Web-based Supplementary Materials 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 manuscript). 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 then final time t_n corresponds to absorbtion of $X(t)$ in state k, we modify the observed data likelihood (eq. (1) in main manuscript) 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 (Section 2.4, eq 1 of the 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 manuscript) 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 2.1 in the main manuscript). 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

$$H_i[a,b] = E\left\{d_t(i) \times I[Y'(t) = (b,0)] | Y'(0) = (a,0)\right\} = \int\limits_0^t \exp\left[(\mathbf{\Lambda} - \mathbf{Q})(u)\right]_{ai} \exp([\Lambda - \mathbf{Q})(t-u)]_{ib}\, du,$$

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

$$M_{ij}[a,b] = E\left\{n_t(i,j) \times I[Y'(t) = (b,0)] | Y'(0) = (a,0)\right\} = \int\limits_0^t \lambda_{ij} \exp\left[(\mathbf{\Lambda} - \mathbf{Q})u\right]_{ai} \exp\left[(\mathbf{\Lambda} - \mathbf{Q})(t-u)\right]_{jb} du.$$

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

$$\frac{\partial}{\partial t}\,\mathrm{E}[d_t(i)I(W_{a0,k0} < t)|Y'(0) = (a,0)] = I(i \not\in A) \sum_{c \notin A} H_i(t)[a,c]\lambda_{ck}.$$

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

$$\frac{\partial}{\partial t}\,\mathrm{E}[n_t(i,j)I(W_{a0,k0} < t)|Y'(0) = (a,0)] = I(i,j \not\in A) \sum_{c \notin A} M_{ij}(t)[a,c]\lambda_{ck} + I(i \not\in A, j = k)S_{ai}(t)\lambda_{ik}.$$

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

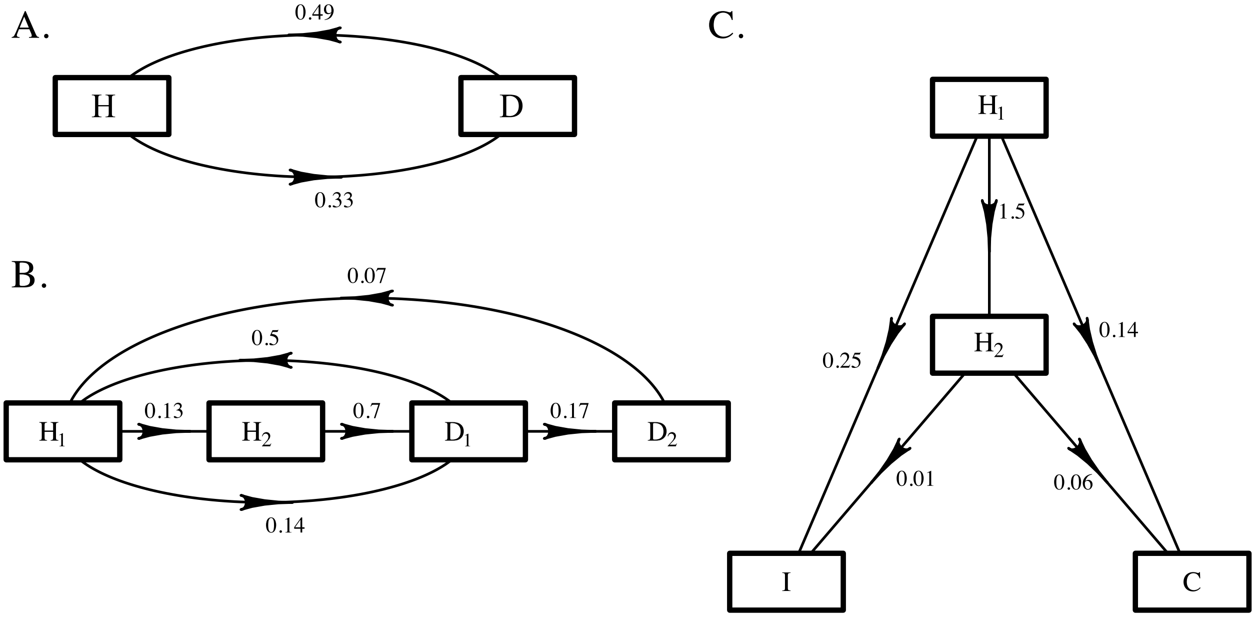

Figure D-1: Data-generating disease models for the simulation study. Data-generating disease models for the simulation study, with transition intensities next to arrows corresponding to each transition. Most studies assumed no covariate effects and common disease models across individuals. A. 2-state 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.

