A multistate survival model of the natural history of cancer using data from screened and unscreened population

Rikesh Bhatt1 | Ardo van den Hout2 | Nora Pashayan1

1Department of Applied Health Research, University College London, London, UK
2Department of Statistical Science, University College London, London, UK

Correspondence
Rikesh Bhatt, Department of Applied Health Research, University College London, London WC1E 6BT, UK.
Email: rikesh.bhatt@ucl.ac.uk

Funding information
Breast Cancer Now, Grant/Award Number: 2017NovPR10241

One of the main aims of models using cancer screening data is to determine the time between the onset of preclinical screen-detectable cancer and the onset of the clinical state of the cancer. This time is called the sojourn time. One problem in using screening data is that an individual can be observed in preclinical phase or clinically diagnosed but not both. Multi-state survival models provide a method of modeling the natural history of cancer. The natural history model allows for the calculation of the sojourn time. We developed a continuous-time Markov model and the corresponding likelihood function. The model allows for the use of interval-censored, left-truncated and right-censored data. The model uses data of clinically diagnosed cancers from both screened and nonscreened individuals. Parameters of age-varying hazards and age-varying misclassification are estimated simultaneously. The mean sojourn time is calculated from a micro-simulation using model parameters. The model is applied to data from a prostate screening trial. The simulation study showed that the model parameters could be estimated accurately.

KEYWORDS
cancer screening, lead time, Markov model, misclassification, prostate cancer, sojourn time, time-varying hazard

1 | INTRODUCTION

The purpose of cancer screening is to detect cancer prior to the onset of symptoms, in order to improve prognosis by earlier intervention. Early in the natural history of cancer, a newly developed tumor would be asymptomatic and undetectable. At some point it becomes detectable by screening. In nonscreened individuals or if the screening test does not detect the cancer, it may progress to become symptomatic and could be diagnosed clinically. The time by which screen detection advances the diagnosis is the lead time. The maximum lead time is the sojourn time. This is the preclinical screen-detectable period (PCDP).

The potential for screening to detect cancer early depends on the sojourn time, the frequency of screening, and the sensitivity of the screening process. Sensitivity refers to the potential for screening to identify a cancer in the PCDP. Estimates of the probability of cancer initiation, the sojourn time, and the sensitivity inform the design and evaluation of a screening program.
Data from cancer screening programs and cancer screening trials provide unique problems in determining sojourn time. Once a tumor in PCDP is screen-detected and treated by surgical resection, then its progression into the clinical phase is aborted. The exact time of entry into PCDP is not observed. In the data, we only observe individuals either in PCDP or with symptomatic clinical cancer but never both. Models used to determine the sojourn time must overcome this problem.

The phases of the cancer can be defined as mutually exclusive states. A multistate Markov model can be used to determine the transitions between states involving survival data.\(^1\) Such a multistate survival model must incorporate interval-censored and left-truncated (delayed entry) data. Left-truncation accounts for prevalence at first observation. Transition rates and test sensitivity can be estimated simultaneously.\(^2\)

Cancer incidence and mortality increase with age.\(^3\) Parmigiani et al.\(^4\) show that the incidence of preclinical cancers varies with age. Weibull or Gompertz hazards are two particular functions used, in multistate models in survival analysis, for age-varying transition rates.\(^2\) However, most studies of multistate models using cancer screening data have constant exponential hazards.\(^5,6\) Piecewise-constant hazards are used to account for age-varying transition rates.\(^7-9\) Models with piecewise constant hazards have a greater number of parameters to fit than continuous hazards, and the transition rates jump between age groups. Other studies use the Weibull hazard function to model nonhomogeneous rates;\(^8,10\) however, those studies fail to take into account deaths as a competing risk. Death from competing causes has been studied in multi-state models using screening data by Wu et al.\(^11\) and Taghipour et al.\(^12\) However, both studies used piecewise constant hazards as opposed to continuous hazards.

Observations of clinical diagnoses are required to derive the transition rate from PCDP to the clinical state. Existing multi-state models using screening data, use either interval cancers (cancers diagnosed clinically in the interval between two screening rounds) only or all clinical diagnoses.\(^5,13\) All clinical diagnoses consist of interval cancers, symptomatic cancers arising among individuals not attending a screening episode or not invited to screening such as individuals in the control arm of a screening trial. Chen et al.\(^13\) used a three- and five-state multistate models for breast cancer screening data; however, the authors did not consider death from other causes or age-varying transition rates. To our knowledge, there are no published models that combine using data on nonscreened individuals while accounting for death from competing causes.

We propose a four-state model to account for competing risks in survival. The model includes age-varying hazards and age-varying misclassification. Additionally, the survival model allows for left-truncated data and requires data from screened and nonscreened participants to estimate the hazard parameters. This study is novel in that it combines all of these elements into a multistate model. In Section 2, we describe the model and its likelihood function. In Section 3, we describe data from randomized prostate cancer screening trial and its application in modeling the natural history of prostate cancer. Lastly, in Section 4, we perform a simulation study.

