojourn time distributions. Moreover, it is easy to extend latent CTMC models into continuous-time hidden Markov models (HMMs) to allow for misclassification error. This is the disease modeling framework we will assume.

Most methods developed for panel observed multistate processes treat visit times as non-informative—an assumption that often does not hold in observational studies. Visits scheduled in advance, even those based on observations at previous time points, are ignorable; but times of patient-initiated, symptom-based visits cannot be ignored in the analysis because these times depend on the underlying disease process

(Gruger, Kay, and Schumacher, 1991). Non-ignorable visit times necessitate joint modeling of the disease process and visit times. However, existing joint models of this sort, capable of analyzing panel data (Chen, Yi, and Cook, 2010; Sweeting, Farewell, and De Angelis, 2010; Chen and Zhou, 2011, 2013), assume pre-designated visits with informative missingness, which is appropriate for clinical trials but not for observational clinical data with random visit times.

In this article, we develop a joint model of a discretely observed multistate disease process and a random observation time process. We treat the random, patient-initiated visit times as a temporal point process, which consists of a time series of binary events that occur in continuous time (Daley and Vere-Jones, 2003). Due to their tractability and flexibility, inhomogeneous Poisson processes are commonly used to model observation time point processes jointly with a longitudinal outcome, including continuous (Sun et al., 2005) and panel-count variables (Li, Zhao, and Sun, 2013). However, in these models the dependence of observation times and the disease process is specified by modeling the disease process conditional on the observation process. In contrast, we flip the conditioning, assuming that the observation process is a doubly stochastic Poisson process with rates that depend on the disease state. Our multistate-disease-driven observation (multistate-DDO) model can be viewed as an extension of the "preferential sampling" approach for spatial data to multistate disease processes (Diggle, Menezes, and Su, 2010).

Our joint modeling framework is as follows. The disease process follows a latent CTMC trajectory. We condition on all scheduled visits and assume that patient-initiated DDO times accrue according to a Markov-modulated Poisson process with rates that depend on the patient's current disease status. The disease process is observed, with possible misclassification error, at informative and non-informative visit times. Our multistate-DDO model is similar to the earthquake timing model of Lu (2012), but our model also allows for observations at non-informative times. We demonstrate that the likelihood of our joint model is computationally tractable. Moreover, we develop an efficient expectation-maximization (EM) algorithm to fit our joint multistate-DDO model to panel data. Via simulations, we demonstrate the importance of accounting for random informative sampling times in preventing bias and increasing precision of estimates of disease process parameters.

To illustrate the multistate-DDO model, we apply it to an observational study of secondary breast cancer events (SBCEs) in women who have had a unilateral primary breast cancer. We use data on screening and diagnostic mammograms subsequent to the primary breast cancer as well as biopsies to characterize transitions between breast cancer states. The disease model has a competing risks framework, with terminal competing events corresponding to ipsilateral SBCE (same side as original cancer), contralateral SBCE (opposite side to original cancer), or death prior to SBCE. Patient visits occur either at scheduled screening examinations or at diagnostic examinations triggered by signs or symptoms of an SBCE, necessitating modeling of informative visit times. While conventional studies of SBCEs view time of diagnosed secondary cancer as the target of inference (Chapman, Fish, and Link, 1999; Geiger et al., 2007; Buist et al., 2010), we focus on latent onset time of a mammographically detectable SBCE prior to diagnosis. Estimates from our model are clinically meaningful, as they provide information about prevalence of undetected SBCEs in the growing population of breast cancer survivors (Siegel et al., 2012) as well as screening accuracy in this population.

2. Modeling Framework

2.1. *Joint Model for Disease Process and Disease Driven Observation Process*

The disease process, denoted $X(t)$ and modeled as a time homogeneous CTMC, has state space $S = \{1, \ldots, s\}$, infinitesimal generator matrix $\mathbf{\Lambda} = \{\lambda_{ij}\}$, and initial distribution $\boldsymbol{\pi}$. Jumps in $X(t)$ correspond to an individual's transitions between states in the disease process. The observation process, denoted $N(t)$, is a Markov-modulated Poisson process with piecewise constant rates $q(t) = q\left(X(t)\right)$ that depend on the underlying disease state. $N(t)$ has state space $\{0, 1, \ldots, \infty\}$, corresponding to the accrual of patient-initiated disease-driven observations (DDOs): the process jumps and the state increases by one each time a DDO occurs. Rates of DDOs corresponding to disease states $\{1, \ldots, s\}$ are denoted $\mathbf{q} = (q_1, \ldots, q_s)$.

Jointly, the disease process and counts of DDOs evolve according to a bivariate time-homogeneous continuous-time Markov chain, $Y(t) = (X(t), N(t))$ (Mark and Ephraim, 2013). The state space for $Y(t)$ is the Cartesian product of the state space of $X(t)$ and $N(t)$,

$$S' = \{(1, 0), (2, 0), \ldots, (s, 0), (1, 1), \ldots (s, 1), \ldots, (1, \infty), \ldots, (s, \infty)\}.$$

Figure 1A shows an example of a joint three-state disease and observation process trajectory. Supposing $\mathbf{Q} = diag(q_1, \ldots, q_s)$, the transition generator matrix for the joint process $Y(t)$ is

$$\mathbf{R} = \begin{bmatrix} \mathbf{\Lambda} - \mathbf{Q} & \mathbf{Q} & \mathbf{0} & \mathbf{0} & \ldots \\ \mathbf{0} & \mathbf{\Lambda} - \mathbf{Q} & \mathbf{Q} & \mathbf{0} & \ldots \\ \mathbf{0} & \mathbf{0} & \mathbf{\Lambda} - \mathbf{Q} & \mathbf{Q} & \ldots \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{bmatrix}.$$

The structure of $\mathbf{R}$ follows from the assumption that DDOs and changes in disease states cannot occur simultaneously. The first $\mathbf{\Lambda} - \mathbf{Q}$ block yields the transition rates between states $(i, 0)$ and $(j, 0)$ and the first $\mathbf{Q}$ block yields the rates between state $(i, 0)$ and $(j, 1)$; the rest of the generator matrix is structured similarly (Fearnhead and Sherlock, 2006).

2.2. *Likelihood for Observed Data*

Our observed data consist of partial observations of the joint disease and DDO process, since we only see an individual's disease status at DDO times or scheduled visit times. The observation times are $t_1, \ldots, t_n$, and DDO times are distinguished from scheduled visit times via indicator functions $\mathbf{h} = (h_1, \ldots, h_n)$. We denote the collection of DDO event times as $\boldsymbol{\tau} = \{t_i : h_i = 1, i = 1, \ldots, n\}$. Disease states at the observation times are $x_1, \ldots x_n$.

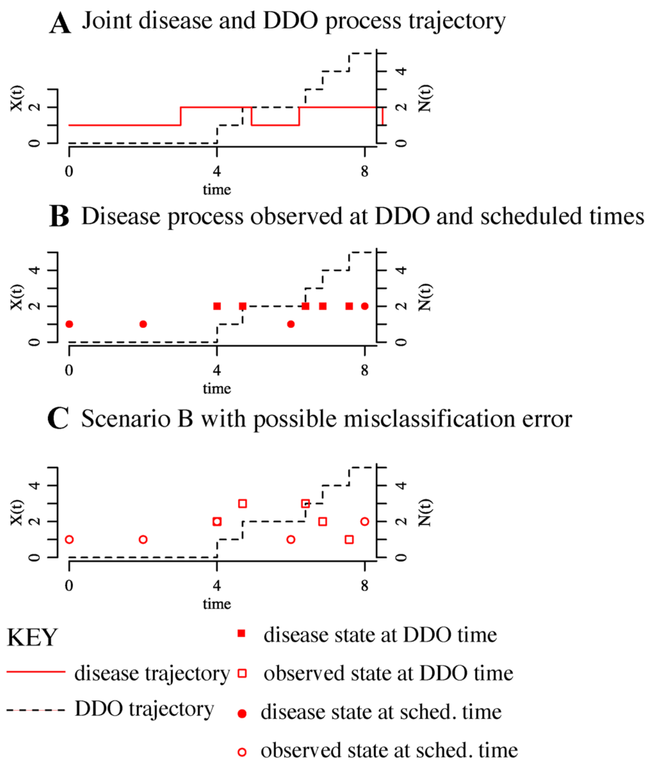

Figure 1. (A) Example of a joint informative observation and disease process, $Y(t) = (X(t), N(t))$. (B) The informative observation time process and the disease process observed at DDO and scheduled times. (C) Same as (B), with misclassification error.