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%

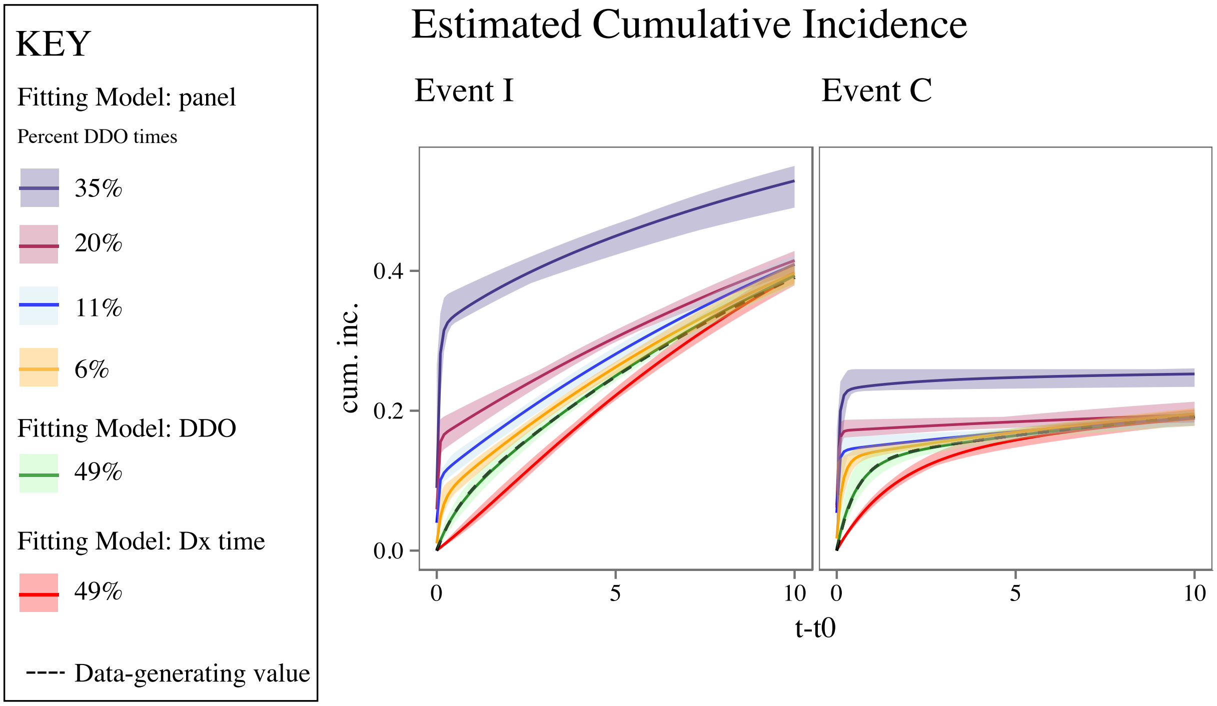

Figure D-2: Functional box plots for simulated data estimates of cumulative incidence for disease events I and C in the latent CTMC competing risks model (Figure D-1C.) Discretely observed data were generated from the disease trajectories according to informative observation times from a DDO model with $q_{H1} = q_{H2} = .25$ 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 how increasing proportions of supplemental scheduled visits mitigates such bias. Also shown is cumulative incidence based on time of diagnosis (Dx time), i.e. the time of the first true positive mammogram.