### 2 NATURAL HISTORY MODEL

We developed a four-state continuous-time Markov model with states \(S_1\), healthy, \(S_2\), screen-detectable, \(S_3\), clinical cancer, \(S_4\), death. State 3, clinical cancer, refers to the point at which a cancer, if it is not screen-detected, will be diagnosed. Screening tests are not 100% accurate, so we include the possibility of misclassification in the hidden Markov model, which we present in Figure 1. Screen-detected cancers are biopsy confirmed diagnosis following testing positive on a screening test. Therefore a subject in true state 1 cannot have an observation of state 2. The time of the observation refers to the time of the screening test and not of the biopsy. We have not accounted for the transition from state 3 to state 4, as the time spent in state 2, sojourn time, is independent of any transition after state 3. Transition rates are elements of an intensity matrix, and the misclassification rates are elements of a misclassification matrix. Misclassification in this context reflects screening uptake, Prostate-Specific Antigen (PSA) performance, screening interval, uptake of diagnostic investigation. Sensitivity, 1 - misclassification, of the screening process would be lower than the test sensitivity.\(^14\)

The intensity matrix \(Q(t)\), at time \(t > 0\), is given by

\[
Q(t) = \begin{bmatrix}
-q_{12}(t) - q_{14}(t) & q_{12}(t) & 0 & q_{14}(t) \\
0 & -q_{23}(t) - q_{24}(t) & q_{23}(t) & q_{24}(t) \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{bmatrix}.
\]
where \( g_{ij}(t) \) are hazard functions. In this paper, we concentrate on two parametric hazard functions, the exponential hazard and the Gompertz hazard. The Gompertz hazard function takes the form,

\[
q_{ij}(t) = \exp(\lambda_{ij} + \beta_{ij} t + \alpha_{ij} \mathbf{Z}),
\]

where \( \beta_{ij} > 0 \), is an age-varying hazard. The exponential is a constant hazard that is, \( 2 \) with \( \beta_{ij} = 0 \). The variable \( \mathbf{Z} \) is the vector of time-constant covariate values, and \( \alpha_{ij} \), \( \beta \) and \( \lambda \) are the parameters of interest. However, in this article, we focus on the case without covariates, \( \mathbf{Z} \), \( \alpha_{ij} \equiv 0 \) without loss of generality. An assumption we have made is that the transition rates to death from healthy and screen-detectable states are equal. That is the hazards \( q_{14}(t) = q_{24}(t) \), and the parameters are constrained such that \( \lambda_{14} = \lambda_{24} \) and \( \beta_{14} = \beta_{24} \). It is difficult to identify all parameters if we relax one (or both) of these restrictions, such that \( \lambda_{14} \neq \lambda_{24} \) or \( \beta_{14} \neq \beta_{24} \).

The probability of observing \( j \) when the true state is \( i \) at age \( t \), \( \text{Pr}(O_i = j|S_i = i) \), is given by the function

\[
e_{ij}(t) = \begin{cases} 
\exp(\nu_m + \gamma_m i) & \text{if } i = 2, j = 1 \\
1 - e_{21}(t) & \text{if } i = j = 2 \\
1 & \text{if } i = j \neq 2 \\
0 & \text{elsewhere}
\end{cases},
\]

where \( \nu_m \in \mathbb{R} \) and \( \gamma_m \in \mathbb{R} \). \( e_{ij}(t) \) is the \( ij \)th element of the \( 4 \times 4 \) misclassification matrix.

The probability of transitioning from state \( i \) to state \( j \) in time interval \((t_0, t_1]\) is \( \text{Pr}(S_{t_1} = j|S_{t_0} = i) \). This is given by the \( ij \)th element of the transition probability matrix, \( P \). For a constant intensity matrix where \( Q(t_0) = Q \), the probability matrix \( P(t_0, t_1) \) is

\[
P(t_0, t_1) = \exp(Q(t_1 - t_0)).
\]

A piecewise constant approximation (section 3.6 in Reference 2) is employed for an intensity matrix, \( Q(t) \). The transition probability matrix, \( P(t_0, t_1) \), is approximated using a grid with grid length \( l \). The probability matrix is

\[
P(t_0, t_1) = \exp\left( Q(t_0)l \right) \exp\left( Q(t_0 + l)l \right) \cdots \exp\left( Q(t_0 + (n - 1)l)l \right) \exp\left( Q(t_0 + nl)l \right) (t_1 - t_0 - nl),
\]

where \( 0 \leq n < \frac{l - t_0}{l} \). As an example, consider \( t_0 = 2 \) and \( t_1 = 4.5 \) with grid length, \( l = 1 \), then

\[
P(2, 4.5) = \exp(Q(2) 1) \exp(Q(3) 1) \exp(Q(4) 0.5).
\]

The grid length is chosen in relation to the screening frequency and the transition rates. We have chosen a grid length of \( l = 1 \) for the trial data and the simulation data.

### 2.1 Likelihood function

In this section, we explain the likelihood function for the model. The data for the model include interval-censored, censored state, misclassification, right censored, and left-truncated observations. The likelihood contribution for an
individual, who has $N$ observations, at ages $t_1$ to $t_N$, is given by

$$L(\theta|t_1 \ldots t_N) = \epsilon_1^T \left( \prod_{n=1}^{N} L_n(\theta|t_{n-1}, t_n) \right) \mathbf{1}, \quad (7)$$

where $\theta$ is the model parameters. Let $t_0$ be the left-truncation age, the age at which it is assumed all individuals are healthy. The likelihood contribution, $L(\theta|t_1 \ldots t_N)$ is a scalar value; $L_n$ is a matrix dependent on the ages $t_{n-1}$, $t_n$, and the corresponding observation $o_n$. Additionally, $\mathbf{1} = [1, 1, 1]^T$, and $\epsilon_1 = [1, 0, 0, 0]^T$, the basis vector, indicates all individuals are in state 1 (healthy) at age $t_0$.

