1-24-2014

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

Jane M. Lange
University of Washington - Seattle Campus, langej@u.washington.edu

Rebecca A. Hubbard
University of Washington - Seattle Campus, rhubb@u.washington.edu

Lurdes Y. T. Inoue
University of Washington - Seattle Campus, linoue@u.washington.edu

Vladimir Minin
University of Washington - Seattle Campus, vminin@uw.edu

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; Titman, 2011; Hubbard et al., 2008) 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 et al., 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 et al., 2010; Chenand and Zhou, 2011; Chen and Zhou, 2013; Sweeting et al., 2010), 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 paper, 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 et al., 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 et al., 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 (MMPP) 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 (BC). 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. In contrast to conventional studies of SBCEs based on diagnosed events (Chapman et al., 1999; Geiger et al., 2007; Buist et al., 2010), we treat the diagnosis time as a left censoring time for onset of mammographically-detectable SBCEs. 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 $q(t) = q(X(t))$ 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 $x_1, \ldots x_n$.

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

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

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 $\mathbf{x}$:

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

One can derive the density and survival functions $f_{ij}(t)$ and $S_{ij}(t)$ explicitly in terms of $\mathbf{\Lambda}$ and $\mathbf{Q}$ using standard CTMC techniques (Freed and Shepp, 1982). First passage time $W_{i0,j1}$ has the same distribution of

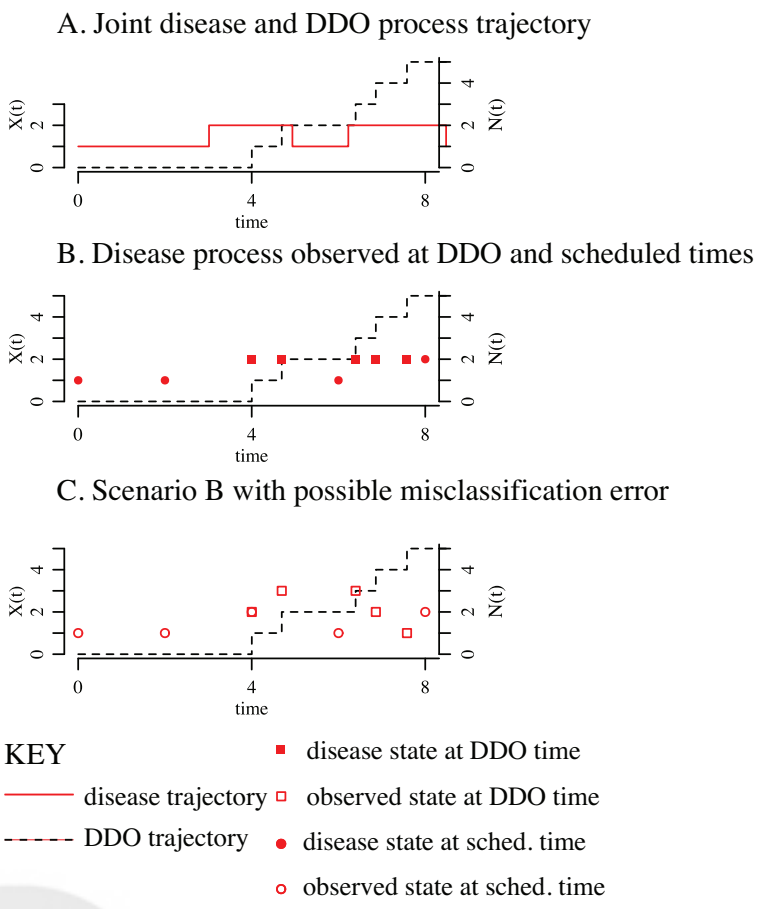

Figure 1: A. Example of a joint informative observation and disease process, $Y(t) = (X(t), N(t))$. B. The informative observation time process and the disease process observed at DDO and scheduled times. C. Same as B, with misclassification error.

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

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

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

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

and the density function is

$$f_{ij}(t) = \frac{d}{dt} P[W_{i0,j1} < t | Y(0) = (i,0)] = \frac{d}{dt} P\left[Y'(t) = (j,1) | Y'(0) = (i,0)\right] = \exp\left[(\mathbf{\Lambda} - \mathbf{Q})t\right]_{ij} q_j,$$

via the Kolmogorov forward equation. Appendix A describes modifications to the observed data likelihood (1) for data containing known transition times to absorbing states, such as death. Appendix B describes efficient methods for calculating the observed data likelihood (1) based on recursions developed for hidden Markov models and MMPPs (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 $\mathbf{\Lambda}$ 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 (Titman and Sharples, 2010; Cumani, 1982). 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, $\log(\lambda_{ij}^k) = \boldsymbol{\zeta}_{ij}^T \mathbf{w}^k$, 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, i.e., $\left\{\lambda_{ij} : i \in \{p_1, \ldots, p_{s_p}\}\right\}$, 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, i.e., $\left\{\lambda_{ij} : i \in \{p_1, \ldots, p_{s_p}\}, j \in \{r_1, \ldots, r_{s_r}\}\right\}$. 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; i.e., $\log(q_i^k) = \boldsymbol{\nu}_i^T \boldsymbol{w}^k$. Initial distributions

and emission distributions are multinomial. Assuming S has s total states, the initial distribution $\boldsymbol{\pi}$ has natural parameters $\{\eta_i = \log(\pi_i/\pi_1)) : i = 2,...,s\}$, and the emission distribution $\mathbf{e}_i$ has natural parameters $\{\eta_{ij} = \log(e(i,j)/e(i,1)) : j = 2,...,g\}$. Subject-level covariates $\mathbf{w}^k$ are added to the multinomial models via a linear predictor, e.g., specifying $\eta_{ij}^k = \boldsymbol{\gamma}_{ij}\mathbf{w}^k$.

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, $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\Lambda}, \mathbf{E}, \mathbf{q})$, 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 MMPPs 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 et al., 1977), as it 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

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

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 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 $\boldsymbol{\Lambda}$.

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 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, e.g., 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 (Appendix Figure D-1). After simulating disease trajectories from these models, we used the MMPP 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 Appendix Tables D-1 and D-2 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* $\rightarrow$ *diseased* transitions and *diseased* $\rightarrow$ *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 2C) and when DDO rates were less discrepant between *healthy* and *diseased* states (Figure 2B vs 2C). Ignoring informative times in the latent CTMC reversible models also led to underestimates of *diseased* $\rightarrow$ *healthy* hazard rates, but *healthy* $\rightarrow$ *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* $\rightarrow I/C$ events were overestimated, yielding left-shifted cumulative incidence curves (Appendix Figure D-2). 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, e.g., 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, i.e., the time of the first true-positive mammogram (Appendix Figure D-2). This is consistent with diagnosis being a left censoring event for screen-detectable disease.