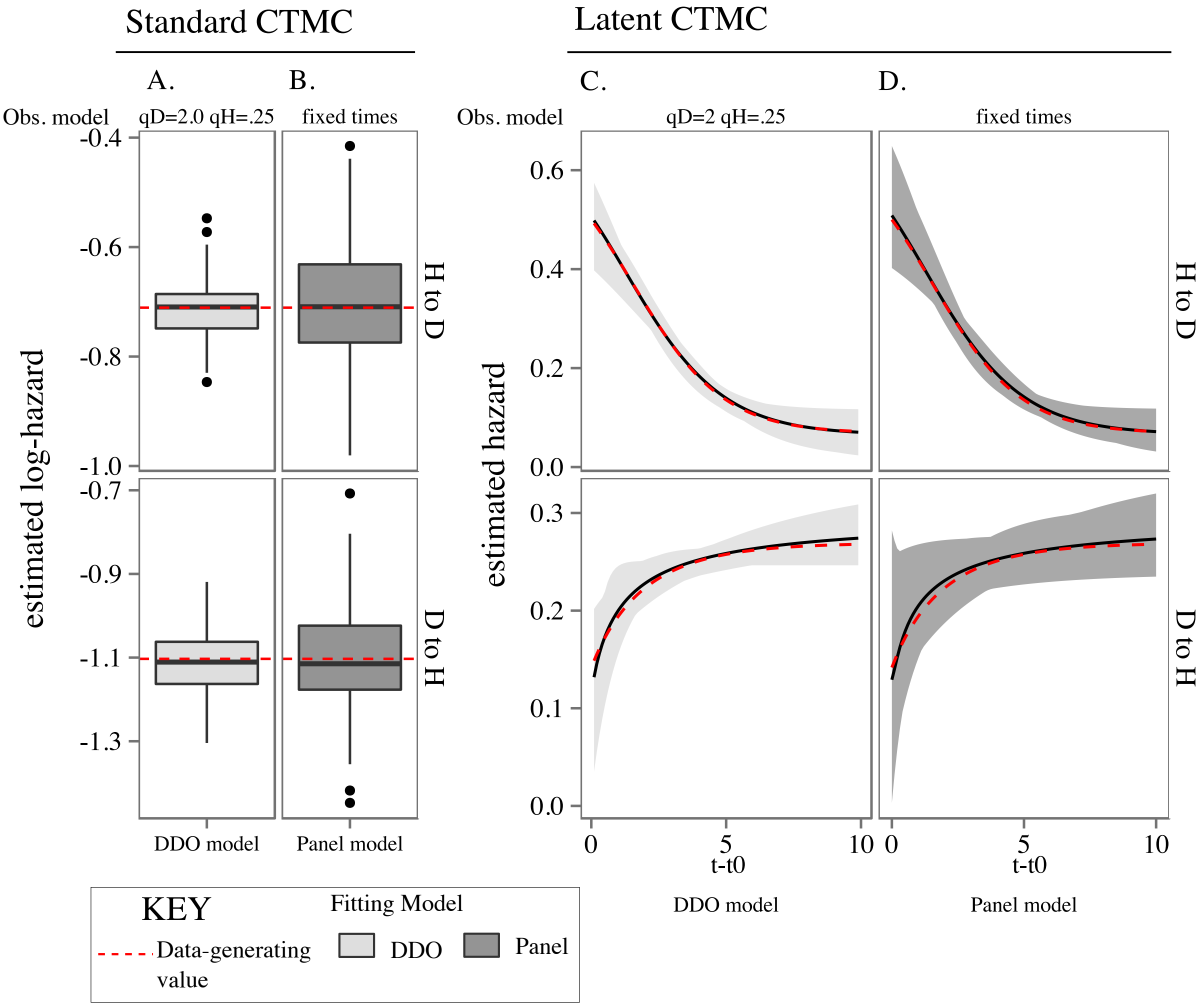

Figure D-3: Box plots/functional box plots for hazard estimates of $H \to D$ and $D \to H$ transitions for standard and latent CTMC reversible disease models (Figure D-1A, D-1B), observed with 15% misclassification error at either DDO times or at fixed times with equal average frequencies. See Table 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.