We first consider the likelihood where we observe $X(t)$ at DDO and scheduled visit times without misclassification error (Figure 1B). The likelihood conditions on scheduled visit times. The random variable h_k is a censoring indicator that denotes whether a DDO observation occurred before or after the next scheduled visit time from time t_{k-1}. The Markov property and time-homogeneity of $Y(t)$ enables us to obtain the likelihood of the observed data as a product of density or survival functions for the first passage time of $Y(t)$ into state $(j, k+1)$, given $Y(t_k) = (i, k)$ across each observation interval $[t_{k-1}, t_k]$. Given the time-homogeneity of $Y(t)$ and the structure of $\mathbf{R}$, it suffices to consider $W_{i0,j1}$, the first passage time into state $(j, 1)$, given state $(i, 0)$ at time 0. When t_k is a DDO time, the contribution to the likelihood for the interval $[t_{k-1}, t_k]$ is the density of $W_{i0,j1}$, $f_{ij}(\Delta t_k)$, where $\Delta t_k = t_{k+1} - t_k$. When t_k is a scheduled visit time, we know that $W_{i0,j1} > \Delta t_k$, and the contribution to the likelihood is the survival function for $W_{i0,j1}$, $S_{ij}(\Delta t_k)$. Thus, the likelihood based on the observed data is

$$\begin{aligned}&\mathrm{P}(x_1, \ldots, x_n, \boldsymbol{\tau}, \mathbf{h})\\ &\quad= \nu_{h_1}\pi_{x_1}(h_1)\prod_{k=2}^{n}\{f_{x_{k-1}x_k}(\Delta t_k)\}^{h_{t_k}}\{S_{x_{k-1}x_k}(\Delta t_k)\}^{1-h_{t_k}}.\end{aligned}$$

More generally, the disease process is observed with misclassification error at scheduled visits and DDO times (Figure 1C). Thus, we observe $\mathbf{o} = (o_1, \ldots, o_n)$ rather than $x_1, \ldots, x_n$. We assume that disease process observations are conditionally independent given $X(t)$. The relationship between observed and latent states is described by an emission matrix $\mathbf{E} = \{e(i, j)\}$ with entries $e(i, j) = \mathrm{P}\{o_t = j | X(t) = i\}$. The likelihood includes emission probabilities and sums $\mathrm{P}(x_1, \ldots, x_n, \mathbf{o}, \boldsymbol{\tau}, \mathbf{h})$ over the possible values of $\mathbf{x}$:

$$\begin{aligned}P(\mathbf{o}, \boldsymbol{\tau}, \mathbf{h}) &= \sum_{x_1}\sum_{x_2}\cdots\sum_{x_n}\nu_{h_1}\pi_{x_1}(h_1)\prod_{k=2}^{n}\{f_{x_{k-1}x_k}(\Delta t_k)\}^{h_k}\\ &\quad\times\{S_{x_{k-1}x_k}(\Delta t_k)\}^{1-h_k}\prod_{i=1}^{n} e(x_i, o_i). \qquad (1)\end{aligned}$$

One can derive the density and survival functions $f_{ij}(t)$ and $S_{ij}(t)$ explicitly in terms of $\boldsymbol{\Lambda}$ and $\mathbf{Q}$ using standard CTMC techniques (Freed and Shepp, 1982). First passage time $W_{i0,j1}$ has the same distribution of the absorption time of an auxiliary process $Y'(t)$, corresponding to $Y(t)$ for $\{t : N(t) \in \{0, 1\}\}$, with state space $\{(1, 0), \ldots (s, 0), (1, 1), \ldots (s, 1)\}$, absorbing states $(1, 1) \ldots (s, 1)$, and rate matrix

$$\bar{\mathbf{R}} = \begin{bmatrix}\boldsymbol{\Lambda} - \mathbf{Q} & \mathbf{Q}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}.$$

The survival function for $W_{i0,j1}$ is

$$\begin{aligned}S_{ij}(t) &= P\{W_{i0,j1} > t | Y(0) = (i, 0)\}\\ &= P\left\{Y'(t) = (j, 0) | Y'(0) = (i, 0)\right\}\\ &= \exp\{(\boldsymbol{\Lambda} - \mathbf{Q})t\}_{ij},\end{aligned}$$

and the density function is

$$\begin{aligned}f_{ij}(t) &= \frac{\mathrm{d}}{\mathrm{d}t}P\{W_{i0,j1} < t | Y(0) = (i, 0)\}\\ &= \frac{\mathrm{d}}{\mathrm{d}t}P\left\{Y'(t) = (j, 1) | Y'(0) = (i, 0)\right\}\\ &= \exp\left\{(\boldsymbol{\Lambda} - \mathbf{Q})\, t\right\}_{ij} q_j,\end{aligned}$$

via the Kolmogorov forward equation. Web Appendix A describes modifications to the observed data likelihood (1) for data containing known transition times to absorbing states, such as death. Web Appendix B describes efficient methods for calculating the observed data likelihood (1) based on recursions developed for hidden Markov models and Markov-modulated Poisson processes (Baum et al., 1970).

2.3. *Latent CTMC Model Parameterization*

Disease process models based on standard CTMCs assume that disease state sojourn times are exponentially distributed. To permit more flexibility, we assume a latent CTMC framework for the disease process. We denote the disease process $V(t)$, with state space $G = \{1, 2, \ldots, g\}$. Underlying $V(t)$ is a latent time-homogeneous CTMC $X(t)$, with transition intensity matrix $\boldsymbol{\Lambda}$ and initial distribution $\boldsymbol{\pi}$ and state space $S = \{1_1, 1_2, \ldots, 1_{s_1}\} \cup \{2_1, 2_2, \ldots, 2_{s_2}\} \cup \cdots \cup$

$\{g_1, g_2, \ldots, g_{s_g}\}$. Each observable disease state corresponds to multiple states in the latent state space, such that $V(t) = j <==> X(t) \in \{j_1, j_2, \ldots, j_{s_j}\}$. The mapping of multiple latent states in S to a single disease state in G yields phase-type sojourn distributions of $V(t)$, which can be used to approximate distributions with hazard functions having different shapes (Aalen, 1995). We assume a Coxian structure for $\mathbf{\Lambda}$ for its flexibility and the fact that, up to trivial permutation of states, it is uniquely parametrized when the latent space has a minimal dimension (Cumani, 1982; Titman and Sharples, 2010). Latent CTMC models can be specified in the framework of the observed data likelihood (1) through use of an emission matrix with observed state space G and hidden state space S that equates emission probabilities $e(j_1, k) = e(j_2, k), \ldots, e(j_{s_j}, k)$ for all $j, k \in G$, permitting the mapping of the latent disease space onto the observed disease space.