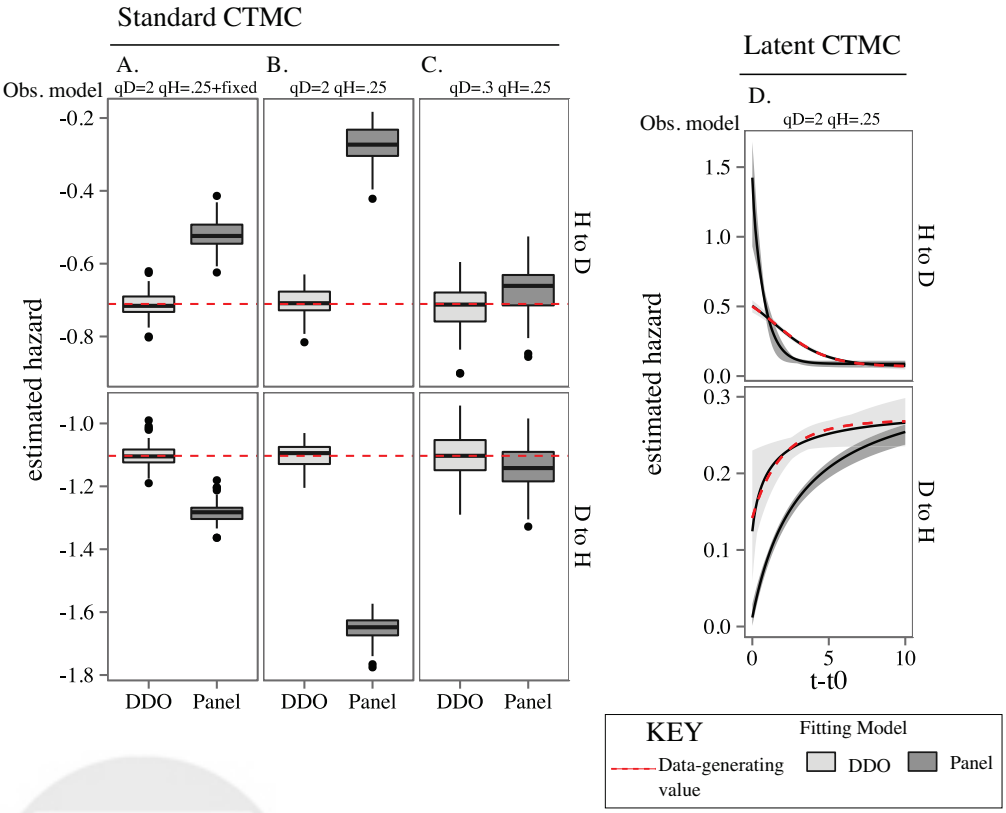

Figure 2: Box plots/functional box plots for hazard estimates of $H \rightarrow D$ and $D \rightarrow H$ transitions using data simulated from discretely observed 2-state standard and latent CTMC multistate-DDO models on the interval t=[0,8]. (See Appendix Table D-1 for simulation details). Data were fit with correctly specified multistate-DDO models and incorrectly specified panel models, demonstrating bias resulting from ignoring informative visits. A. DDO rates are $q_D = 2, q_H = .25$; data also included fixed observation times $t = (0,2,4,6,8)$. B. DDO rates are $q_D = 2, q_H = .25$. C. DDO rates are $q_D = .35, q_H = .25$. D. DDO rates are $q_{H1} = q_{H2} = .25$ and $q_{D1} = q_{D2} = 2$.

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 (Appendix Figure D-1) and simulated misclassified observations either in data sampled at DDO times or at pre-designated visit times with equivalent average observation frequencies (Appendix Table D-1). 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 (Appendix Figure D-3).

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 BC between 1994 and 2009 who were members of Group Health (GH), 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 GH 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. Appendix E provides additional details on outcome variable definitions and exclusion criteria.

6.1 Data description

There are 2,936 women in the analysis sample, with a median follow-up time of 5.8 years (IQR 2.8-9.2). Appendix Table E-1 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 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.

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

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 BC diagnosis, reflecting recurrences of the primary BC, 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.

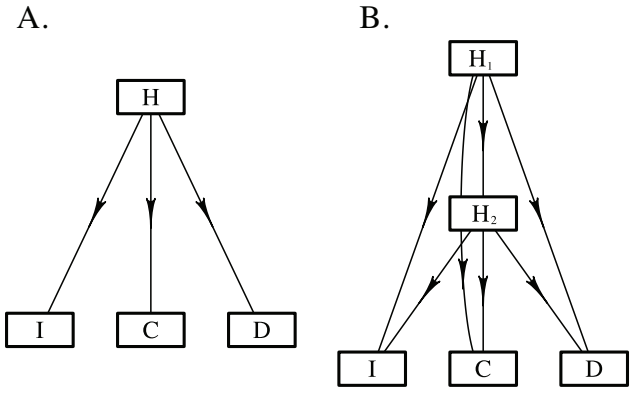

Figure 3: SBCE competing risks disease models. A. Standard CTMC, where H=healthy, C=contralateral SBCE, I=ipsilateral SBCE, and D=death before SBCE. B. Latent CTMC with Coxian structure. States H_1 and H_2 map to the healthy state.

Covariates were added to the disease model assuming an additive effect on the log-rates, i.e., $log(\lambda_{ij}) = \boldsymbol{\zeta}_{ij}^T \mathbf{X}$, 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 $H_1 \rightarrow H_2$ 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 BC (0=in-situ, 1, 2+); adjuvant endocrine therapy for the original cancer (yes or no); and race (White versus 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, i.e. assuming that the rate was the same in more than one state (for details, see Appendix Table E-2). 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; e.g., 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 Appendix Table E-3 for model comparison). The estimated DDO rate in 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 BC 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 five 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 five 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 five to ten 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 (BCSC), of which GH 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. More-

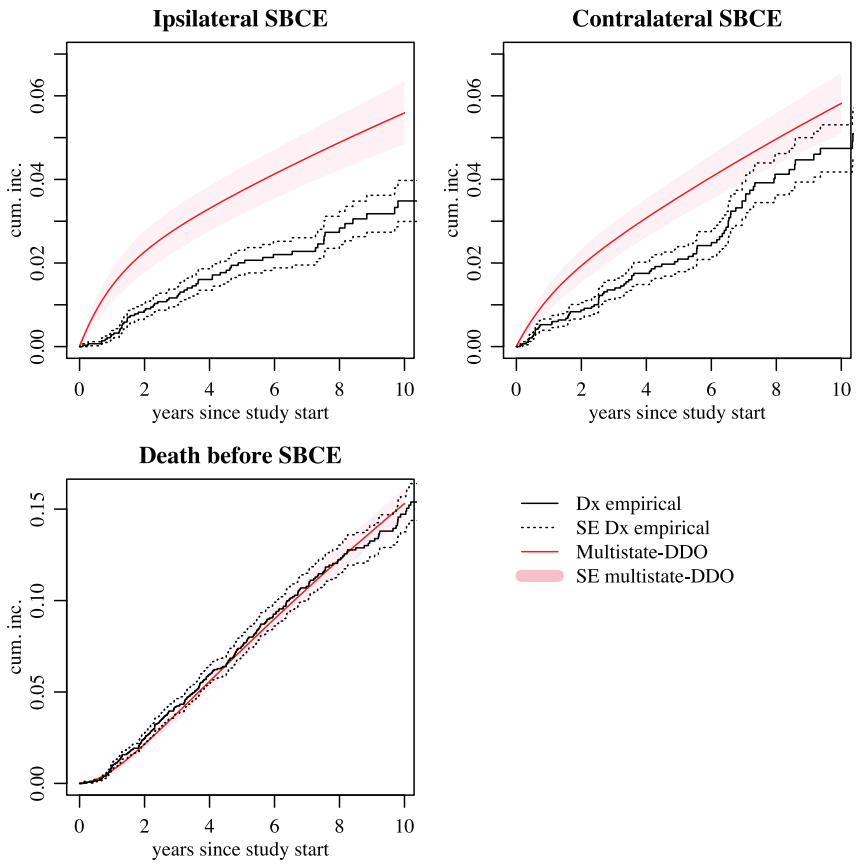

Figure 4: Estimated cumulative incidence for ipsilateral and contralateral SBCEs and death, via empirical estimates of the diagnosis times or using the BIC-selected multistate-DDO model (Appendix Table E-2, model 6). The bands are point-wise standard errors. Abbreviations: Dx empirical=empirical estimate of cumulative incidence of diagnosed SBCE events; SE=standard error.

over, 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 five years after BC diagnosis (Appendix Figure E-1). Further, assuming no informative observations yielded left-shifted cumulative incidence estimates relative to models allowing for DDO rates to differ across disease states. While these results are consistent with the simulation studies examining bias due to ignoring informative sampling times (Appendix Figure D-2), 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 (Appendix Table E-4). 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 (Appendix Figure E-2). 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 0.6-0.84). 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 (Appendix Figure E-3), as well as the BCSC's study on diagnosed SBCE events (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.

Table 2: Coefficient estimates for a covariate-parameterized version of the BIC-selected SBCE multistate-DDO (M-DDO) model (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	0.4	-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	-0.8	-1.20	-0.40
Stage 1	Dx	-0.6	-1.18	-0.02	0.32	-0.31	0.95	0.5	0.07	0.93
(ref stage 0)	M-DDO	-0.84	-1.4	-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	-0.8	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

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.