The random variable time to state 2 is interval-censored since it is only known that the individual enters state 2 between observations. If the observation at age $t_n$ is $o_n = 1$ then there exists a possibility of misclassification defined by (3). The probability of transitioning from state 1 or 2 to state $s$ but observing $o_n$

$$Pr(O_n = o_n|S_{n-1} = \{1 \text{ or } 2\}) = \sum_{s_n \in \{1, 2\}} Pr(S_n = s_n|S_{n-1} = \{1 \text{ or } 2\}) Pr(O_n = o_n|S_n = s_n). \quad (8)$$

Not accounting for the sensitivity of cancer screening tests would under-estimate the length of sojourn time. Satten and Longin proposed methods in determining underlying dynamics of diseases with measurement error. We follow this method along with the assumption that misclassification is independent of previous misclassification. The contributing matrix, $L_n$, in (7) is

$$L_n = P(t_{n-1}, t_n)E(t_n), \quad (9)$$

where

$$E(t_n) = \begin{bmatrix} e_{1o_n}(t_n) & 0 & 0 & 0 \\
0 & e_{2o_n}(t_n) & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \end{bmatrix}. \quad (10)$$

As an example, let us say an individual is first observed at age $t_1$. The second observation at age $t_2$ is a negative screening result, $o_2 = 1$. In this scenario (9) is

$$L_2 = P(t_1, t_2)E(t_2), \quad (11)$$

$$= \begin{bmatrix} p_{11}(t_1, t_2)e_{11}(t_2) & p_{12}(t_1, t_2)e_{21}(t_2) & 0 & 0 \\
0 & p_{22}(t_1, t_2)e_{21}(t_2) & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \end{bmatrix},$$

where $p_{ij}$ is the $ij$th element of the $P$ matrix. Nonscreened individuals would be in either healthy (true state 1) or screen-detectable (true state 2) states if alive and not clinically diagnosed. The contributing matrix for the likelihood Equation 7 for an individual with a censored state observation is $L_n = P(t_{n-1}, t_n)$.

The age at death, state 4, is always known and is not interval-censored. The other absorbing state is clinical cancer, state 3, and is also assumed to be not interval-censored. Hence, observations of states 3 and 4 are exact time transitions, and without misclassification, the contribution to (7) for observations of these states, $d$, is

$$L_N = P(t_{N-1}, t_N)Q(t_{N-1})\delta_d \quad (12)$$

where

$$\delta_3 = \begin{bmatrix} 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 \end{bmatrix}$$
The method above assumes the intensity matrix is constant in the time interval \((t_{N-1}, t_N]\) and transition instantaneously at age \(t_N\). The method aligns with the \(\text{msm}\) package in R.\(^{16}\) The \(\text{msm}\) package does not account for left-truncation, we construct the likelihood function with left-truncation.

The first observation is a censored state for individuals in the control arm or nonscreened participants in screening data. A censored state is when we do not know whether the true state of these individuals is 1 or 2. Screened individuals that have a negative screen result at the first observation may be the result of incorrect observation of healthy state when the true state is screen-detectable. To model the first observation we followed the method described by van den Hout (section 8.4 in "Multi-state models for interval censored data").\(^2\) The left-truncation method assumes all individuals are in true state 1 at some initial age, \(t_0\). Furthermore, individuals who had entered state 3 or had died before \(t_1\) were censored from the analysis. The probability of a left truncated observation is

\[
Pr(O_1 = o_1 | S_0 = 1, S_1 \notin \{3, 4\}) = \sum_{s_1 = (1,2)} \frac{Pr(S_1 = s_1 | S_0 = 1) Pr(O_1 = o_1 | S_1 = s_1) Pr(S_1 \notin \{3, 4\} | S_0 = 1)}{P(S_1 \notin \{3, 4\} | S_0 = 1)}.
\]

The contribution of the first observation, \(L_1\), in (7) is

\[
L_1 = \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0
\end{bmatrix}
\end{equation}

The likelihood function is a product of individual contributions \(L^{(k)}\) for individual \(k\),

\[
L(\theta | t_1 \ldots t_N) = \prod_k L^{(k)}(\theta | t_1 \ldots t_N).
\]

We maximize the log-likelihood function to estimate model parameters. The SEs of the estimates are determined from the resulting Hessian.

## 3 PROSTATE CANCER SCREENING DATA AND MODEL

### 3.1 Data

We used data from the Prostate Lung Colorectal and Ovarian screening trial.\(^{17}\) A total of 76 685 men, aged 55 to 74 years, were randomly assigned to intervention (annual screening with the blood test, PSA for 6 years, 38 340) or control (usual care, which included opportunistic PSA testing, 38 345) and followed up for a median of 13 years.\(^{18}\) The individuals enter and leave the trial at different ages, either due to death, positive cancer diagnosis or end of the follow-up period, see Figure 2. The state table in Table 1 illustrates the problem of deriving the sojourn time from cancer screening data; that is there is no observed transition from state 2 to state 3.