To incorporate baseline subject-level covariates $\mathbf{w}^{(k)}$ in the disease model, we relate log-rates to a linear predictor, $\log(\lambda_{ij}^{(k)}) = \boldsymbol{\zeta}_{ij}^{\mathrm{T}}\mathbf{w}^{(k)}$, where k denotes the individual. In latent CTMCs, different constraints on covariate effects provide different interpretations. Adding the same covariate parameter to all latent transitions originating from disease state p, that is, $\left\{\lambda_{ij} : i \in \{p_1, \ldots, p_{s_p}\}\right\}$, implies a multiplicative effect on the sojourn time in state p. To represent covariate effects on cause-specific hazard functions, one can add a separate covariate parameter for each transition out of disease state p to disease state r, that is, $\left\{\lambda_{ij} : i \in \{p_1, \ldots, p_{s_p}, j \in \{r_1, \ldots, r_{s_r}\}\right\}$. This specification does not, however, represent a proportional hazards parameterization without additional non-linear constraints (Lindqvist, 2013).

One can also add covariates to DDO, emission, and initial distribution parameterizations. This is achieved by relating log rates of DDOs to a linear predictor; that is, $\log(q_i^{(k)}) = \boldsymbol{\nu}_i^{\mathrm{T}}\boldsymbol{w}^{(k)}$. Initial distributions and emission distributions are multinomial. Assuming S has s total states, the initial distribution $\boldsymbol{\pi}$ has natural parameters $\{\eta_i = \log\{\pi_i/\pi_1)\} : i = 2, \ldots, s\}$, and the emission distribution $\mathbf{e}_i$ has natural parameters $\left\{\eta_{ij} = \log\left\{e(i, j)/e(i, 1)\right\} : j = 2, \ldots, g\right\}$. Subject-level covariates $\mathbf{w}^{(k)}$ are added to the multinomial models via a linear predictor, for example, specifying $\eta_{ij}^{(k)} = \boldsymbol{\gamma_{ij}}^T\mathbf{w}^{(k)}$.

3. Model Selection

We recommend selecting models via the Bayesian information criterion (BIC), given its good performance for selecting general mixture models (Steele and Raftery, 2010) and applicability to comparing non-nested models. The BIC can assist in choosing the dimension of latent space as well assessing parameter constraints in the DDO rates. Finally, hypothesis tests for covariate effects based on likelihood ratio or Wald tests are appropriate, provided parameter identifiability holds under the null model (Sundberg, 1973), which is achievable by constraining covariate effects rather than estimating them separately for each latent disease state.

4. Parameter Estimation

The parameters of interest in the multistate-DDO model, $\boldsymbol{\theta} = (\boldsymbol{\pi}, \mathbf{\Lambda}, \mathbf{E}, \mathbf{q})$, characterize the initial distribution, the disease process, the misclassification probabilities, and the DDO process rates, respectively; we will condition on h_1 rather than estimating its distribution. The standard approach for Markov-modulated Poisson processes and partially observed bivariate CTMCs (Ryden, 1996; Mark and Ephraim, 2013) is to use an EM algorithm to arrive at the maximum likelihood estimates (MLEs) of model parameters (Dempster, Laird, and Rubin, 1977), as this algorithm exploits the ease of maximizing a "complete data" likelihood compared to the observed data likelihood.

In the multistate-DDO model, the complete data are $(\mathbf{x}, \boldsymbol{\tau}, \mathbf{o})$, the full disease trajectory, the DDO trajectory, and observed disease statuses at the discrete times, respectively. The complete data log-likelihood has exponential family form and is a linear function of complete data sufficient statistics. These sufficient statistics include $n_T(i, j)$, the total counts of transitions from state i to state j; $d_T(i)$, the total time spent in state i; z_i, the initial disease state indicator; $u_T(i) = \sum_{l=2}^{n} I(x_l = i)I(h_l = 1)$, the total number of DDOs that have occurred while X(t) was in state i; and $o_T(i, j) = \sum_{l=1}^{n} I(x_l = i)I(o_l = j)$, the total co-occurrences of latent state i and observed state j. As described by Lu (2012), the complete data log-likelihood for an individual is

$$
\begin{aligned}
l(\boldsymbol{\theta}; \mathbf{o}, \boldsymbol{\tau}, \mathbf{x}|h_1) &= l(\boldsymbol{\pi}; x_1|h_1) + l(\mathbf{\Lambda}, \mathbf{q}; \mathbf{x}, \boldsymbol{\tau}|x_1) + l(\mathbf{E}; \mathbf{o}|\mathbf{x}, x_1) \\
&= \sum_{i}^{s} z_i \log\{\pi_i(h_1)\} + \sum_{i=1}^{s} \sum_{j \neq i}^{s} n_T(i, j) \log(\lambda_{ij}) \\
&\quad - \sum_{i=1}^{s} d_T(i) \left(\sum_{j \neq i}^{s} \lambda_{ij} \right) + \sum_{i=1}^{s} u_T(i)(q_i) \\
&\quad - \sum_{i=1}^{s} q_i d_T(i) + \sum_{i=1}^{s} \sum_{j=1}^{r} o_T(i, j) \log[e(i, j)].
\end{aligned} \tag{2}
$$

This likelihood is additive across multiple independent individuals, yielding the complete data likelihood for an entire sample.

The expectation step (E-step) requires computing the expectation of the complete data log-likelihood (2) conditional on observed data $(\mathbf{o}, \boldsymbol{\tau}, \mathbf{h})$. Methods for obtaining these expectations are described in Web Appendix C. The maximization step (M-step) maximizes the conditional expectation of the complete data likelihood, calculated in the E-step, with respect to $\boldsymbol{\theta}$. Covariate-free models admit closed-form M-steps (Lu, 2012). For covariate-parameterized models, we optimize the complete data likelihood via the Newton–Raphson method. Lange and Minin (2013) provide a full description of such a numeric M-step in the context of discretely observed latent CTMCs; the extension to multistate-DDOs is straightforward, as complete-data score and information functions for the $\mathbf{q}$ parameters are identical to those for $\mathbf{\Lambda}$.

We provide an implementation of the EM algorithm in R (R Core Team, 2013), in the form of the R package `cthmm`, available at `http://r-forge.r-project.org/projects/multistate/`. As with all local optimization methods, convergence to the true maximum log-likelihood is not guaranteed, and the method is sensitive to starting values. To make it likely that the true maximum is obtained, we run