Simulation studies examining estimated covariate effects on disease transition parameters

These simulations used the competing risks latent CTMC disease model as their disease model framework (Figure D-1C). Informative observation times were generated at a rate of 2/year in states I and C, and .25/year in states H_1, and H_2. Each dataset consisted of 500 individuals, observed at informative sampling times between $t = 0$ and $t = 8$ years. We generated a binary covariate, X, and assumed that relative to those with X=0, those with X=1 had the log intensity rates $H_1 \rightarrow C$ and $H_2 \rightarrow C$ increased by 1.5, and the $H_1 \rightarrow I$ and $H_2 \rightarrow I$ transitions reduced by 1.5. We then fit a correctly specified multistate-DDO model or an analogous, but incorrectly specified, panel model to 100 simulated datasets.

The results summarizing the estimates of these two parameters (denoted β_1 and β_2, with data generating values of 1.5 and -1.5, respectively) are summarized in Table D-3. Also shown are the intercept terms for the intensity rates. For the correctly specified models, the estimates of β_1 and β_2 and their standard errors demonstrate little apparent bias, and the coverage of 95% confidence intervals was close to the nominal 95%. Interestingly, the estimates from the panel model are similar in sign, and not too far off in magnitude, to their data-generating values, despite the obvious bias of other estimates of intensity parameters in the latent model. Moreover, the confidence interval coverage for the estimated covariate effects is not too far from the nominal 95%.

Simulation studies examining usefulness of Bayesian information criterion for model selection

To verify the usefulness of the BIC in latent state selection, we have conducted simulation experiments. In these studies, we used the competing risks latent CTMC disease model (Figure D-1C) and generated informative sampling times with rates of 2/year in states I and C and .25/year in states H_1, and H_2. We generated 50 data sets with 1000 individuals each and fit models that either correctly or incorrectly specified the latent CTMC disease and informative sampling time models. The alternative, incorrect models, varied either the disease model structure or the informative sampling time model. Table D-4 provides details of the additional models fit to the data. After fitting each model to the simulated data, we calculated and ranked the BIC for each of the models fit to the data. Across each of the 50 datasets, the ranking of the BIC was consistent: BIC was lowest for the correctly specified model (Model 3, Table D-4), followed by Models 4, 5, 1, and 2. Thus, using the criterion of selecting a model based on the lowest BIC, the correctly specified model was selected for 50/50 simulated datasets.

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

Table D-3: Summary of covariate estimates based on correctly specified multistate-DDO (M-DDO) or incorrectly specified panel models using simulated data.

Model	Param.	True value	Mean of est.	Sd of est.	Ave. SE of est.	Coverage of 95% CI
M-DDO	β_1	1.50	1.53	0.21	0.19	0.92
M-DDO	β_2	-1.50	-1.51	0.30	0.30	0.97
M-DDO	$\log(\lambda_{12})$	.41	.38	.25	.23	.92
M-DDO	$\log(\lambda_{13})$	-1.39	-1.46	.26	.24	.97
M-DDO	$\log(\lambda_{14})$	-1.95	-2.00	.32	.33	.97
M-DDO	$\log(\lambda_{23})$	-4.53	-4.60	.28	.25	.90
M-DDO	$\log(\lambda_{24})$	-2.87	-2.90	.12	.14	.98
Panel	β_1	1.50	1.66	0.25	0.22	0.92
Panel	β_2	-1.50	-1.29	0.29	0.31	0.90
Panel	$\log(\lambda_{12})$	.41	.95	.51	.38	.78
Panel	$\log(\lambda_{13})$	-1.39	-.13	.54	.40	.12
Panel	$\log(\lambda_{14})$	-1.95	.06	.52	.38	<.01
Panel	$\log(\lambda_{23})$	-4.53	-6.31	.82	.95	.65
Panel	$\log(\lambda_{24})$	-2.87	-3.30	.28	.29	.82

Table D-4: Multistate DDO models fit to simulated competing risks data. Model 3 in this table is the correctly specified multistate-DDO model.