Figure 3 shows more screen detected cancers at first screening compared to subsequent screenings, over 28% of all the screen-detected cancer. Participants with screen-detected cancer in the remaining five screenings represented 1.03% of all participants. To perform the analysis of the prostate cancer data, we took a random sample of 10 000 participants, 5000 from the screening arm and 5000 from the control arm.
FIGURE 2  Number of individuals by age (rounded to closest year) during the trial. Arm 1 is the screen-tested arm and Arm 2 is the control arm. Individuals leave the trial after screen detection and clinical diagnosis. The individuals are in the trial for a median of 13 years. The average age during the trial is 68 years

FIGURE 3  Detection of cancers per screening round

TABLE 1  PLCO data state table: −1 refers to a censored state where we only know the individual is not clinically diagnosed or dead

| Arm    | From state | To state | Total       |
|--------|------------|----------|-------------|
|        |            | 1        | 2           | 3           | 4           |             |
| Screening | 1         | 139925   | 1475       | 2175       | 5544       | 35871 men   |
|        | 2         |          | 574\textsuperscript{a} |            |            |             |
| Control | −1        | 27692    | 4040       | 6591       | 38323 men  |

\textsuperscript{a}First and only observation.

3.2  Model

We present three models of the natural history of prostate cancer, using different hazard distributions on the transitions. Firstly, we apply the model (Figure 1), with exponential hazards on all transitions, to the trial data. Let us call this Model A. Secondly, we employ a Gompertz hazard on all transitions. Let us call this Model B. Lastly, a model with Gompertz hazard on transitions from state 1 to state 2 and state 1 to state 4 (states 2 to 4), and an exponential hazard on state 2 to state 3. Let us call this Model C. We estimate model parameters, by maximizing the likelihood function (15).

We assume all patients are healthy (in state 1) at some age \( t_0 \) before the trial starts. Healthy refers to no detectable cancer. We assumed that every individual of the PLCO trial was in state 1 at age 40. Secondly, the hazard rate of death from competing causes is equal when patients are in state 1 or state 2. Lastly, we assumed that hazards are monotonically increasing or are constant with respect to age.

From (2) and (3) the hazards and misclassification functions are

\[
q_{12}(t) = \exp(\lambda_{12} + \beta_{12} t)
\]
TABLE 2  Estimated parameters for Model C on prostate cancer screening data. Model C has a Gompertz hazard on transition from state 1 to state 2 ($q_{ij} = \exp(\lambda_{ij} + \beta_{ij} t)$), and on transitions to state 4, exponential on transition on transition from state 2 to state 3 ($q_{ij} = \exp(\lambda_{ij})$), and with age-varying misclassification ($\exp(\nu_m + \gamma_m t) + \exp(\nu_m + \gamma_m t)$). Age was centered in this model where an age of 68 was set to 0. The programs sensitivity for a 55 year old is 18.4% and for a 80-year old the sensitivity is 92.5%.

| Parameters | Value (95% CI)   | SE  |
|------------|-----------------|-----|
| $\lambda_{12}$ | $-4.47$ ($-4.55 : -4.39$) | 0.04 |
| $\lambda_{14} = \lambda_{24}$ | $-4.34$ ($-4.40 : -4.29$) | 0.03 |
| $\lambda_{23}$ | $-1.47$ ($-1.84 : -1.10$) | 0.19 |
| $\beta_{12}$ | 0.04 ($0.03 : 0.05$) | 0.00 |
| $\beta_{14} = \beta_{24}$ | 0.09 ($0.08 : 0.10$) | 0.00 |
| $\nu_m$ | $-0.59$ ($-0.26 : 0.51$) | 0.56 |
| $\gamma_m$ | $-0.16$ ($-0.26 : -0.07$) | 0.05 |

The hazards and misclassification functions for Model A are Equations (16) to (20), where $\beta_{12} = \beta_{14} = \beta_{24} = \beta_{23} = \gamma_m = 0$. The hazards and misclassification functions for Model B are Equations (16) to (20), where $\beta_{ij} > 0$, and $\gamma_m \neq 0$. Model C is similar to Model B except $\beta_{23} = 0$.

The parameter estimates of Model C are given in Table 2, and the graphs of the hazards and misclassification are presented in Figures A1 and A2. The results indicate that constant hazards yield worse results than time-varying hazards. The AIC of Model A was 30243, and the AIC of Model C was significantly lower with 29412. The SE of the misclassification parameter, $\nu_m$, was very high (50.20) for Model A. The high SE suggests the exponential model, Model A, was not able to correctly identify the true misclassification rate. Model B yielded a negative $\beta_{23}$. However, since we restricted our analysis to $\beta_{23} \geq 0$ we concentrate on Model C.

3.2.1  Model application

The sojourn time can be calculated from a micro-simulation using the model parameter estimates in Table 2. The length of time in state 2 (the screen-detectable phase) given that the individual enters state 3 (clinical cancer) is the sojourn time. The algorithm for the micro-simulation is in Box 1. To generate survival times we follow the method described by Bender et al.19 The time to an event, $t_n - t_{n-1}$, is given by

$$t_n = H^{-1}[-\ln(U) + H(t_{n-1})],$$

where $H$ is the cumulative hazard function and $U \sim U[0, 1]$. The micro-simulation consisted of 10 000 individuals, 3134 of whom entered state 3. The mean sojourn time, 3.79 (95% confidence interval: 3.58-3.85) years, is found by taking the mean of the time in state 2 of the 3134 simulated individuals.