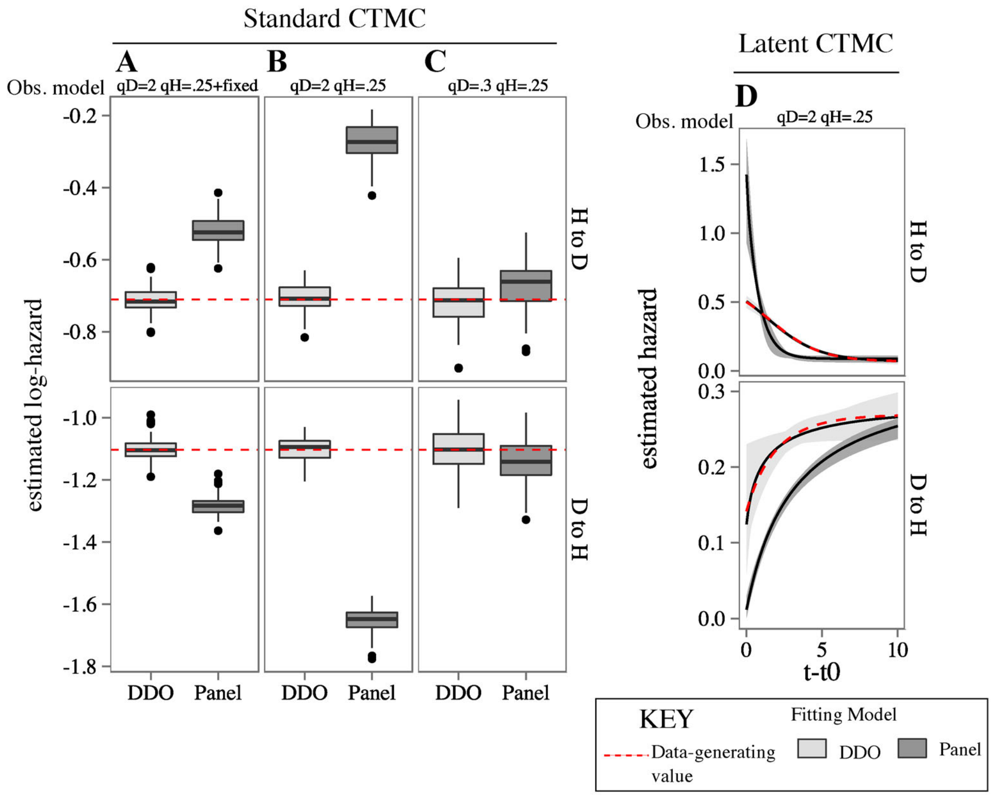

Figure 2. Simulation results demonstrating bias that occurs when informative visit times are ignored. Data were simulated from discretely observed two-state standard and latent CTMC multistate-DDO models on the interval $t = [0, 8]$ at DDO times or a combination of DDO and scheduled visits (see Web Appendix Figure D1 and Table D1 for simulation details). Data were fit with correctly specified multistate-DDO models and incorrectly specified panel models. Box plots/functional box plots are shown for hazard estimates of $H \rightarrow D$ and $D \rightarrow H$ transitions from both DDO and panel models. The different DDO rates in the model states varied across simulations, with more discrepant rates inducing more bias under model misspecification. (A) DDO rates are $q_D = 2, q_H = 0.25$; data also included fixed observation times $t = (0, 2, 4, 6, 8)$. (B) DDO rates are $q_D = 2, q_H = 0.25$. (C) DDO rates are $q_D = 0.35, q_H = 0.25$. (D) DDO rates are $q_H = 0.25$ and $q_D = 2$.

the EM algorithm from multiple sets of initial values, such as random deviates around sensible values based on prior knowledge or MLEs obtained from fitting simpler, for example, covariate-free, models. Finally, we use numerical differentiation, implemented in the R package "NumDeriv" (Gilbert and Varadhan, 2012), to obtain standard errors for parameter estimates from the observed Fisher information matrix.

5. Simulation Study

We used simulated data to characterize the bias incurred by fitting models that condition on the visit times rather than jointly modeling them with the disease trajectory. We considered three disease models: (1) a standard CTMC reversible disease model with two states (*healthy* and *diseased*); (2) a latent CTMC reversible disease model; and (3) a latent CTMC competing risks model similar to the SBCE application, where absorbing states I and C correspond to mammographically detectable ipsilateral and contralateral SBCEs (Web Appendix Figure D1). After simulating disease trajectories from these models, we used the Markov-modulated Poisson process DDO models to generate discretely observed datasets with informative observation times, specifying comparatively higher DDO rates in the diseased states than in the healthy states. The competing risks model allowed for potentially misclassified observations, corresponding to disease surveillance tests with 70% sensitivity and 98% specificity. See Web Appendix Tables D1 and D2 for details.

To investigate bias resulting from ignoring DDO times, we fit data generated from the reversible models with correctly specified multistate-DDO models and with misspecified panel data models that condition on the observations times. The multistate-DDO models yielded unbiased estimates of the disease hazards. Under the misspecified panel models, bias in rate estimates from the reversible standard CTMC followed a consistent pattern: hazard rates for *healthy* → *diseased* transitions and *diseased* → *healthy* transitions were over- and under-estimated, respectively, (Figure 2). Intuitively, informative observation times lead to more observations in the *diseased* state and fewer in the *healthy* state than would be expected under scheduled visits. Bias declined when non-informative times were included with the informative observations (Figure 2A vs. C) and when DDO rates were less discrepant between *healthy* and *diseased* states (Figure 2B vs. C). Ignoring informative times in the latent CTMC reversible models also led to underestimates of *diseased* → *healthy* hazard rates, but *healthy* → *diseased* hazard rates were overestimated only near the state origin time.

In the competing risks disease model similar to the SBCE application, we focused on estimates of the cumulative incidence functions of disease of events I and C. Again, to investigate bias, we either fit correctly specified multistate-DDO models or misspecified panel data models. The correctly specified multistate-DDO model produced unbiased cumulative incidence estimates. The bias resulting from ignoring informative visit times was consistent with results from reversible models: the hazard rates for *healthy* → I/C events were overestimated, yielding left-shifted cumulative incidence curves (Web Appendix Figure D2). Moreover, bias decreased with increasing numbers of scheduled visits added to supplement informative visits. Misspecification of the informative sampling times also dramatically underestimated mammography sensitivity estimates, for example, sensitivity was estimated at 40% when 20% of visits were informative, versus the data-generating sensitivity of 70%. Finally, in addition to investigating bias given model misspecification, we also observed that cumulative incidence estimates based on the properly specified DDO model were shifted left relative to those based on a simulated time of diagnosis, that is, the time of the first true-positive mammogram (Web Appendix Figure D2). This is consistent with diagnosis being a left censoring event for screen-detectable disease.

Via simulation, we also examined the precision of estimates of disease process parameters under informative and non-informative observation schemes. Informative visit times mitigate the uncertainty about the underlying disease states at discrete observations with misclassification error, enabling more precise estimates. We generated data from the reversible standard and latent CTMC disease models (Web Appendix Figure D1) and simulated misclassified observations either in data sampled at DDO times or at pre-designated visit times with equivalent average observation frequencies (Web Appendix Table D1). The simulated data were fit with correctly specified multistate-DDO models or panel models, and we observed less variability in multistate-DDO estimates than their in panel model equivalents (Web Appendix Figure D3).

Covariate effects on disease transition parameters are often a study's scientific target. We used data simulated from the latent CTMC competing risks disease model to consider the sensitivity of estimated covariate effects to correctly specifying the informative sampling time model versus ignoring informative sampling times (see Web Appendix D for setup details and results). Under the correctly specified multistate-DDO model, the MLEs of covariate effects appear valid in terms of bias and confidence interval coverage (Web Appendix Table D3). Interestingly, under misspecified models, estimates of covariate effects retained the same sign and order of magnitude as their data-generating values, and the nominal 95% confidence interval coverage was ~90%. These results are plausible as the covariate effects reflect relative rates of transitions between states across covariate levels rather estimates of the absolute rates. While these simulations are limited in scope, they support the idea that covariate effect estimates may be relatively robust to misspecification of the sampling time model. That said, bias in the intercept terms will still yield biased predictions of state occupancies for different covariate levels.

Finally, all of our simulation studies have assumed that we have correctly specified the number of latent states in the disease models. In general, choosing the number of latent states is an important component of model selection, and we have recommended using the BIC for this purpose. To evaluate the performance of the BIC, we conducted simulation experiments based on data generated from the latent CTMC competing risks multistate-DDO model (see Web Appendix D for details). Upon fitting models that varied in the specification of latent disease and DDO model, we found that the BIC was able to correctly select the data-generating model for 50 out of 50 simulated data sets.

6. Application

We apply the multistate-DDO model to a study of secondary breast cancer events (SBCEs) in women with a history of unilateral breast cancer. The target of inference is onset of mammographically detectable ipsilateral or contralateral SBCE, which are unobserved events that occur prior to diagnosis. The dataset consists of the sequence of mammograms and biopsies following completion of treatment for a primary breast cancer. These data are suited for multistate-DDO models, as mammograms have misclassification error, and observation times include both scheduled screening and patient-initiated visits. Scientifically, we are interested in differences in estimates of cumulative incidence of mammographically detectable versus diagnosed SBCEs, estimates of mammography misclassification, and estimates of covariate effects on disease process parameters.