The model's contribution to methodology for discretely-observed disease process data is to accommodate informative patient-initiated visit times by jointly modeling the random informative observation and disease processes. Via simulations, we showed the need for such an approach to avoid bias. Ignoring the informative sampling led to overestimated rates of transitions into and underestimated rates out of preferentially sampled disease states, as well as biased estimates of misclassification probabilities. We also showed that multistate-DDO models can improve precision of estimates of disease process parameters with misclassified data, as informative visit times disambiguate individuals' disease statuses at sampled times.

Our application of the multistate-DDO model to the study of SBCEs represents a new analysis method in this setting. Existing studies of secondary BCs 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 BC 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 et al., 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).

In our SBCE study, the BIC-selected latent CTMC disease model is likely reasonable. In women with

unilateral primary BC, ipsilateral SBCEs reflect both recurrences and new primary cancers, which is consistent with hazard functions that are relatively high near the primary BC diagnosis and flatten out over time (Demicheli et al., 1996). Contralateral SBCEs reflect only new primary cancers, stochastic events with approximately constant rates over time. Hazard estimates from the selected latent CTMC were consistent with this basic pattern, although they did depict a slight decline in contralateral SBCE rates rather than suggesting they are merely constant. Moreover, the estimated mammography sensitivity from the model agrees quite well with empirical estimates from other studies (Houssami et al., 2011; Robertson et al., 2011), providing additional support that the disease model is not grossly misspecified.

The multistate-DDO model's limitations suggest alternative disease modeling approaches may be desirable in some contexts. For example, in the SBCE model sojourn duration in the healthy state coincides with the external times scale of time since diagnosis. Thus, one could use an inhomogeneous standard CTMC disease model rather than a latent CTMC. In theory, the multistate-DDO model could be modified to accommodate this framework, but additional machinery would be required for likelihood calculations. 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.

The multistate-DDO model also makes assumptions about the observation time process that may not adequately capture patient behavior. In particular, the MMPP model assumes that DDO events in non-overlapping intervals are independent conditional on the latent disease process history. In reality, patient-initiated visit times likely display dependence on upcoming scheduled and prior visit times. One can evaluate the reasonability of the MMPP assumptions using goodness of fit tests based on time-rescaling methods that transform general point processes into a standard homogeneous Poisson process with unit intensity (Brown et al., 2002; Lu, 2012). We also note the possibility of expanding our DDO model to accommodate prior and future visit times as time-dependent covariates, allowing for additional temporal dependence in the DDO process.

Acknowledgments

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.

Brown, E. N., Barbieri, R., Ventura, V., Kass, R. E., and Frank, L. M. (2002). The time-rescaling theorem and its application to neural spike train data analysis. *Neural Computation* **14,** 325–346.

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. (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.

Chenand, 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.

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*. Springer, 2nd edition.

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 (Applied Statistics)* **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.

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.

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.

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*, pages 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.

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*. R Foundation for Statistical Computing, Vienna, Austria. 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 Chen, M.-H., Muller, P., Sun, D., Ye, K., and Dey, D., editors, *Frontiers of Statistical Decision Making and Bayesian Analysis*, pages 113–130. 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). Multi-state 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. 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* pages in press, doi: 10.1007/s10549–013–2816–3.

Appendix for "A joint model for multistate disease processes and random informative observation times, with applications to electronic medical records data"

by Jane M. Lange, Rebecca A. Hubbard, Lurdes Y. T. Inoue, Vladimir N. Minin

Appendix A: Accommodating known times of absorption in observed data likelihood

Known times of death must be accounted for in the observed data likelihood (eq. (1) in main text). Let A be the set of all absorbing states in disease state space S. Assuming that absorption in other states and informative observation events are competing risks, the density of the time of absorption in state $k \in A$, designated by the random variable $W_{i0,k0}$, is given by

$$g_{ik}(t) = \frac{d}{dt}P[W_{i0,k0} < t | Y(0) = (i,0)] = \frac{d}{dt}P[Y'(t) = (k,0) | Y'(0) = (i,0)] = \sum_{j \notin A} S_{ij}(t)\lambda_{jk},$$

where i is a transient state.

When the final time t_n corresponds to absorption of $X(t)$ in state k, we modify the observed data likelihood (eq. (1) in main text) by replacing the terms $f_{x_{n-1}x_n}(\Delta t_n)$ or

$$[f_{x_{n-1}x_n}(\Delta t_n)]^{h_{t_n}}[S_{x_{n-1}x_n}(\Delta t_n)]^{1-h_n}$$

with $g_{x_{n-1}x_n}(\Delta t_n)$.

Appendix B: Forward and backward functions

We use the abbreviation $\mathbf{x}_{1:k}$ for $x_1, \ldots, x_k$, $\mathbf{o}_{1:k}$ for $o_1, \ldots o_k$, $\mathbf{h}_{1:k}$ for $h_1, \ldots, h_k$. The sequence of DDO times up to observation time t_k is denoted $\boldsymbol{\tau}(1,k) = \{t_i : h_i = 1, i = 1, \ldots, k\}$. Forward functions are defined as $\alpha_{t_k}(u) = P[\mathbf{o}_{1:k}, \boldsymbol{\tau}(1,k), \mathbf{h}_{1:k}, X_k = u]$ and backward functions as $\beta_{t_k}(u) = \mathrm{P}[\mathbf{o}_{k+1:n}, \boldsymbol{\tau}(k+1,n), \mathbf{h}_{k+1:n} | X_k = u]$. The forward function is initialized with

$$\alpha_{t_1}(u) = \mathrm{P}(O_1 = o_1, X_1 = u, H_1 = h_1) = e(u, o_1)\nu_{h_1}\pi_{x_1}(h_1),$$

and the recursion for $k = 2, \ldots, n-1$ is

$$\alpha_{t_k}(u) = \sum_i \alpha_{t_{k-1}}(i)e(u, o_k)[f_{iu}(\Delta t_k)]^{h_k}[S_{iu}(\Delta t_k)]^{1-h_k}.$$

The backward function is initialized with $\beta_{t_n}(u) = 1$, and the recursion for $k = 1, \ldots, n-1$ is

$$\beta_{t_k}(u) = \sum_i \beta_{t_{k+1}}(i)e(i, o_{k+1})[f_{ui}(\Delta t_{k+1})]^{h_{k+1}}[S_{ui}(\Delta t_{k+1})]^{1-h_{k+1}}.$$

Observed data likelihood

The observed data likelihood (eq (1) in main text) is $P(\mathbf{o}, \boldsymbol{\tau}, \mathbf{h}) = \sum_u \alpha_{t_n}(u)$, via the forward algorithm; by the backward algorithm, it is
$P(\mathbf{o}, \boldsymbol{\tau}, \mathbf{h}) = \sum_u \beta_{t_1}(u) e(u, o_1) \nu_{h_1} \pi_{x_1}(h_1)$. The forward and backward recursions make the likelihood evaluation practical because, similarly to the standard HMM forward-backward algorithm, the algorithmic complexity of both recursions is $O(ns^2)$.