Model label	Disease model	DDO constraints	No. params
1	Standard CTMC	$q_I = q_C$	6
2	Latent CTMC (2 states)	$q_{H_1} = q_{H_2} = q_I = q_C$	8
3	Latent CTMC (2 states)	$q_{H_1} = q_{H_2}, q_I = q_C$	9
4	Latent CTMC (2 states)	$q_{H_1} = q_{H_2}$	10
5	Latent CTMC (3 states)	$q_{H_1} = q_{H_2} = q_{H_3}, q_I = q_C$	13

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 breast cancer 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 breast cancer diagnosis (IQR 52, 71). Approximately one fifth of the sample had a stage 0 (DCIS) primary breast cancer (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 breast cancer.

Table E-1: Characteristics of the Group Health patients with a history of primary breast cancer, 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

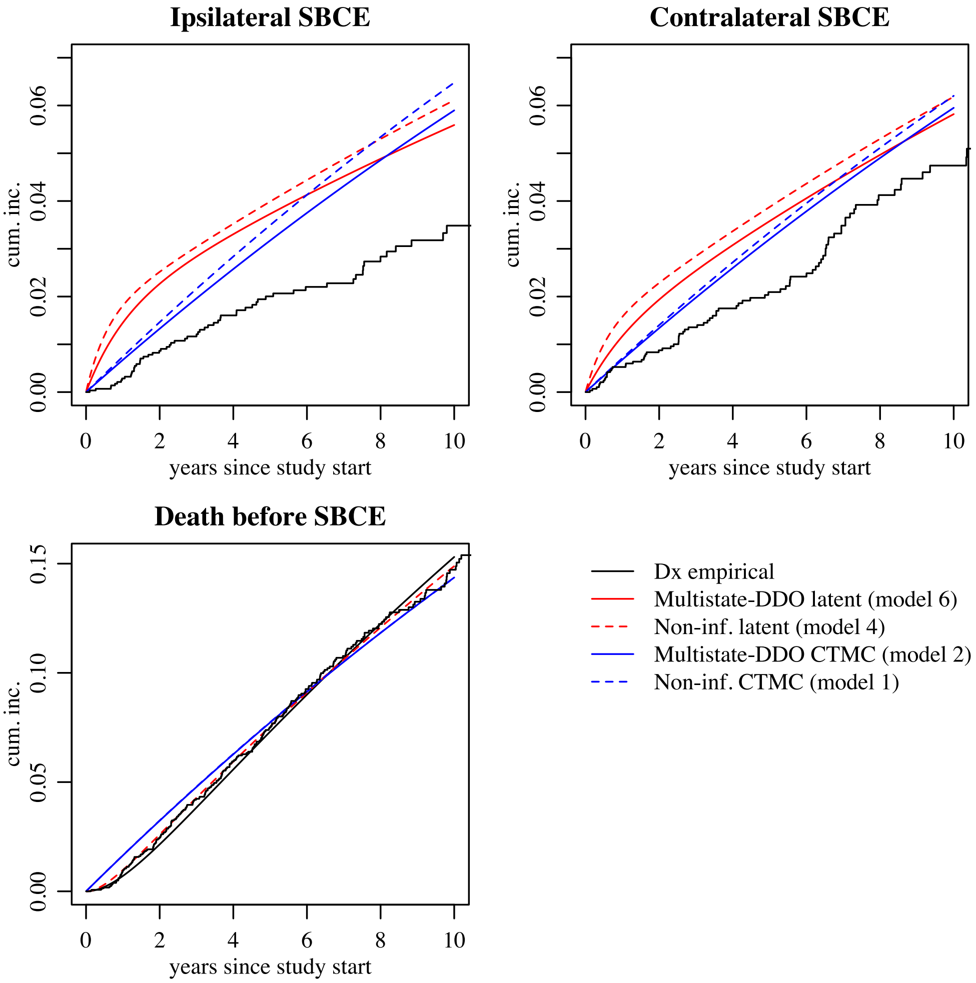

Figure E-1: Sensitivity of SBCE cumulative incidence estimates to choice of disease and observation model. Table E-2 shows model details. Models include informative multistate-DDO models (models 2 and 6), and misspecified non-informative observation models (models 1 and 4). Abbreviations: Dx empirical=empirical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