Overdiagnosis is the proportion of screen-detected cases that would not have been diagnosed clinically within a person’s lifetime. Using the model, we can calculate overdiagnosis by employing a micro-simulation. Varying the start and stop age, and the frequency of screening, one could determine a screening strategy that minimizes overdiagnosis while detecting harmful cancers.

3.2.2  Model validation using cumulative incidence function

To validate the model, we compare nonparametric and model-based cumulative incidence functions (CIF). For state $k$, the CIF $I_k$ is the probability, $\Pr(T \leq t, S = k)$ of experiencing the absorbing state $k$ before age $t$.20 A comparison of the CIF $I_3$ and $I_4$ provides a heuristic measure for the goodness of fit.
Algorithm 1. Algorithm to generate data, used to calculate sojourn time

Micro-simulation for sojourn time calculation:

1. Initial state is 1 the participant’s age is \( t_0 \in U[40, 45] \).
2. From this state the individual can transfer to 2 and 4. We simulate the age of entry into these states and take the minimum. The simulated ages in state 2 and 4 is given by

   \[
   t_{12} = \frac{1}{\beta_{12}} \ln (-\beta_{12} \ln(U) + \exp(\lambda_{12} + \beta_{12} t_0)) - \frac{\lambda_{12}}{\beta_{12}}
   \]
   \[
   t_{14} = \frac{1}{\beta_{14}} \ln (-\beta_{14} \ln(U) + \exp(\lambda_{14} + \beta_{14} t_0)) - \frac{\lambda_{14}}{\beta_{14}}
   \]

   respectively, where \( U \sim U[0, 1] \).
3. If \( t_{14} \leq t_{12} \) Stop. Otherwise move to the next step.
4. From this state, the individual can transfer to 3 and 4. We simulate the age of entry into these states and take the minimum. The simulated ages in state 3 and 4 are given by

   \[
   t_{23} = -\ln(U) \exp(-\lambda_{23}) + t_{12}
   \]
   \[
   t_{24} = \frac{1}{\beta_{24}} \ln (-\beta_{24} \ln(U) + \exp(\lambda_{24} + \beta_{24} t_0)) - \frac{\lambda_{24}}{\beta_{24}}
   \]

   respectively, where \( U = U[0, 1] \).

The nonparametric estimate of the CIF for state \( k \) is given by

\[
\hat{I}_k(t) = \sum_{j: t_{j-1} \leq t} \frac{d_{kj}}{n_j} \prod_{j-1 \leq j' \leq t} \left( 1 - \frac{d_{kJ-1} + d_{kj-1}}{n_{j-1}} \right),
\]

where \( n_j \) is the number at risk at age \( t_j \) and \( d_{kj} \) is the number of observations of state \( k \) at age \( t_j \). We estimate \( I_3 \) and \( I_4 \) for the control group in the PLCO data. The model-based estimate of \( \Pr(T \leq t, S = k) \) is computed using the transition probabilities from the model.

We use 38,323 individuals of the control group to validate the model. The model CIF replicates the ages at baseline of the PLCO data, see Figure 4. The estimates for \( I_3 \), in Figure 5A, shows good agreement between the model and the PLCO data. However, at an advanced age the accuracy of the model becomes less accurate. The estimates for \( I_4 \) of the data and model in Figure 5B match closely. The graphs of the CIFs indicate there is a good fit between the model and the data.

![Figure 4](image-url)  
**Figure 4** Number of individuals by age at the start of the trial. Arm 1 is the screen tested arm and Arm 2 is the control arm. The average age at the start of the trial is 62 years.
4 | SIMULATION STUDY

Data generation

To validate the models, we performed a simulation study, estimating model parameters and comparing them to the input parameters. The simulation study involved creating 200 sets of panel data. Each panel data were created with distributions with known parameters. To create the data, we followed the algorithm in Box 2.

Algorithm 2. Algorithm to generate simulation data

Simulation of cancer screening trial data:

1. Initial state is 1 the participants age is between 20 and 39, say \( t_{-1} \).
2. From this state, the individual can transfer to 2 and 4, from Equation (21). We simulate the ages at these states, taking the minimum of the simulated ages and the corresponding state. If individual enters state 4 go to step 4.
3. Simulate ages at states from state 2, the individual can transfer to states 3 or 4.
4. If the entry to state 2 is before age 40, or the entry to the absorbing states is before \( t_1 = t_{-1} + 21 \), go back to step 1.
5. Amend the data such that the first observation is at \( t_1 \).
6. Implement screening procedure, including misclassification, for the screening arm. We have chosen annual-screening to mimic the PLCO data.
7. For control arm from age \( t_1 \) add in the censored state, which we label \( -1 \).
8. Remove observations of state 3 and state 4, where an observation of state 2 exists.
9. To implement the grid structure, we add "observations" of censored state \( (-1) \), where the time between observations is more than one year.
10. Implement maximum follow-up period, remove all observations after 12 years.
11. Add right censoring state a year after the last observations if the last observation is 1 or \(-1\).