Hidden state smoothing probabilities

One can generalize the forward and backward functions to an arbitrary time t. That is, we can define $\alpha_t(u) = P[\mathbf{o}_{1:k}, \boldsymbol{\tau}(1,k), \mathbf{h}_{1:k}, X(t) = u]$, for $t \in [t_k, t_{k+1}]$, which is given by

$$\alpha_t(u) = \sum_i \alpha_{t_k}(i) S_{iu}(t - t_k).$$

Similarly, we define $\beta_t(u) = P[\mathbf{o}_{k+1:n}, \boldsymbol{\tau}(k+1,n), \mathbf{h}_{k+1:n} | X(t) = u]$, for $t \in [t_{k-1}, t_k]$, which is given by

$$\beta_t(u) = \sum_i \beta_{t_k}(i) S_{ui}(t_k - t).$$

The general versions of the forward and backward functions also allow us to calculate the smoothing probability $\mathrm{P}[X(t) = i | \boldsymbol{o}, \boldsymbol{\tau}, \boldsymbol{h}]$ for any $t \in [t_1, t_n]$, which predicts the hidden disease state at an arbitrary time conditional on the observed data. This probability is given by

$$\mathrm{P}[X(t) = i | \boldsymbol{o}, \boldsymbol{\tau}, \boldsymbol{h}] = \frac{\beta_t(i)\alpha_t(i)}{\mathrm{P}(\mathbf{o}, \boldsymbol{\tau}, \mathbf{h})}. \tag{B-1}$$

Appendix C: Expectation step

To compute the expectation step (E-step) for the EM algorithm, we note that an individual's log-likelihood contribution (eq. (2) in main text) is additive across time intervals $T_l = [t_l, t_{l+1}]$. Thus,

$$\begin{aligned}
\mathrm{E}[l(\boldsymbol{\theta}; \mathbf{o}, \boldsymbol{\tau}, \mathbf{x}) | \mathbf{o}, \boldsymbol{\tau}, \mathbf{h}] &= \sum_{i=1}^{s} \mathrm{E}[z_i | \mathbf{o}, \boldsymbol{\tau}, \mathbf{h}] \log(\pi_i) \\
&+ \sum_{l=1}^{n-1} \sum_{i=1}^{s} \sum_{j \neq i} \mathrm{E}[n_{T_l}(i,j) | \mathbf{o}, \boldsymbol{\tau}, \mathbf{h}] \log(\lambda_{ij}) - \sum_{l=1}^{n-1} \sum_{i=1}^{s} \mathrm{E}[d_{T_l}(i) | \mathbf{o}, \boldsymbol{\tau}, \mathbf{h}] \left(\sum_{j \neq i} \lambda_{ij} \right) \\
&+ \sum_{l=2}^{n-1} \sum_{i=1}^{s} E[u_{T_l}(i) | \mathbf{o}, \boldsymbol{\tau}, \mathbf{h}] \log(q_i) - \sum_{l=1}^{n-1} \sum_{i=1}^{s} E[d_{T_l}(i) | \mathbf{o}, \boldsymbol{\tau}, \mathbf{h}] q_i \\
&+ \sum_{l=1}^{n-1} \sum_{i=1}^{s} \sum_{j=1}^{r} \mathrm{E}[o_{T_l}(i,j) | \mathbf{o}, \boldsymbol{\tau}, \mathbf{h}] \log\left[e(i,j)\right].
\end{aligned}$$

Computing the E-step therefore requires conditional expectations of the complete data sufficient statistics across T_l. Conditional expectations for z_i, $o_{T_l}(i,j)$, and $u_{T_l}(i)$ are computed using the smoothing probabilities $\mathrm{P}(X_l = m|\mathbf{o}, \boldsymbol{\tau}, \boldsymbol{h})$ (B-1).

Hence,

$$\mathrm{E}[z_i|\mathbf{o}, \boldsymbol{\tau}, \mathbf{h}] = \mathrm{P}(X_1 = i|\mathbf{o}, \boldsymbol{\tau}, \mathbf{h}) = \frac{\beta_{t_1}(i)\alpha_{t_1}(i)}{\mathrm{P}(\mathbf{o}, \boldsymbol{\tau}, \mathbf{h})},$$

$$\mathrm{E}[o_T(j,m)|\mathbf{o}, \boldsymbol{\tau}, \mathbf{h}] = \sum_{l=1}^{n} I(o_l = m)\,\mathrm{P}(X_l = j|\mathbf{o}, \boldsymbol{\tau}, \mathbf{h}) = \sum_{l=1}^{n} I(o_l = m)\frac{\beta_{t_l}(j)\alpha_{t_l}(j)}{\mathrm{P}(\mathbf{o}, \boldsymbol{\tau}, \mathbf{h})},$$

and

$$\mathrm{E}[u_T(j)|\mathbf{o}, \boldsymbol{\tau}, \mathbf{h}] = \sum_{l=2}^{n} I(h_l = 1)\,\mathrm{P}(X_l = j|\mathbf{o}, \boldsymbol{\tau}, \mathbf{h}) = \sum_{l=2}^{n} I(h_l = 1)\frac{\beta_{t_l}(j)\alpha_{t_l}(j)}{\mathrm{P}(\mathbf{o}, \boldsymbol{\tau}, \mathbf{h})}.$$

Note that the sum in the last set of identities is over 2 to n, as the first time should not be considered an observed DDO event.

Expectations of CMTC sufficient statistics $C_{T_l} = d_{T_l}(i)$ or $C_{T_l} = n_{T_l}(i,j)$ can be obtained by first conditioning on x_l, x_{l+1}:

$$\mathrm{E}[C_{T_l}|\mathbf{o}, \boldsymbol{\tau}, \mathbf{h}] = \mathrm{E}\left[\mathrm{E}\left(C_{T_l}|\mathbf{o}, \boldsymbol{\tau}, \mathbf{h}, X_l = a, X_{l+1} = b\right)\right] = \mathrm{E}\left[\mathrm{E}\left(C_{T_l}|X_l = a, X_{l+1} = b, H_{l+1} = h_{l+1}\right)|\mathbf{o}, \boldsymbol{\tau}, \mathbf{h}\right]. \tag{C-1}$$

This follows due to conditional independence of $X(t)$ on $[t_l, t_{l+1}]$ given knowledge of the joint disease and DDO process at the interval endpoints. The task of computing the expectation can be broken down into computing "inner" expectations $\mathrm{E}\left[C_{T_l}|X_l = a, X_{l+1} = b, H_{l+1} = h_{l+1}\right]$ and "outer" expectations. We describe the "inner" and "outer" expectations in turn.

Inner expectations for CTMC sufficient statistics

The formulae for the "inner expectations" are based on conditional expectations for CTMC sufficient statistics with absorbing states (Asmussen et al., 1996). We derive the desired quantities by considering conditional expectations of sufficient statistics $C = n_{ij}(t)$ or $C = d_t(i)$ for a generic homogeneous CTMC X(t) on the interval $[0,t]$, conditional on X(t) at interval endpoints and the informative observation status h_t at time t.

To obtain these expectations, recall that $W_{a0,b1}$ is the first passage time of the bivariate CTMC $Y(t) = (X(t), N(t))$ from state (a,0) to state (b,1). $W_{a0,b1}$ has the same distribution as the time to absorption in state $(b,1)$ of the auxiliary process $Y'(t)$, given $Y'(0) = (a,0)$ and has survival function $S_{ab}(t) = \exp(\boldsymbol{\Lambda} - \mathbf{Q})_{ab}$ and density function $f_{ab}(t) = \exp\left[(\boldsymbol{\Lambda} - \mathbf{Q})\,t\right]_{ab} q_b$ (Section 2.2 in the main text). We will use conditional expectation formulae applicable to $Y'(t)$ to derive the desired quantities.

When the endpoint t is a scheduled visit ($h_t = 0$), we seek the conditional expectation

$$E[C|X(0) = a, X(t) = b, h_t = 0] = \frac{E\left\{C \times I[Y'(t) = (b,0)]|Y'(0) = (a,0)\right\}}{S_{ab}(t)}. \tag{C-2}$$

Our bivariate representation of the process $Y'(t)$ enables us to use standard methods for computing expectations for CTMCs (Hobolth and Jensen, 2011). Thus, for $C = d_t(i)$, the numerator in C-2 is the joint expectation

$$H_i[a,b] = E\left\{d_t(i) \times I[Y'(t) = (b,0)] | Y'(0) = (a,0)\right\} = \int_0^t \exp\left[(\mathbf{\Lambda} - \mathbf{Q})(u)\right]_{ai} \exp([\Lambda - \mathbf{Q})(t-u)]_{ib}\, du,$$

and for $C = n_t(i,j)$, the joint expectation

$$M_{ij}[a,b] = E\left\{n_t(i,j) \times I[Y'(t) = (b,0)] | Y'(0) = (a,0)\right\} = \int_0^t \lambda_{ij} \exp\left[(\mathbf{\Lambda} - \mathbf{Q})u\right]_{ai} \exp\left[(\mathbf{\Lambda} - \mathbf{Q})(t-u)\right]_{jb} du.$$

When t corresponds to a DDO ($h_i = 1$), we seek the conditional expectation

$$\begin{aligned} E[C|X(0) = a, X(t) = b, h_t = 1] &= E[C|W_{a0,b1} = t, Y'(0) = (a,0)] \\ &= \frac{\frac{\partial}{\partial t}\,\mathrm{E}[C, I(W_{a0,b1} < t)|Y'(0) = (a,0)]}{f_{ab}(t)}. \end{aligned} \tag{C-3}$$

To calculate the numerator, we employ expectation formulae derived for CTMCs with absorbing states (Asmussen et al., 1996). For $C = d_t(i)$, the numerator in (C-3) is given by the differentiated joint expectation

$$\frac{\partial}{\partial t}\,\mathrm{E}[d_t(i), I(W_{a0,b1} < t)|Y'(0) = (i,0)] = H_i[a,b]q_b,$$

and for $C = n_t(i,j)$, by

$$\frac{\partial}{\partial t}\,\mathrm{E}[n_t(i,j), I(W_{a0,b1} < t)|Y'(0) = (a,0)] = M_{ij}[a,b]q_b,$$

where $H_i[a,b]$ and $M_{ij}[a,b]$ are defined as before.