The study population consists of women diagnosed with unilateral primary breast cancer between 1994 and 2009 who were members of Group Health, an integrated health care system in Washington state, at the time of their primary cancer diagnosis. Women were followed from 180 days after their first cancer until the earliest of the first positive biopsy for a SBCE, death, or disenrollment from the Group Health cohort. Women in this population were recommended to undergo annual screening mammograms in an effort to detect SBCEs before they become symptomatic. Women were also recommended to receive diagnostic evaluations for symptoms that arise in between scheduled surveillance intervals. Mammograms that are positive were followed up with further imaging workup, and, if warranted, biopsies. Mammography visit times were considered to be scheduled screening visits unless the woman and radiologist reported that the visit was for "evaluation of a breast problem," or only the radiologist coded it as such, but the woman endorsed an additional variable indicating symptoms. Web Appendix E provides additional details on outcome variable definitions and exclusion criteria.

6.1. *Data Description*

There are 2936 women in the analysis sample, with a median follow-up time of 5.8 years (IQR 2.8–9.2). Web Appendix Table E1 provides a description of baseline sample characteristics. There were 14,288 contralateral and 10,468 ipsilateral mammograms and 241 contralateral and 212 ipsilateral biopsies. There are fewer ipsilateral than contralateral mammograms because some women were treated for their primary

Table 1
Outcomes for mammograms and biopsies by procedure laterality

		Observed result			
Procedure type	Laterality	Total	Healthy	Ipsi.	Contra.
Mamm.	Contra.	14,288	13,305	0	983
	Ipsi.	10,468	9,800	668	0
Biopsy	Contra.	241	157	0	84
	Ipsi.	212	148	64	0

cancer with mastectomy and thus no longer require disease surveillance on the ipsilateral side. The results of the mammograms and biopsies are shown in Table 1. There were 84 women diagnosed with contralateral SBCEs and 64 diagnosed with contralateral SBCEs. Approximately 7% of all mammograms and 33% of biopsies were positive. Overall, there were 280 days coded as patient-initiated informative visits. On average, women had 0.98 scheduled mammogram visits per person-year. In contrast, rates of informative visits were low: 0.018 per person-year.

6.2. *SBCE Models*

The disease model is a competing risks model with three absorbing states: ipsilateral SBCE, contralateral SBCE, and death before SBCE. We considered both a standard CTMC with state space $\{H = \text{healthy}, I = \text{Ipsilateral SBCE}, C = \text{contralateral SBCE}, D = \text{death before SBCE}\}$ and a latent model with state space $\{H_1, H_2, I, C, D\}$, where H_1, and H_2 are two latent states that map to the healthy disease state. The latent model is biologically plausible as it allows for SBCE hazard rates to be higher near the time of primary breast cancer diagnosis, reflecting recurrences of the primary breast cancer, and to level out over time, reflecting novel cancer events (Demicheli et al., 1996). The transitions in the two models are depicted in Figure 3. All women are assumed to be disease free at the beginning of the study, and start in either the H or H_1 state, depending on the disease model.

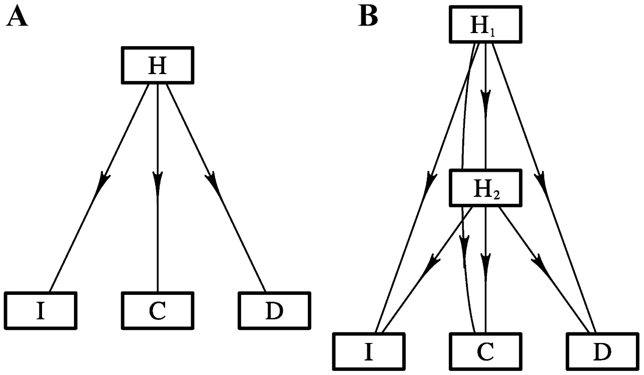

Figure 3. SBCE competing risks disease models. (A) Standard CTMC, where H=healthy, C=contralateral SBCE, I=ipsilateral SBCE, and D=death before SBCE. (B) Latent CTMC with Coxian structure. States H_1 and H_2 map to the healthy state.

Covariates were added to the disease model assuming an additive effect on the log-rates, that is, $\log(\lambda_{ij}) = \boldsymbol{\zeta}_{ij}^{\mathrm{T}}\mathbf{X}$, where $\mathbf{X}$ are the covariates and $\boldsymbol{\zeta}_{ij}$ the coefficients for transition i, j. To ensure parameter identifiability, we constrained parameters in the latent model $\boldsymbol{\zeta}_{H_1,j} = \boldsymbol{\zeta}_{H_2,j}, j \in \{I, C, D\}$ and did not add covariates to the $H_1 \to H_2$ transition. Thus, for each covariate, there is one parameter each affecting transition rates from the healthy state to ipsilateral SBCEs, contralateral SBCEs and death prior to SBCE. The specific covariates we focused on included age at diagnosis, dichotomized to age <50 versus age >50; American Joint Committee on Cancer, Version 6, stage of the primary breast cancer (0 =in situ, 1, 2+); adjuvant endocrine therapy for the original cancer (yes or no); and race (White vs. non-White), based on prior evidence in the literature (de Bock et al., 2006; Andreetta and Smith, 2007; Moran et al., 2008).

The DDO models specify rates of informative sampling times according to the individual's underlying disease state. For model comparison and sensitivity analysis we considered different restrictions on these DDO rates, that is, assuming that the rate was the same in more than one state (for details, see Web Appendix Table E2). All models assumed that the DDO rate in the death state was zero. Models that assume DDO rates are identical across the healthy and ipsilateral and contralateral states suggest that the sampling times are not informative about the disease process: this assumption yields estimates that are quite similar to models that condition on the times, but allows for model comparison via the BIC.

Each mammogram and biopsy was classified as ipsilateral or contralateral. To model mammography misclassification, we assumed a zero probability of detecting an SBCE with a discordant procedure laterality; for example, detecting an ipsilateral SBCE via a mammogram on the contralateral side. In order to promote parameter identifiability in the overall model, we estimated mammography sensitivity and specificity but fixed the biopsy false negative rate at 0.02 and false positive rate at 0, which are reasonable given modern biopsy accuracy rates (Dillon et al., 2005). To accommodate different misclassification probabilities depending on the procedure type and side, we used a time-dependent emission distribution.

6.3. *Model Fitting Results*

The BIC is lowest for the latent CTMC disease model and $H_1/H_2/I, C$ DDO model, where rates of DDO times are allowed to vary in the two healthy states, but are equal in ipsilateral and contralateral SBCE states (see Web Appendix Table E3 for model comparison). The estimated DDO rate in