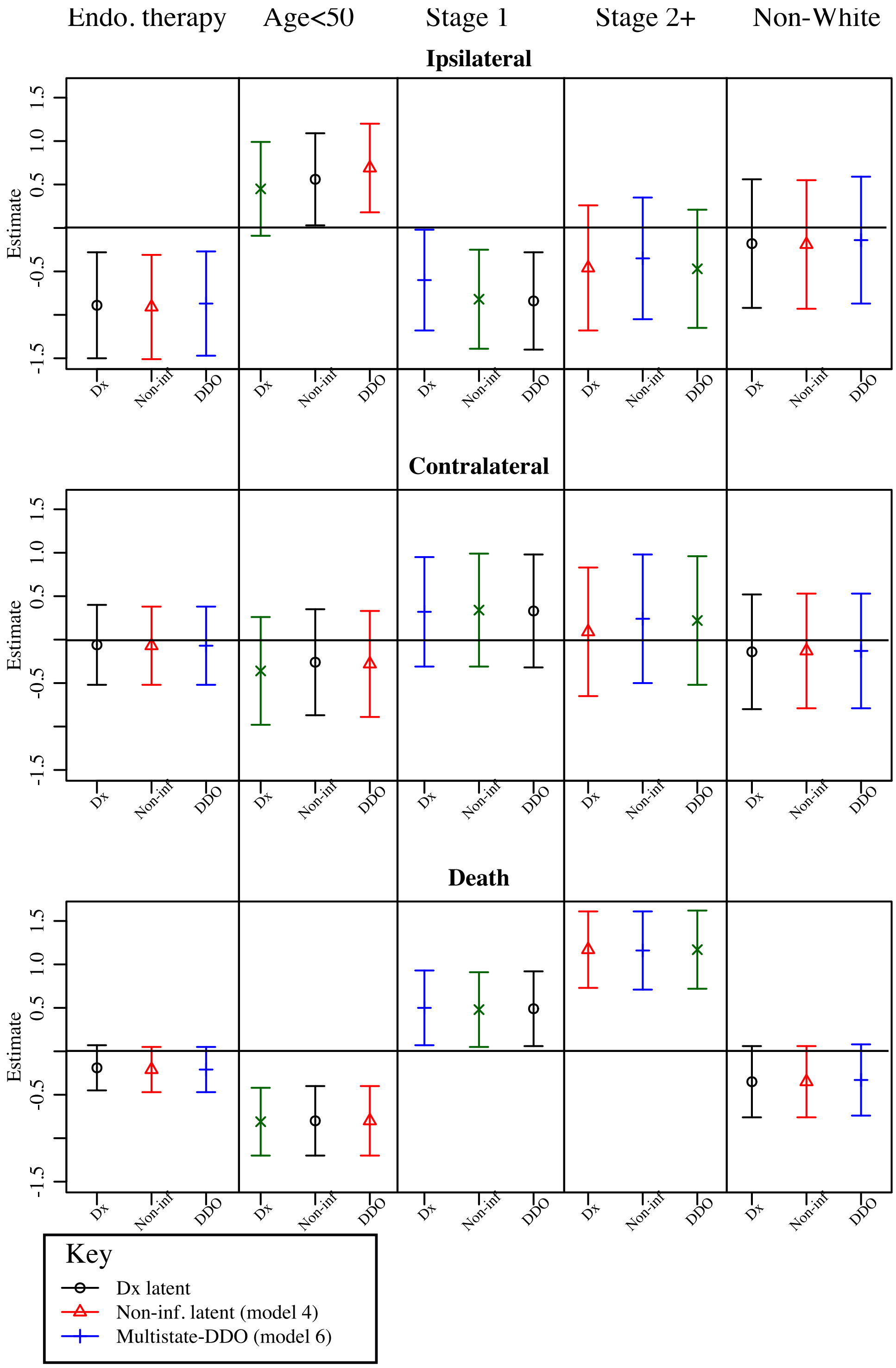

Figure E-2: Point esitmates and 95% confidence intervals for covariate effects via a latent diagnosis time model and different multistate-DDO models (Table E-2). For *Stage 1* and *Stage 2+*, the reference cancer stage is *Stage 0*.