We also need to consider the special case of computing conditional expectations for $d_t(i)$ and $n_t(i,j)$ when the interval endpoint t corresponds to a known absorption time in the disease process, such as a time of death. Let A be the set of all absorbing states in S. Treating DDO events as a competing risk, suppose $W_{a0,k0}$ is the time of absorption of $Y'(t)$ in state $k \in A$, given $Y'(0) = (a,0)$, with density $g_{ak}(t) = \sum_{j \notin A} S_{ij}(t)\lambda_{jk}$. In this case, we need the conditional expectation

$$E[C|W_{a0,k0} = t, Y'(0) = (a,0)] = \frac{\frac{\partial}{\partial t}\,\mathrm{E}[C, I(W_{a0,k0} < t)|Y'(0) = (a,0)]}{g_{ak}(t)}. \tag{C-4}$$

When the complete-data statistic of interest is $C = d_t(i)$, the numerator in C-4 is the differentiated joint expectation

$$\frac{\partial}{\partial y}\,\mathrm{E}[d_t(i) I(W_{a0,k1} < t)|Y'(0) = (a,0)] = I(i \notin A) \sum_{c \notin A} H_i(t)[a,c]\lambda_{ck}.$$

For $C = n_t(i,j)$, the numerator in C-4 is the differentiated joint expectation

$$\frac{\partial}{\partial y}\,\mathrm{E}[n_t(i,j) I(W_{a0,k1} < t)|Y'(0) = (a,0)] = I(i,j \notin A) \sum_{c \notin A} M_{ij}(t)[a,c]\lambda_{ck} + I(i \notin A, j = k) S_{ai}(t)\lambda_{ik}.$$

One can use eigenvalue decomposition or the uniformization approach to computing the integrals in each of the joint expectation formulae (Hobolth and Jensen, 2011). Our implementation uses the efficient matrix-based methods from (Minin and Suchard, 2008).

Outer expectations for CTMC sufficient statistics

After computing the "inner expectations," using the described formulae, one can compute "outer" expectations (C-1) for sufficient statistics $C_{T_l} = d_{T_l}(i)$ or $C_{T_l} = n_{T_l}(i,j)$ on the interval T_l using Baum-Welch's bivariate smoothing probabilities

$$\mathrm{P}(X_l = a, X_{l+1} = b|\mathbf{o}, \boldsymbol{\tau}, \mathbf{h}) = \frac{e(b, o_{l+1})\alpha_{t_l}(a)\beta_{t_{l+1}}(b)[f_{ab}(\Delta t_{l+1})]^{h_{l+1}}[S_{ab}(\Delta t_l)]^{1-h_{l+1}}}{\mathrm{P}(\mathbf{o}, \boldsymbol{\tau}, \mathbf{h})}.$$

Thus, the expression for the conditional expectation of the complete data sufficient statistic C_T across the entire time interval $T = [t_1, t_n]$ is

$$\mathrm{E}[C_T|\mathbf{o}, \boldsymbol{\tau}, \mathbf{h}] = \sum_{l=1}^{n-1}\sum_{a=1}^{s}\sum_{b=1}^{s}\mathrm{E}[C_{T_l}|X_l = a, X_{l+1} = b, H_{l+1} = h_{l+1}]\,\mathrm{P}(X_l = a, X_{l+1} = b|\mathbf{o}, \boldsymbol{\tau}, \mathbf{h}).$$

Appendix D: Simulation study

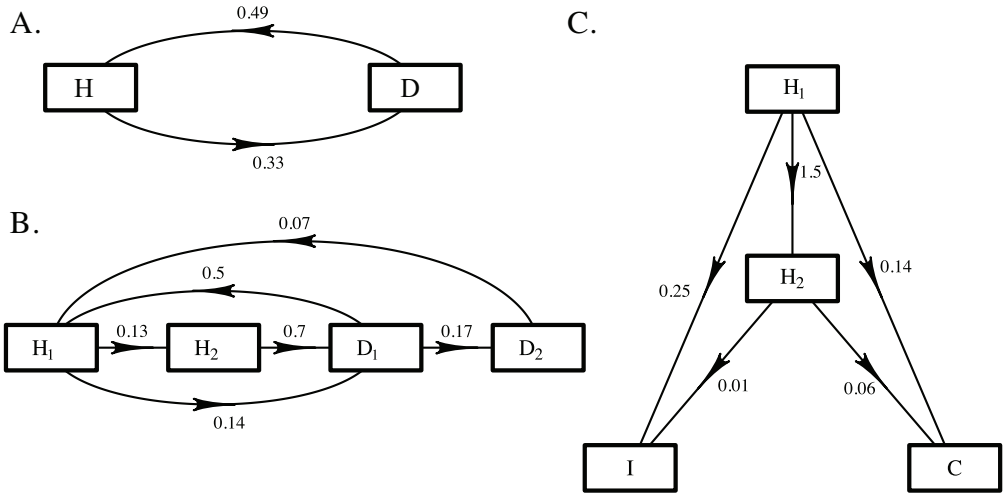

Figure D-1: Data-generating disease models for simulation study. A. 2-state standard CTMC disease model. B. 2-state latent CTMC disease model, where latent states (H_1, H_2) and (D_1, D_2) map to *diseased* and *healthy* states, respectively. C. Competing risks disease model similar to the SBCE model. Latent states (H_1, H_2) map to the *healthy* state; I and C are two absorbing diseased states, corresponding to ipsilateral and contralateral SBCEs.

Table D-1: Data descriptions for discretely-observed datasets simulated from reversible disease models (Figures D-1A and D-1B), including DDO rates, fixed observation times, and misclassification probabilities. These data specifications pertain to experiments summarized in Figure 2 in the main text and in Figure D-3. Each experiment consisted of 100 simulated datasets with 1000 independent individuals.

Figure	Disease model	q_D	q_H	e(H,D)	e(D,H)	Obs. interval	Fixed times	DDOs observed
2A	A	2	.25	0	0	[0,8]	0,2,4,6,8	Y
2B	A	2	.25	0	0	[0,8]	0,8	Y
2C	B	.3	.25	0	0	[0,8]	0,8	Y
2D	B	2	.25	0	0	[0,8]	0,8	Y
D-3A	A	2	.25	.15	.15	[0,7.9]	0,7.9	Y
D-3B	A	0	0	.15	.15	[0,7.9]	0,7.9+10 obs.	N
D-3C	B	2	.25	.15	.15	[0,.8.2]	0,8.2	Y
D-3D	B	0	0	.15	.15	[0,8.2]	0,8.2+8 obs	N

Table D-2: Data descriptions for simulated data from discretely-observed competing risks model (Figure D-1C), including DDO rates, fixed observations, and misclassification probabilities. Notation: $q_{I/C} = q_I = q_C$ and $e(H, I/C) = e(H, I) = e(H, C)$. These data specifications pertain to experiments summarized in Figure D-2. Each experiment consisted of 100 simulated datasets with 1000 independent individuals.