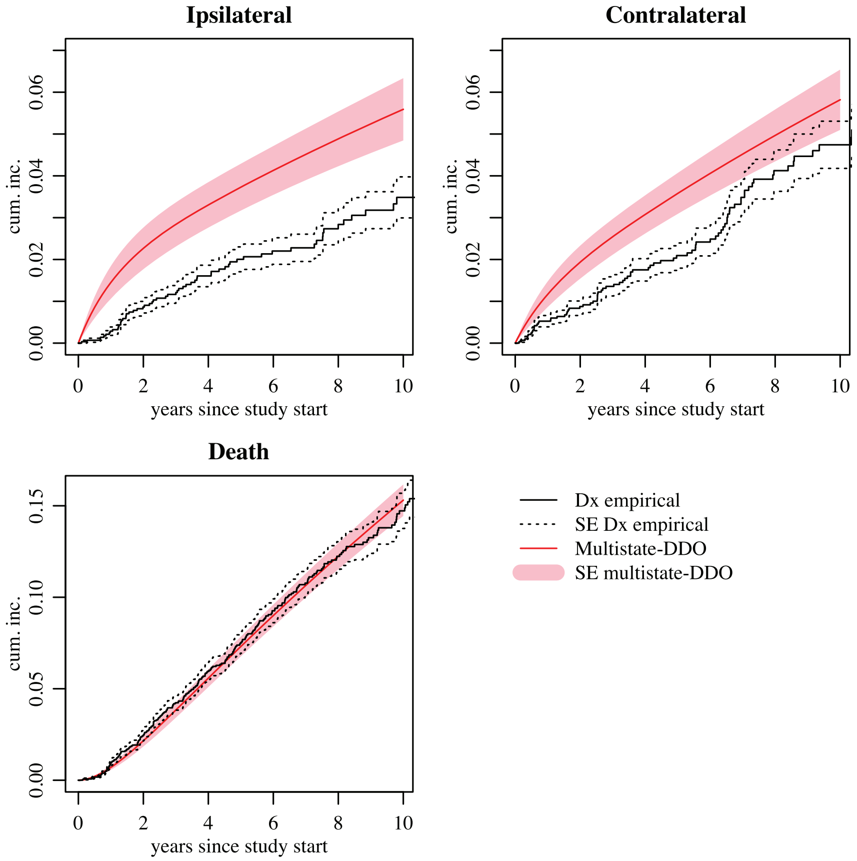

Figure 4. Estimated cumulative incidence for ipsilateral and contralateral SBCEs and death, via empirical estimates of the diagnosis times or using the BIC-selected multistate-DDO model (Web Appendix Table E2, model 6). The bands are pointwise standard errors. Abbreviations: Dx empirical = empirical estimate of cumulative incidence of diagnosed SBCE events; SE = standard error.

state H_1 is 0.046/person-year (95% CI (0.036,0.058)); in H_2 it declines to 0.009/person-year (95% CI (0.007,0.012)); and in the SBCE disease states it is 0.076/person-year (95% CI (0.047,0.11)). These rate estimates are plausible given that patients may be more likely to exhibit symptoms or to initiate visits close to their primary breast cancer diagnosis, as well as after they have developed an SBCE.

Figure 4 plots estimates of cumulative incidence of mammographically detectable SBCEs based on the BIC-preferred multistate-DDO model, in addition to empirical cumulative incidence of diagnosed SBCE events. The multistate-DDO model estimates that at 5 years after diagnosis 3.7% (95% CI [2.6,4.8]) of women will have a mammographically detectable ipsilateral SBCE, whereas 2% (95% CI [1.14,2.6]) will have been diagnosed. Likewise, at 5 years, the multistate-DDO model estimates 3.6% (95% CI [2.6,4.5]) will have a contralateral SBCE, whereas 2.4% (95% CI [1.9, 2.9]) will have been diagnosed. In general, the BIC-preferred DDO model estimates that a range of 25–45% of prevalent SBCEs are undiagnosed from 5 to 10 years after the primary BC, demonstrating the potential benefit of a more sensitive test for improvement of early disease detection.

The multistate-DDO models allow us to estimate true and false positive rates for mammograms. Based on the BIC-selected multistate-DDO model, the estimate of the true positive rate is 69% (95% CI (55%,81%)), and the false positive rate is 5.6% (95% CI (5.3%, 5.9%)). These results are comparable with empirical estimates of mammography sensitivity of 65.4% (95% CI, (61.5%, 69.0%)) and specificity of 98.3% (95%CI (98.2%, 98.4%)) from the Breast Cancer Surveillance Consortium, of which Group Health is a participating institution (Houssami et al., 2011), as well as a recent meta analysis reporting mammography sensitivity ranges of 64-67% and specificity ranges of 85–97% across studies (Robertson et al., 2011).

The multistate-DDO models are parametric, and results are sensitive to model parameterization. Moreover, misspecification of either the observation time, misclassification, or disease model will affect estimates of all components. We examined how results differed if we had assumed a CTMC disease model or a non-informative observation model for the patient-initiated visit times. Unlike the BIC-selected latent disease model, the standard CTMC disease model was unable to capture higher SBCE cumulative incidence in the first 5 years after breast cancer diagnosis (Web Appendix Figure E1). Further, assuming no informative observations yielded left-shifted cumulative incidence estimates relative to models allowing for DDO rates to differ across disease states.

Table 2

Coefficient estimates for a covariate-parameterized version of the BIC-selected SBCE multistate-DDO (M-DDO) model (Web Appendix Table E-2, model 6) and an analogous latent CTMC competing risks disease model based on time of diagnosis (Dx)

		Ipsilateral			Contralateral			Death		
			95% CI			95% CI			95% CI	
	Model	Est.	Low.	Upp.	Est.	Low.	Upp.	Est.	Low.	Upp.
Endocrine	Dx	−0.89	−1.50	−0.28	−0.06	−0.52	0.40	−0.19	−0.45	0.07
therapy	M-DDO	−0.87	−1.47	−0.27	−0.07	−0.52	0.38	−0.21	−0.47	0.05
Age <50	Dx	0.45	−0.09	0.99	−0.36	−0.98	0.26	−0.81	−1.20	−0.42
	M-DDO	0.69	0.18	1.20	−0.28	−0.89	0.33	−0.80	−1.20	−0.40
Stage 1	Dx	−0.60	−1.18	−0.02	0.32	−0.31	0.95	0.50	0.07	0.93
(ref stage 0)	M-DDO	−0.84	−1.40	−0.28	0.33	−0.32	0.98	0.49	0.06	0.92
Stage 2+	Dx	−0.46	−1.18	0.26	0.09	−0.65	0.83	1.17	0.73	1.61
(ref stage 0)	M-DDO	−0.47	−1.15	0.21	0.22	−0.52	0.96	1.17	0.72	1.62
Non-white	Dx	−0.18	−0.92	0.56	−0.14	−0.80	0.52	−0.35	−0.76	0.06
ethnicity	M-DDO	−0.14	−0.87	0.59	−0.13	−0.79	0.53	−0.33	−0.74	0.08

While these results are consistent with the simulation studies examining bias due to ignoring informative sampling times (Web Appendix Figure D2), the magnitude of the shift is much more subtle, probably attributable to the low incidence of DDO times. Estimates of mammography true positive rates are also sensitive to choice of disease and DDO model (Web Appendix Table E4). Indeed, higher sensitivity estimates are associated with lower estimates of the cumulative incidence of SBCEs across the observation period.

6.4. *Covariate Effects*

Point estimates for the covariate parameters within the BIC-selected multistate-DDO model are shown in Table 2. For the purpose of comparison, we also estimated covariate effects for an analogous latent CTMC disease model based on time of diagnosis, the modeled event in conventional studies of SBCEs. Estimates for covariate effects were quite similar between the multistate-DDO and diagnosis-time models, with the exception of effect sizes for age and primary cancer stage on ipsilateral SBCEs. Interestingly, covariate effects were not only similar between diagnosis and multistate-DDO models, they also were relatively robust to misspecification of the informative sampling time model (Web Appendix Figure E2). The models indicated overall significant covariate effects on rates of ipsilateral disease (Wald test ($p<0.001$), but not contralateral SBCEs (Wald p-values ranged from 0.6 to 0.84). Our findings on covariate effects are compatible with an exploratory data analysis we conducted looking at the marginal effects of covariates on cumulative incidence of diagnosed SBCEs (Web Appendix Figure E3), as well as the Breast Cancer Surveillance Consortium's study on diagnosed SBCEs (Buist et al., 2010). Further, although the chosen covariate parameterization does not imply proportional hazards, inspection of estimated hazard ratios revealed they were very near constant over time. Thus exponentiated coefficient estimates are approximately interpretable as having multiplicative effects on hazards. For example, hormone treatment for primary cancer was associated with a reduced hazard of ipsilateral SBCEs, by a factor of $\exp(-0.89) = 0.41$ (95% CI [0.23,0.76]), adjusting for other covariates.

7. Discussion

The increasing availability of electronic medical resources presents new opportunities for modeling multistate diseases. However, as patients' disease statuses are only assessed at discrete clinic visit times – and visit times may be informative about the patients' disease histories – these data pose challenges for inference. The multistate-DDO model provides a novel and flexible approach for modeling such data: it applies to a broad class of disease models, including chronic diseases with reversible transitions and duration-dependent hazard functions; allows for covariate effects; and accommodates both patient-initiated random visit times and scheduled non-informative visits.