We present a study of the model in Figure 1, using two different hazard functions on the transition from state 2 to state 3. Firstly, the model with an exponential hazard between state 2 and state 3, and a Gompertz hazard on the remaining transition. This is Model C from Section 3.2. Each dataset contains 1600 individuals, 800 simulated individuals in the screening arm and 800 in the control arm. Secondly, the model employed Gompertz hazards on all transitions. This is Model B from Section 3.2. The simulation study for this model used data sets containing 1000 individuals, 500 in each arm. Both models include age-varying misclassification. The averages of the parameters in 200 simulations, estimated by a Nelder-Mead algorithm, are compared numerically to the input parameters. We assumed the simulation
TABLE 3 Simulation results for a Model C. Model C has a Gompertz hazard on transition from state 1 to state 2, and on transitions to state 4 \((q_{ij} = \exp(\lambda \beta t))\), exponential on transition on transition from state 2 to state 3 \((q_{ij} = \exp(\lambda t))\) and with age-varying misclassification \((\exp(\nu \gamma t))\). The true parameter values were chosen to be similar to the prostate cancer screening trial in Table 2.

| Parameters | True value | Mean | Rel. bias (mean) | Median | Rel. bias (median) | Average SE | Emp. SE |
|------------|------------|------|------------------|-------|-------------------|------------|--------|
| \(\lambda_{12}\) | −4.00 | −3.98 | 0.00 | −3.99 | 0.00 | 0.15 | 0.15 |
| \(\lambda_{14} = \lambda_{24}\) | −4.50 | −4.45 | 0.01 | −4.44 | 0.01 | 0.13 | 0.11 |
| \(\lambda_{23}\) | −1.50 | −1.54 | −0.03 | −1.50 | 0.00 | 0.53 | 0.55 |
| \(\beta_{12}\) | 0.05 | 0.05 | 0.07 | 0.05 | 0.03 | 0.02 | 0.02 |
| \(\beta_{14} = \beta_{24}\) | 0.10 | 0.10 | 0.00 | 0.10 | 0.00 | 0.02 | 0.02 |
| \(\nu_m\) | −0.50 | −0.85 | −0.69 | −0.50 | 0.01 | 4.32 | 3.11 |
| \(\gamma_m\) | −0.20 | −0.14 | 0.32 | −0.20 | 0.01 | 0.78 | 0.62 |

TABLE 4 Estimates of the model parameters and relative bias with misspecifying the left truncation age, based on simulation study with left truncation age of 40 years: Some parameters have not been identified but the effect on parameters is small.

| Parameters | True value | Left truncation age 40 | Left truncation age 30 | Left truncation age 50 |
|------------|------------|------------------------|------------------------|------------------------|
| \(\lambda_{12}\) | −4.00 | −3.99 | −3.95 | −3.95 |
| \(\lambda_{14} = \lambda_{24}\) | −4.50 | −4.44 | −4.45 | −4.47 |
| \(\lambda_{23}\) | −1.50 | −1.50 | −1.49 | −1.64 |
| \(\beta_{12}\) | 0.05 | 0.05 | 0.04 | 0.05 |
| \(\beta_{14} = \beta_{24}\) | 0.10 | 0.10 | 0.10 | 0.10 |
| \(\nu_m\) | −0.50 | −0.50 | −0.50 | −0.43 |
| \(\gamma_m\) | −0.20 | −0.20 | −0.19 | −0.20 |

correctly identifies the parameters if the relative bias is less than 0.1 for every parameter in the model. The models were implemented in R and simulated with the use of a high performance computing cluster.

4.1 Results of the simulation study

The piecewise approximation method accurately estimates the parameters of age-varying hazards in the multistate model in Figure 1 with (16) to (20). The simulations of Model C correctly identified all parameters. Table 3 shows the median relative bias is less than 0.1 for all the parameters. However, not all parameter estimates have a mean relative bias of less than 0.1. Of the 200 simulations, 162 simulations yielded a positive definite Hessian and converged. Table 4 compares results of truly specifying the left-truncation age, and wrongly assuming the left-truncation age to be 10 years lower, and 10 years higher. The relative bias in the estimates was small as seen in Table 4. Simulation results for Model C where the number of subjects is 300, 600, 900, and 1200 are presented in Table B1. The results show that if the number of subjects is 900 or below that not all the parameters are identifiable, with a relative bias of less than 0.1. The SEs prominently decrease as the number of subjects increase. Secondly, Model B has a relative bias of less than 0.1 for all parameters (see Table 5). A positive definite Hessian was exhibited in 173 of 200 simulations.

The coverage is the proportion of simulations in which the true parameter value is contained within the confidence interval. The coverage is calculated only with simulations with a positive definite Hessian. The 95% confidence interval coverage of both models is presented in Figures 6 and 7. The true parameter values were well covered in the simulations. The coverage was between 85% and 97% for all parameters.
The coverage graphs show in some simulations the SEs of the misclassification parameters were large. Even though the true parameter values lie within the 95% confidence interval, the wide confidence interval in these simulations indicates the parameter estimates were not precise. The confidence intervals that do not contain the true parameter value in Figures 6B,E and 7A,D; the misclassification parameters Figure 7G,H, may indicate that overestimating the $\lambda$’s in the hazard functions is compensated by underestimating the effect of age, $\beta$’s.