Figure	Disease model	$q_{I/C}$	q_H	e(H,I/C)	e(I/C,H)	Obs. interval	Fixed times	%DDO times
D-2	C	2	.25	.01	.3	[0,8]	0,8	49%
D-2	C	2	.25	.01	.3	[0,8]	0,2,4,6,8	35%
D-2	C	2	.25	.01	.3	[0,8]	0,1,2,...,7,8	20%
D-2	C	2	.25	.01	.3	[0,8]	0,.5,1,...,7.5,8	11%
D-2	C	2	.25	.01	.3	[0,8]	0,.25,.5,...,7.75,8	6%

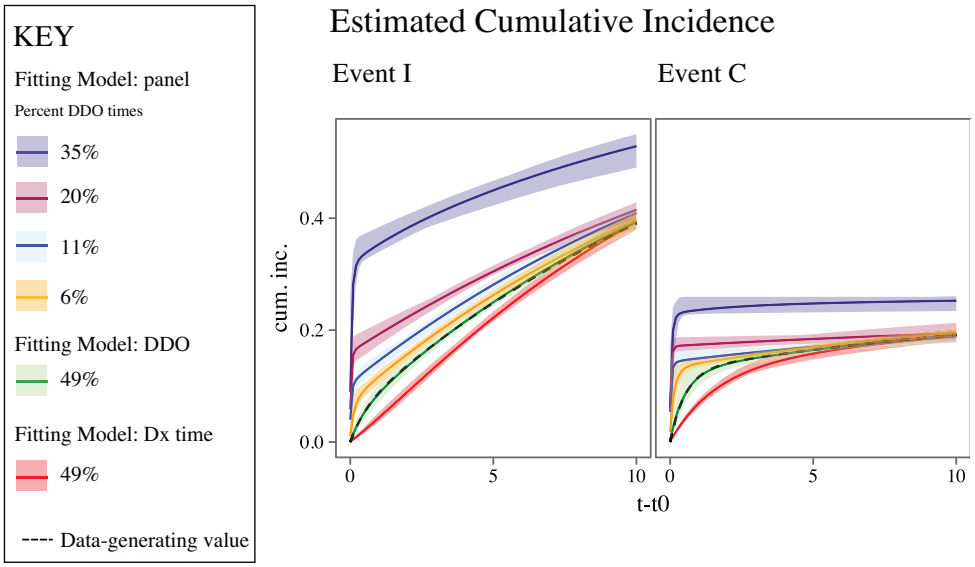

Figure D-2: Functional box plots for simulated data estimates of cumulative incidence for disease events I and C in the latent CTMC competing risks model (Figure D-1C.) Discretely observed data were generated from the disease trajectories according to informative observation times from a DDO model with $q_{H1} = q_{H2} = .25$ and $q_I = q_C = 2$, and varying proportions of supplemental non-informative times. Observations had 70% sensitivity and 98% specificity, corresponding to mammography data. See Table D-2 for further dataset details. Data were fit with panel models or multistate-DDO models, demonstrating bias incurred by ignoring informative observations, and showing how increasing proportions of supplemental scheduled visits mitigates such bias. Also shown is cumulative incidence based on time of diagnosis (Dx time), the time of the first true positive mammogram.

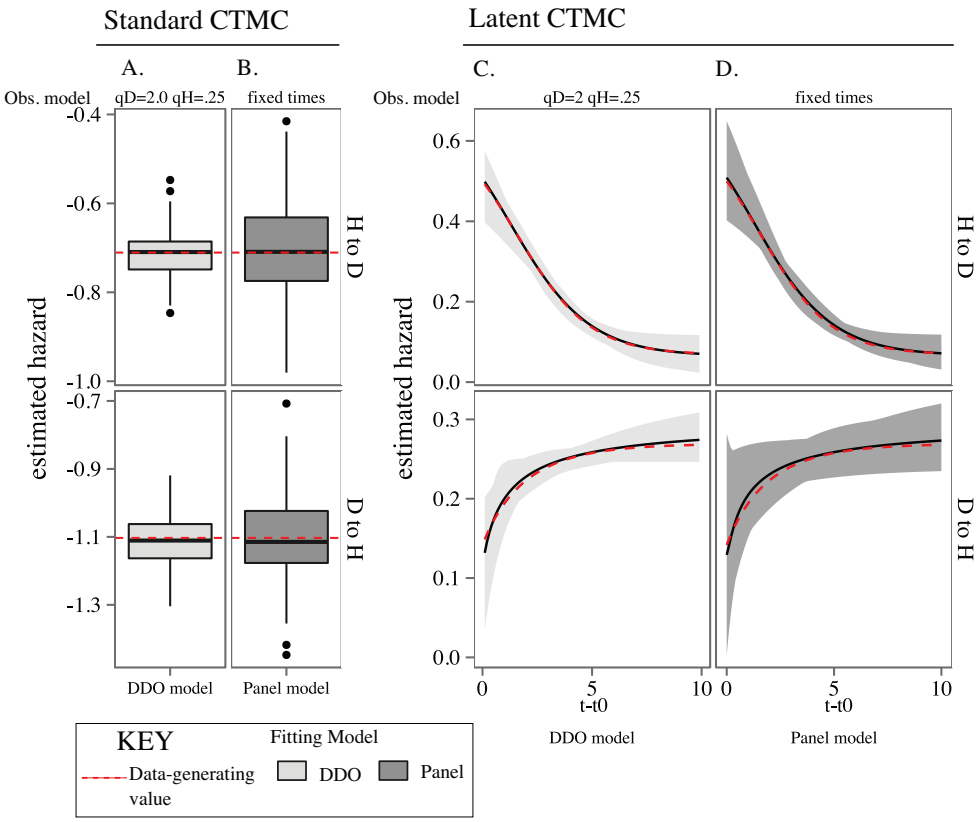

Figure D-3: Box plots/functional box plots for hazard estimates of $H \to D$ and $D \to H$ transitions for standard and latent CTMC reversible disease models (Figure D-1A, D-1B), observed with 15% misclassification error at either DDO times or at fixed times with equal average frequencies. See Table D-1 for further details. Data are fit with correctly specified multistate-DDO or panel models. These result demonstrate the gains in precision in hazard estimates via jointly modeling informative sampling times in the presence of misclassification error.

Appendix E: Second Breast Cancer Event Application

Mammography and biopsy outcomes

Mammograms were positive if the BI-RADS (Breast Imaging-Reporting and Data System) score was 0="more imaging needed," 4="suspicious abnormality," 5="highly suggestive of malignancy," or 6="known malignancy" American College of Radiology (2003). Biopsies with a result of invasive malignancy or ductal carcinoma in situ (DCIS) were considered positive; negative findings included benign growths and benign hyperplasias.

Dataset exclusions

There were 4,133 women with primary unilateral breast cancers diagnosed from 1994-2009 who subsequently received mammography at Group Health. We applied sequential exclusions to obtain an analysis dataset. We excluded women with a mammographically-detectable SBCE within 180 days following the primary BC diagnosis (N=94), since events prior to that time likely reflect progression of the primary disease. We also excluded women if they had a biopsy record not preceded by a mammogram within the preceding 100 days (N=352), as well as those with any missing laterality for mammograms or biopsy procedures (N=424), and those missing any of the covariates of interest (N=327). In total, these exclusions reduced the dataset from 4,133 to 2,936 women, removing 49% percent of ipsilateral cases, 32% of contralateral cases, 37% of those who died prior to an SBCE, and 27% of those who were alive and SBCE-free at the time they were last seen. More ipsilateral cases were dropped since they were more likely to have biopsies not preceded by mammograms within the study period.

Sample characteristics

The 2,936 women in the sample used for analysis, as well as the 1,197 excluded from the sample, are described in Table E-1. The sample was predominantly white (84.7%, N=2,488), with a median age of 61 at primary BC diagnosis (IQR 52, 71). Approximately one fifth of the sample had a stage 0 (DCIS) primary BC (18.6%, N=548), whereas half had stage 1 (49.6%, N=1,456), and the rest, stage 2 or higher. The main difference between included and excluded women is that excluded individuals were more likely to have stage 2 or higher cancer. This is related to our exclusion of individuals with biopsies not preceded by mammograms within the study period being more likely to have advanced stage primary BC.