Our application of the multistate-DDO model to the study of SBCEs represents a new analysis method in this setting. Existing studies of secondary breast cancers focus on diagnosis as the primary outcome (Chapman et al., 1999; Geiger et al., 2007; Buist et al., 2010), our method uses patient mammography data to model onset of mammographically detectable disease, a clinically relevant outcome that indicates the fraction of a screened population at a given time with undetected disease. Further, others have studied mammography visit patterns in breast cancer survivors (Wirtz et al., 2014), as well as the relationship between screening mammography and mortality (Buist et al., 2013), but our approach is unique in its joint modeling of disease and mammography visit processes.

The multistate-DDO approach for the SBCE data bears similarities to models developed for disease screening trials (Boer, Plevritis, and Clarke, 2004); both model onset of screen-detectable disease and estimate screen sensitivity. However, there are important differences between the two approaches. Disease screening models consider progression to a single disease state that is divided into symptom-free pre-clinical and symptomatic clinical sub-states. In contrast, the multistate-DDO model can handle more complicated disease frameworks, such as the SBCE model's competing risk scenario, but does not distinguish between pre-clinical and clinical sub-states. Indeed, the multistate-DDO model reflects symptom-development implicitly through the informative visit process; DDOs based on symptoms occur more frequently in diseased states but may also occur when the patient is healthy. Ultimately, while estimating pre-clinical sojourn duration is desirable for developing screening protocols, the multistate-DDO model's flexibility invites its use in contexts where screening models do not apply.

The multistate-DDO model also has limitations. For one, the latent structure means parameters are not always identifiable: model building requires compromises between parameterizations that retain estimability but are rich enough to describe the disease process. Furthermore, the model's parametric assumptions make it sensitive to model-misspecification. In particular, misspecification of the disease model impacts both estimates of disease cumulative incidence and mammography sensitivity — an observation also made in reference to disease screening models (Etzioni and Shen, 1997). To probe parametric assumptions of the multistate-DDO model, it will be important to develop goodness of fit evaluation strategies. The informative sampling times mean that the methods aimed at goodness of fit assessment for discretely observed multistate models are no longer applicable (Titman and Sharples, 2008). In our setting, the observed disease states at the disease driven observation times can be construed as a multivariate point process (Gerhard, Haslinger, and Pipa, 2011). Transforming event times in a multivariate point process by the events' cumulative hazard functions yields independent Poisson processes, one for each event category (Meyer, 1971), enabling goodness of fit evaluation via testing the Poisson process assumptions. We plan to adapt this strategy to the multistate-DDO model in the future.

Another concern is the requisite of identifying patient-initiated visits. In absence of such information, we advise against modeling all visit times via the Markov-modulated Poisson process, due to clustering of scheduled visit times. Visit indication is often available from insurance claims as well as clinical records, although different sources may conflict (Fassil et al., 2014; Fenton et al., 2014). In this situation, we recommend performing sensitivity analyses using various visit definitions. There is also the potential for expanding the multistate-DDO model to include visit status as an additional latent component.

We note the potential for other model extensions, including allowing disease transition parameters to have time-dependent covariates. Accommodating piecewise constant covariate effects is straightforward, since one can split individual records on times that the covariate values change. More generally, the models could also include time-dependent covariates that vary in a continuous fashion, but such an approach would require calculating transition probabilities by numerically solving the Kolmogorov forward equations and numeric optimization to obtain MLEs (Titman, 2011). It would also be possible to expand the DDO model to accommodate prior and future visit times as time-dependent covariates, allowing for additional temporal dependence in the DDO process. In general, estimation in a Bayesian framework might also be useful, as it would allow incorporation of prior information about the disease process or misclassification probabilities and might mitigate concerns about parameter identifiability.

8. Supplementary Materials

The R package `cthmm` is found at `http://r-forge.r-project.org/projects/multistate/`. Web Appendices, Tables, and Figures referenced in Sections 2, 4, 5, and 6, are available with this paper at the *Biometrics* website on Wiley Online Library. Sample code and simulated data demonstrating the package are also included.

Acknowledgements

We thank Jim Hughes for helpful discussions. VNM was supported by the NIH grant R01-AI107034. JML, RAH, and LYTI were supported in part by the NIH grant R01-CA160239. Collection of data by the Group Health Breast Cancer Surveillance Registry was supported by NIH grant U01-CA063731.

References

Aalen, O. O. (1995). Phase type distributions in survival analysis. *Scandinavian Journal of Statistics* **22**, 447–463.

Andersen, P. K. and Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research* **11**, 91–115.

Andreetta, C. and Smith, I. (2007). Adjuvant endocrine therapy for early breast cancer. *Cancer Letters* **251**, 17–27.

Baum, L., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics* **41**, 164–171.

Boer, R., Plevritis, S., and Clarke, L. (2004). Diversity of model approaches for breast cancer screening: A review of model assumptions by the Cancer Intervention and Surveillance Network (CISNET) Breast Cancer Groups. *Statistical Methods in Medical Research* **13**, 525–538.

Buist, D. S. M., Abraham, L. A., Barlow, W. E., Krishnaraj, A., Holdridge, R. C., Sickles, E. A., Carney, P. A., Kerlikowske, K., and Geller, B. M. (2010). Diagnosis of second breast cancer events after initial diagnosis of early stage breast cancer. *Breast Cancer Research and Treatment* **124**, 863–873.

Buist, D. S. M., Bosco, J. L. F., Silliman, R. A., Gold, H. T., Field, T., Yood, M. U., Quinn, V. P., Prout, M., and Lash, T. L. (2013). Long-term surveillance mammography and mortality in older women with a history of early stage invasive breast cancer. *Breast Cancer Research and Treatment* **142**, 153–163.

Chapman, J., Fish, E., and Link, M. (1999). Competing risks analyses for recurrence from primary breast cancer. *British Journal of Cancer* **79**, 1508–1513.

Chen, B., Yi, G. Y., and Cook, R. J. (2010). Analysis of interval-censored disease progression data via multi-state models under a nonignorable inspection process. *Statistics in Medicine* **29**, 1175–1189.

Chen, B. and Zhou, X.-H. (2011). Non-homogeneous Markov process models with informative observations with an application to Alzheimer's disease. *Biometrical Journal* **53**, 444–463.

Chen, B. and Zhou, X.-H. (2013). A correlated random effects model for non-homogeneous Markov processes with nonignorable missingness. *Journal of Multivariate Analysis* **117**, 1–13.

Chen, P.-L. and Tien, H.-C. (2004). Semi-Markov models for multistate data analysis with periodic observations. *Communications in Statistics—Theory and Methods* **33**, 475–486.

Cumani, A. (1982). On the canonical representation of homogeneous Markov processes modelling failure-time distributions. *Microelectronics and Reliability* **22**, 583–602.

Daley, D. J. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes*, 2nd edition. New York: Springer.

de Bock, G. H., van der Hage, J. A., Putter, H., Bonnema, J., Bartelink, H., and van de Velde, C. J. (2006). Isolated loco-regional recurrence of breast cancer is more common in young patients and following breast conserving therapy: Long-term results of European Organisation for Research and Treatment of Cancer studies. *European Journal of Cancer* **42**, 351–356.

Dean, B. B., Lam, J., Natoli, J. L., Butler, Q., Aguilar, D., and Nordyke, R. J. (2009). Use of electronic medical records for health outcomes research: A literature review. *Medical Care Research and Review* **66**, 611–638.

Demicheli, R., Abbattista, A., Miceli, R., Valaguss, P., and Bonadonna, G. (1996). Time distribution of the recurrence risk for breast cancer patients undergoing masectomy: Further support about the concept of tumor dormancy. *Breast Cancer Research and Treatment* **41**, 177–185.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* **39**, 1–38.