The model can identify parameters similar to the estimates for prostate cancer in Table 2. However, the hazard rate, from state 2 to state 3, might be monotonically increasing for other different cancers. We have shown these parameters are identifiable too.

## DISCUSSION

We have developed a model using cancer screening data and the corresponding likelihood function. The simulation study shows that a maximum likelihood estimation can retrieve the correct parameter values. The simulation study also shows that the piecewise approximation method to determine the probability matrix, (5), is appropriate for age-varying hazards.

We use a cancer screening trial to model the natural history of prostate cancer. The mean sojourn time is calculated from the prostate cancer model parameters by a micro-simulation. The mean sojourn time (3.79 years) is lower than might be expected for prostate cancer. However, this is due to the high amount of contamination in the trial, where men in the control arm have taken the screening test.\(^\text{21}\) Our model assumes that participants in State 1 and State 2 are random sample of the general population. However, PLCO trial, as any other screening trial, has healthy volunteer effect, where the study participants might be in better health than the general population.\(^\text{22}\) Compliance with screening and survival patterns could be different between the screening trial participants and the general population. As we use data from both arms of the screening trial to derive the transition probabilities, healthy volunteer effect would not affect the internal validity of our findings but could affect its generalizability. Correction for healthy volunteer effect is important but would require additional data and further assumptions.\(^\text{23}\) Additionally we assumed all subjects were in state 1 at age 40. However, we show that misspecifying the left-truncation age by 10 years would have a minimal effect on the parameters estimates.

The extent of overdiagnosis is important in the evaluation of a screening strategy. A benefit of the model is that it can be used to determine the effects of different screening programs on the cancer. Using the model we can determine the proportion of cancers detected by a screening program, and the proportion of those cases that are non-overdiagnosed harmful cancers.

Existing software is available for many of the elements needed for the natural history model but not all. For example, models in the msm package in R can have time-varying hazards, and hidden Markov models with time-varying misclassification.\(^\text{16}\) However, we developed software for models that include left-truncated data, which msm is unable to do.
Figure 6  Coverage graphs of model C simulations. The graphs show the coverage of the simulations of the model with Gompertz hazard (1 → 2, 1 → 4, 2 → 4), exponential hazard (2 → 3), and age-varying misclassification. The x-axis is $\hat{CI}_p - \theta_p$, where $\hat{CI}_p$ is the 95% confidence interval, $p$ is the parameter, and $\theta_p$ is true parameter value for $p$. The grey lines are where the true parameter values lie in the 95% confidence intervals. The black lines are where the values do not lie in the 95% confidence interval. A, $\lambda_{12}$ has a coverage of 94%; B, $\lambda_{14} = \lambda_{24}$ has a coverage of 93%. Note all the black lines in figure is larger than 0. C, $\lambda_{23}$ has a coverage of 89%. D, $\beta_{12}$ has a coverage of 97%; E, $\beta_{14} = \beta_{24}$ has a coverage of 96%. Note all the black lines are less than 0. F, $\lambda_m$ has a coverage of 93%; G, $\beta_m$ has a coverage of 88%.
FIGURE 7  Coverage graphs of model B simulations. The graphs show the coverage of the simulations of a model with Gompertz hazard on all transitions, and age-varying misclassification. The x-axis is $\hat{CI}_p - \theta_p$, where $\hat{CI}_p$ is the 95% confidence interval, $p$ is the parameter, and $\theta_p$ is true parameter value for $p$. The grey lines are where the true parameter values lie in the 95% confidence intervals, and the black lines are where the values do not lie in the 95% confidence interval. A, $\lambda_{12}$ has a coverage of 85%. Note almost all the black lines are greater than 0; B, $\lambda_{14} = \lambda_{24}$ has a coverage of 90%; C, $\lambda_{23}$ has a coverage of 85%; D, $\beta_{12}$ has a coverage of 93%. Note almost all black lines are less than 0; E, $\beta_{14} = \beta_{24}$ has a coverage of 94%; F, $\beta_{23}$ has a coverage of 89%; G, $\lambda_m$ has a coverage of 88%. Note all the black lines are greater than 0; H, $\beta_m$ has a coverage of 89%. Note all the black lines are less than 0.
We have derived the transition to the clinical state by using data on clinical cases from both the screened and unscreened population. Using only the interval cases would capture only fast progressing cancers and cancers that are missed by screening. The sojourn time distribution derived from interval cancers would not represent the sojourn time distribution of all cancers. Not using an unscreened cohort will cause an unrepresentative mix of the cancers in these studies. Most other models that use screening data neglect the control arm or use it to determine prevalent cases, test sensitivity, or used to test the validity of the model. Not using clinically diagnosed cancers from an unscreened population will bias results.

Some studies calculate prevalence from a control arm or assume prevalence from other data sources. Figure 3 highlights the need for accurately describing prevalent cases. The left-truncation method is the most suitable method.

We estimate hazard rates simultaneously with age-varying misclassification. In many models, test sensitivity is an input though there are some notable exceptions. In a study of breast cancer, Taghipour et al show test sensitivity changes with age. However, the calculated sensitivity is not continuous but piecewise constant. The authors do not use the left-truncation method and diagnosed cases in the first 6 months of enrolment are removed from the study. While both Uhry et al and Wu et al have assumed constant sensitivity without competing risks.