Table E-1: Characteristics of the GH patients with a history of primary BC, either included in or excluded from the analysis sample. Percentages do not include missing data. Abbrevations: ER+=estrogen receptor positive, PR+=progesterone receptor positive.

	Included (N=2,936)		Excluded (N=1,197)	
	N	%	N	%
Age at diagnosis				
<50	557	19	264	22.1
50-59	801	27.3	330	27.6
60-69	757	25.8	281	23.5
70+	821	28	322	26.9
Missing	0		0	
Race				
White	2488	84.7	1005	86.6
Black	83	2.8	34	2.9
Asian	189	6.4	48	4.1
Other	176	6	73	6.3
Missing	0		37	
Stage of primary cancer				
0	548	18.7	138	14.1
1	1456	49.6	425	43.4
2+	932	31.7	417	42.6
Missing	0		217	
ER+ or PR+ for primary cancer				
No	386	16.3	165	17.5
Yes	1984	83.7	779	82.5
Missing	556		253	
Treatment of primary breast cancer				
Mastectomy				
None	18	0.6	24	2.3
Partial	1925	66.4	711	66.9
Complete unilateral	955	33	328	30.9
Missing	38		134	
Radiation				
No	943	33.3	323	30.9
Yes	1891	66.7	723	69.1
Missing	102		151	26.9
Chemotherapy				
No	2054	70.2	704	63.3
Yes	874	29.8	409	36.7
Missing	8		84	
Adjuvant endocrine therapy				
No	1464	49.9	500	50.8
Yes	1472	50.1	485	49.2
Missing	0		212	

Table E-2: Informative sampling time models for the SBCE data. Non-informative models assume the same DDO rate in all states.

Model label	Disease model	DDO model		
			No. DDO params	Constraints
1	Standard CTMC	non-informative	1	$q_H = q_I = q_C$
2		H/I,C	2	$q_H, q_I = q_C$
3		H/I/C	3	q_H, q_I, q_C
4	Latent CTMC	non-informative	1	$q_{H_1} = q_{H_2} = q_I, q_C$
5		H1,H2/I,C	2	$q_{H_1} = q_{H_2}, q_I = q_C$
6		H1/H2/I,C	3	$q_{H_1}, q_{H_2}, q_I = q_C$
7		H1/H2/I/C	4	$q_{H_1}, q_{H_2}, q_I, q_C$

Table E-3: Model fitting results for SBCE disease and informative sampling time models.

	Disease Model						
	Standard CTMC			Latent CTMC			
	DDO model			DDO model			
	non-inf.	H/I,C	H/I/C	non-inf.	H1,H2/I,C	H1/H2/I,C	H1/H2/I/C
Model label	1	2	3	4	5	6	7
LL	-9,166	-9,155	-9,154	-9,141	-9,131	-9,103	-9,102
no. params	6	7	8	10	11	12	13
BIC	18,381	18,366	18,373	18,362	18,349	18,302	18,308

Table E-4: Mammography misclassification estimates for different DDO and disease models.

True positive rate				95% CI	
Model label	Disease model	DDO model	Estimate	Lower	Upper
1	Standard CTMC	Non-inf.	0.77	0.63	0.86
3	Standard CTMC	H/I/C	0.81	0.68	0.90
4	Latent CTMC	Non-inf.	0.61	0.46	0.74
6	Latent CTMC	H1/H2/I,C	0.69	0.55	0.81

False positive rate				95% CI	
Model label	Disease model	DDO model	Estimate	Lower	Upper
1	Standard CTMC	Non-inf.	0.056	0.053	0.059
3	Standard CTMC	H/I/C	0.056	0.053	0.059
4	Latent CTMC	Non-inf.	0.055	0.053	0.058
6	Latent CTMC	H1/H2/I,C	0.056	0.053	0.059

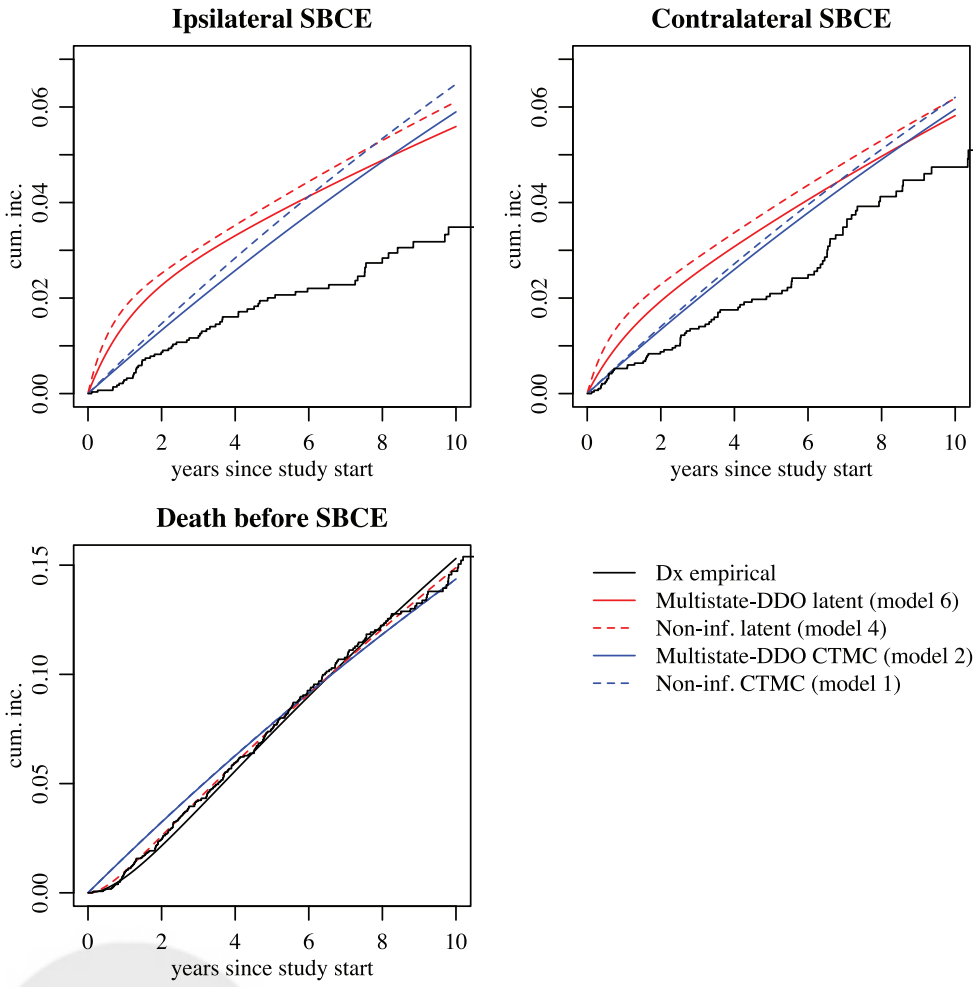

Figure E-1: Sensitivity of SBCE cumulative incidence estimates to choice of disease and observation model. Table E-2 shows model details. Models include informative multistate-DDO models (models 2 and 6), and misspecified non-informative observation models (models 1 and 4). Abbreviations: Dx empirical=empirical estimate of cumulative incidence of diagnosed SBCE events.