Diggle, P., Menezes, R., and Su, T.-l. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society, Series C* **59**, 191–232.

Dillon, M. F., Hill, A. D. K., Quinn, C. M., O'Doherty, A., McDermott, E. W., and O'Higgins, N. (2005). The accuracy of ultrasound, stereotactic, and clinical core biopsies in the diagnosis of breast cancer, with an analysis of false-negative cases. *Annals of Surgery* **242**, 701–707.

Etzioni, R. and Shen, Y. (1997). Estimating asymptomatic duration in cancer: The AIDS connection. *Statistics in Medicine* **16**, 627–644.

Fassil, H., Adams, K. F., Weinmann, S., Doria-Rose, V. P., Johnson, E., Williams, A. E., Corley, D. A., and Doubeni, C. A. (2014). Approaches for classifying the indications for colonoscopy using detailed clinical data. *BioMed Central Cancer* **14**, 95.

Fearnhead, P. and Sherlock, C. (2006). An exact Gibbs sampler for the Markov-modulated Poisson process. *Journal of the Royal Statistical Society* **68**, 767–784.

Fenton, J. J., Zhu, W., Balch, S., Smith-Bindman, R., Fishman, P., and Hubbard, R. A. (2014). Distinguishing screening from diagnostic mammograms using medicare claims data. *Medical Care* **52**, 44–51.

Freed, D. S. and Shepp, L. A. (1982). A Poisson process whose rate is a hidden Markov process. *Advances in Applied Probability* **14**, 21–36.

Geiger, A. M., Thwin, S. S., Lash, T. L., Buist, D. S. M., Prout, M. N., Wei, F., Field, T. S., Ulcickas Yood, M., Frost, F. J., Enger, S. M., and Silliman, R. A. (2007). Recurrences and second primary breast cancers in older women with initial early-stage disease. *Cancer* **109**, 966–974.

Gerhard, F., Haslinger, R., and Pipa, G. (2011). Apply the multivarate time-rescaling theorem to neural population models. *Neural Computation* **23**, 1452–1483.

Gilbert, P. and Varadhan, R. (2012). *numDeriv: Accurate Numerical Derivatives.* R package version 2012.9-1.

Gruger, J., Kay, R., and Schumacher, M. (1991). The validity of inferences based on incomplete observations in disease state models. *Biometrics* **47**, 595–605.

Houssami, N., Abraham, L. A., Miglioretti, D. L., Sickles, E. A. Kerlikowske, K., Buist, D. S. M., Geller, B. M., Muss, H. B., and Irwig, L. (2011). Accuracy and outcomes of screening mammography in women with a personal history of early-stage breast cancer. *Journal of the American Medical Association* **305**, 790–799.

Hubbard, R. A., Inoue, L. Y. T., and Fann, J. R. (2008). Modeling nonhomogeneous Markov processes via time transformation. *Biometrics* **64**, 843–850.

Kalbfleisch, J. D. and Lawless, J. F. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association* **80**, 863–871.

Kang, M. and Lagakos, S. W. (2007). Statistical methods for panel data from a semi-Markov process, with application to HPV. *Biostatistics* **8**, 252–264.

Kay, R. (1986). A Markov model for analysing cancer markers and disease states in survival studies. *Biometrics* **42**, 855–865.

Lange, J. M. and Minin, V. N. (2013). Fitting and interpreting continuous-time latent Markov models for panel data. *Statistics in Medicine* **32**, 4581–4595.

Li, N., Zhao, H., and Sun, J. (2013). Semiparametric transformation models for panel count data with correlated observation and follow-up times. *Statistics in Medicine* **32**, 3039–3054.

Lindqvist, B. H. (2013). Phase-type distributions for competing risks. In *Proceedings of the 59th ISI World Statistics Congress*, 25–30. Hong Kong, China.

Longini, I. M. and Clark, S. W. (1989). Statistical analysis of the stages of HIV infection using a Markov model. *Statistics in Medicine* **8**, 831–843.

Lu, S. (2012). Markov modulated Poisson process associated with state-dependent marks and its applications to the deep earthquakes. *Annals of the Institute of Statistical Mathematics* **64**, 87–106.

Mark, B. L. and Ephraim, Y. (2013). An EM algorithm for continuous-time bivariate Markov chains. *Computational Statistics & Data Analysis* **57**, 504–517.

Meira-Machodo, L., de Una-Alvarez, J., Cadarso-Suarez, C., and Andersen, P. K. (2009). Multi-state models for the analysis of time-to-event data. *Statistical Methods in Medical Research* **18**, 195–222.

Meyer, P.-A. (1971). Démonstration simplifiée d'un théorème de Knight. *Séminaire de probabilités de Strasbourg* **5**, 191–195.

Moran, M. S., Yang, Q., Harris, L. N., Jones, B., Tuck, D. P., and Haffty, B. G. (2008). Long-term outcomes and clinicopathologic differences of African-American versus white patients treated with breast conservation therapy for early-stage breast cancer. *Cancer* **113**, 2565–2574.

R Core Team (2013). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Robertson, C., Ragupathy, S. K. A., Boachie, C., Fraser, C., Heys, S. D., Maclennan, G., Mowatt, G., Thomas, R. E., and Gilbert, F. J. (2011). Surveillance mammography for detecting ipsilateral breast tumour recurrence and metachronous contralateral breast cancer: A systematic review. *European Radiology* **21**, 2484–2491.

Ryden, T. (1996). An EM algorithm for estimation in Markov-modulated Poisson processes. *Computational Statistics & Data Analysis* **21**, 431–447.

Saint-Pierre, P., Combescure, C., Daurès, J. P., and Godard, P. (2003). The analysis of asthma control under a Markov assumption with use of covariates. *Statistics in Medicine* **22**, 3755–3770.

Siegel, R., DeSantis, C., Virgo, K., Stein, K., Mariotto, A., Smith, T., Cooper, D., Gansler, T., Lerro, C., Fedewa, S., Lin, C., Leach, C., Cannady, R. S., Cho, H., Scoppa, S., Hachey, M., Kirch, R., Jemal, A., and Ward, E. (2012). Cancer treatment and survivorship statistics, 2012. *CA: A Cancer Journal for Clinicians* **62**, 220–241.

Steele, R. and Raftery, A. (2010). Performance of Bayesian model selection criteria for Gaussian mixture models. In: *Frontiers of Statistical Decision Making and Bayesian Analysis*, Chen, M.-H., Muller, P., Sun, D., Ye, K., Dey, D. (eds), 113–130. New York: Springer.

Sun, J., Park, D.-H., Sun, L., and Zhao, X. (2005). Semiparametric regression analysis of longitudinal data with informative observation times. *Journal of the American Statistical Association* **100**, 882–889.

Sundberg, R. (1973). Maximum likelihood theory for incomplete data from an exponential family. *Scandinavian Journal of Statistics* **1**, 49–58.

Sweeting, M. J., Farewell, V. T., and De Angelis, D. (2010). Multistate Markov models for disease progression in the presence of informative examination times: An application to hepatitis C. *Statistics in Medicine* **29**, 1161–1174.

Titman, A. C. (2011). Flexible nonhomogeneous Markov models for panel observed data. *Biometrics* **67**, 780–787.

Titman, A. and Sharples, L. (2008). A general goodness-of-fit test for Markov and hidden Markov models. *Statistics in Medicine* **27**, 2177–2195.

Titman, A. C. and Sharples, L. D. (2010). Semi-Markov models with phase-type sojourn distributions. *Biometrics* **66**, 742–752.

Wirtz, H. S., Boudreau, D. M., Gralow, J. R., Barlow, W. E., Gray, S., Bowles, E. J. A., and Buist, D. S. M. (2014). Factors associated with long-term adherence to annual surveillance mammography among breast cancer survivors. *Breast Cancer Research and Treatment* **143**, 541–550.

Received January 2014. *Revised July* 2014.
Accepted September 2014.