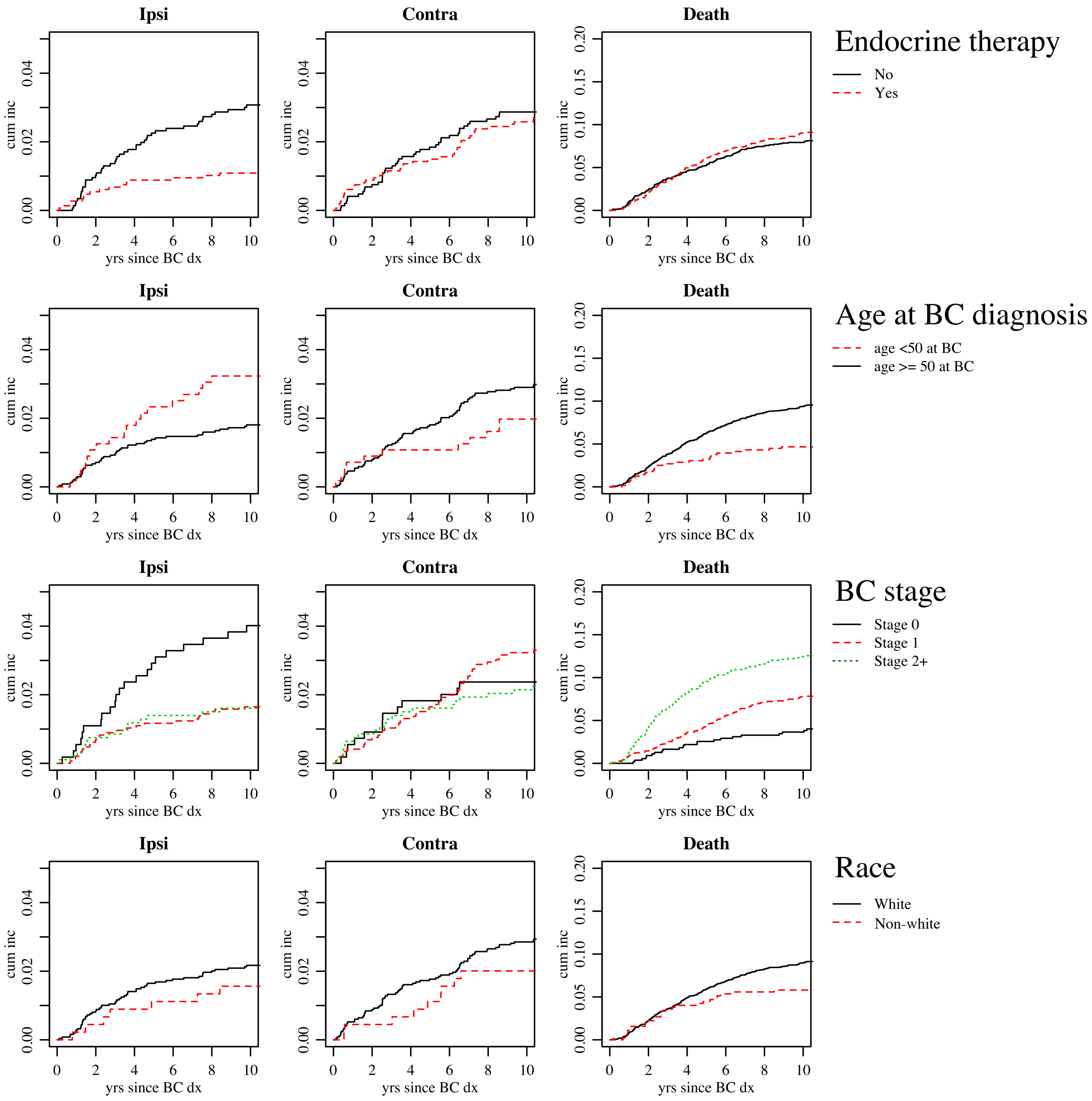

Figure E-3: Empirical cumulative incidence estimates for diagnosis of ipsilateral and contralateral SBCEs and death prior to SBCE, stratified by covariate levels.

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.