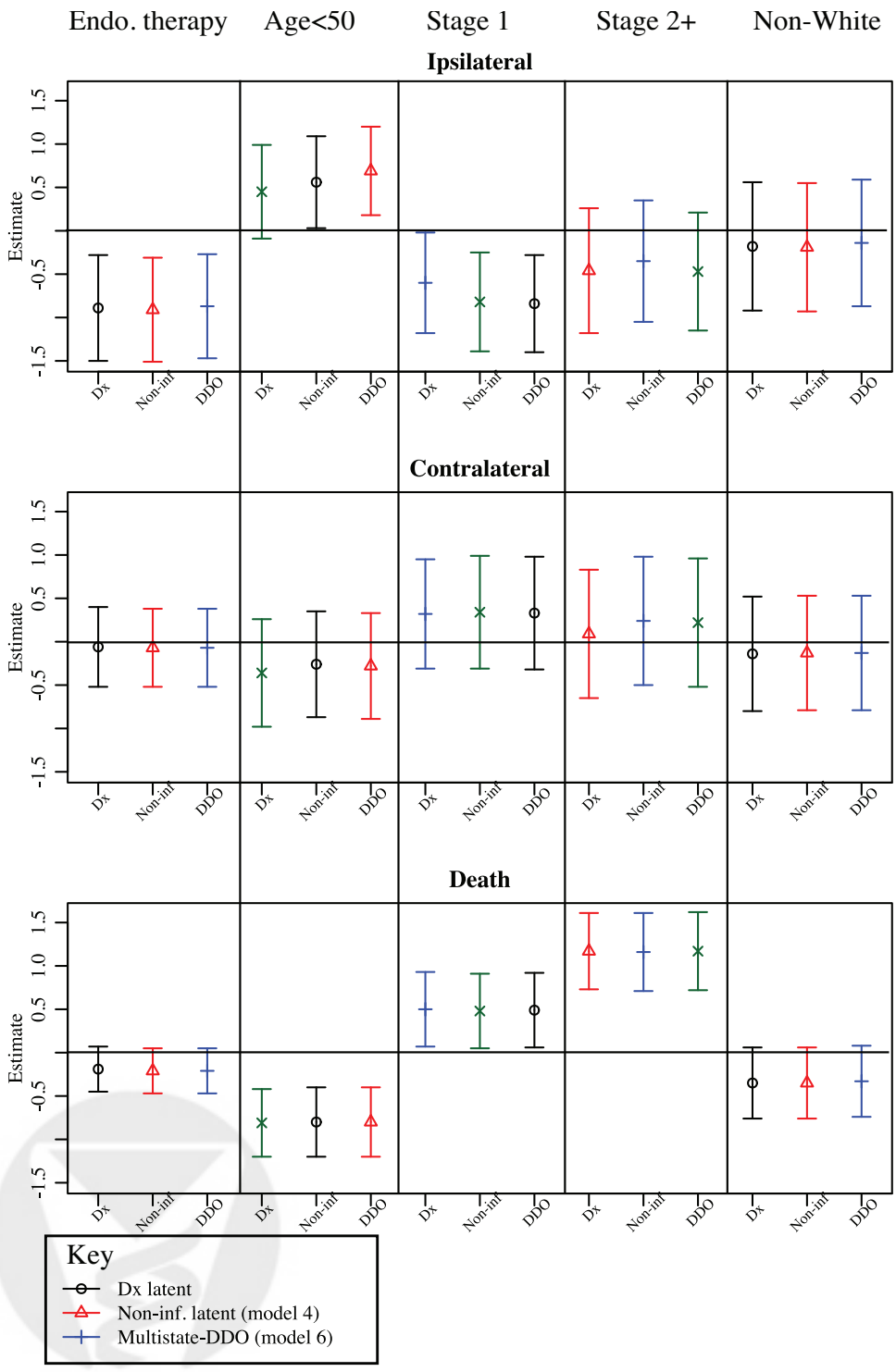

Figure E-2: Point esitmates and 95% confidence intervals for covariate effects via a latent diagnosis time model and different multistate-DDO models (Table E-2). For *Stage 1* and *Stage 2+*, the reference cancer stage is *Stage 0*.

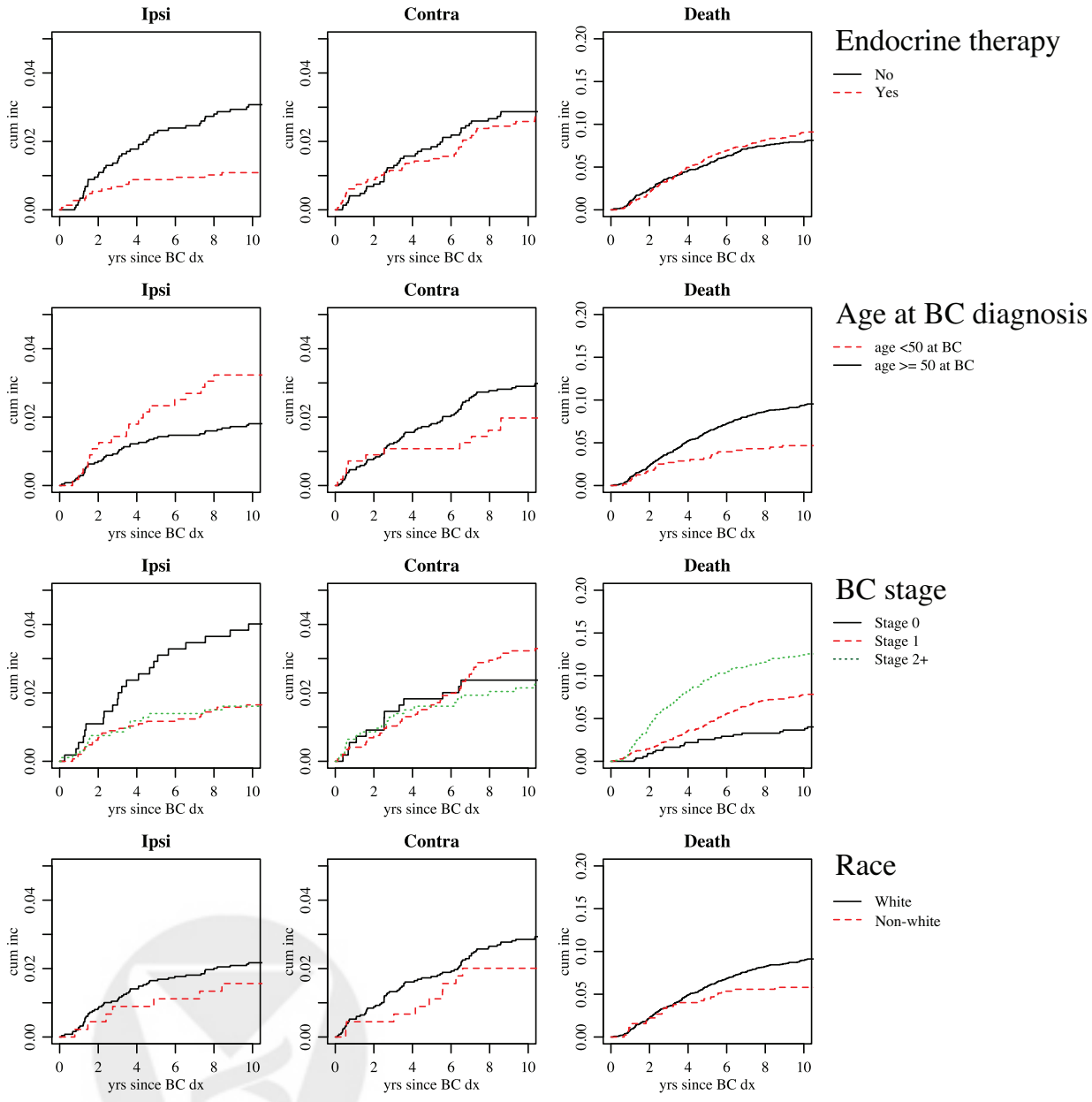

Figure E-3: Empirical cumulative incidence estimates for diagnosis of ipsilateral and contralateral SBCEs and death prior to SBCE, stratified by covariate levels.

References

American College of Radiology (2003). Breast Imaging Reporting and Data System (BI-RADS). American College of Radiology, Reston, Va, 4 edition.

Asmussen, S., Nerman, O., and Olsson, M. (1996). Fitting phase-type distributions via the EM algorithm. Scandinavian Journal of Statistics **23,** 419–441.

Hobolth, A. and Jensen, J. (2011). Summary statistics for endpoint-conditioned continuous-time Markov chains. Journal of Applied Probability **48,** 911–924.

Minin, V. N. and Suchard, M. A. (2008). Counting labeled transitions in continuous-time Markov models of evolution. Journal of Mathematical Biology **56,** 391–412.