In our analysis, we use a Gompertz hazard, but this does not prohibit other time-varying hazards being used in the model. For the transition from state 2 to state 3, a Gompertz hazard implies mean sojourn time becomes shorter with age. Other hazards that have non-monotonic functions may also be employed, but the additional complexity may require a further simulation study. Incorporating a hazard function of time since entry into state 2 could improve the model. Since state 2 is interval censored this semi-Markov model would require integrating through all possible entry times. While this is possible and perhaps may increase the accuracy of the model, the CIFs indicate that there is a good fit between the model and the data. In the model, we assumed that individuals in the healthy and in screen-detectable state experience the same other cause mortality rate. Another assumption we made is that a false-negative screening is independent of previous false negatives. These assumptions are unlikely to impact findings.

Our approach solves many issues in modeling the natural history of cancer using cancer screening data. The model is particularly novel in that it incorporates age-varying hazards and age-varying misclassification, with a state of death as a competing risk, while using data from non-screened individuals.

ACKNOWLEDGEMENTS
The authors thank the National Cancer Institute for access to NCI’s data collected by the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. The statements contained herein are solely those of the authors and do not represent or imply concurrence or endorsement by NCI. The study is funded by Breast Cancer Now (Grant reference number: 2017NovPRI0241).

ORCID
Rikesh Bhatt https://orcid.org/0000-0003-3924-3619

REFERENCES
1. Andersen PK. Multistate models in survival analysis: a study of nephropathy and mortality in diabetes. Stat Med. 1988;7(6):661-670. https://doi.org/10.1002/sim.4780070605.
2. Van den Hout A. Multi-State Survival Models for Interval-Censored Data. Boca Raton, FL: Chapman & Hall; 2017.
3. Pike MC, Krailo MD, Henderson BE, Casagrande JT, Hoel DG. Hormonal risk factors, breast tissue age and the age-incidence of breast cancer. Nature. 1983;303:767-770. https://doi.org/10.1038/303767a0.
4. Parmigiani G, Skates S. Estimating distribution of age of the onset of detectable asymptomatic cancer. Math Comput Model. 2001;33(12):1347-1360. https://doi.org/10.1016/S0895-7177(00)00320-4.
5. Uhry Z, Hedelin G, Colonna M, et al. Multi-state Markov models in cancer screening evaluation: a brief review and case study. Stat Methods Med Res. 2010;19(5):463-486. https://doi.org/10.1177/0962280209359848.
6. Shih HC, Chou P, Liu CM, Tung TH. Estimation of progression of multi-state chronic disease using the Markov model and prevalence pool concept. BMC Med Inform Decis Mak. 2007;7(1):34. https://doi.org/10.1186/1472-6947-7-34.
7. Wu WYY, Nystrom L, Jonsson H. Estimation of overdiagnosis in breast cancer screening using a non-homogeneous multi-state model: a simulation study. J Med Screen. 2018;25(4):183-190. https://doi.org/10.1177/0969141317733294.
8. Hsieh HJ, Chen THH, Chang SH. Assessing chronic disease progression using non-homogeneous exponential regression Markov models: an illustration using a selective breast cancer screening in Taiwan. Stat Med. 2002;21(22):3369-3382. https://doi.org/10.1002/sim.1277.
9. Wu GHM, Auvinen A, Maattanen L, et al. Number of screens for overdetection as an indicator of absolute risk of overdiagnosis in prostate cancer screening. *Int J Cancer*. 2012;131(6):1367-1375. https://doi.org/10.1002/ijc.27340.
10. Chen TH, Yen M, Shiu M, Tung T, Wu H. Stochastic model for non-standard case-cohort design. *Stat Med*. 2004;23(4):633-647. https://doi.org/10.1002/sim.1610.
11. Wu GHM, Auvinen A, Yen AMF, Hakama M, Walter SD, Chen HH. A stochastic model for survival of early prostate cancer with adjustments for leadtime, length bias, and over-detection. *Biom J*. 2011;54(1):20-44. https://doi.org/10.1002/bimj.201000107.
12. Taghipour S, Banjevic D, Miller AB, Montgomery N, Jardine AKS, Harvey BJ. Parameter estimates for invasive breast cancer progression in the Canadian national breast screening study. *Br J Cancer*. 2013;108:542-548.
13. Chen THH, Kuo HS, Yen MF, Lai MS, Tabar L, Duffy SW. Estimation of sojourn time in chronic disease screening without data on interval cases. *Biometrics*. 2000;56(1):167-172.
14. Hakama M, Auvinen A, Day N, Miller A. Sensitivity in cancer screening. *J Med Screen*. 2007;14(4):174-177.
15. Satten GA, Longini IM. Markov chains with measurement error: estimating the 'true' course of a marker of the progression of human immunodeficiency virus disease. *J R Stat Soc Ser C Appl Stat*. 1996;45(3):275-309.
16. Jackson C. Multi-state models for panel data: the MSM package for R. *J Stat Softw*. 2011;38(1):1-28. https://doi.org/10.18637/jss.v038.i08.
17. Prorok PC, Andriole GL, Bresalier RS, et al. Design of the prostate, lung, colorectal and ovarian (PLCO) cancer screening trial. *Control Clin Trials*. 2000;21(6, Suppl 1):273S-309S. https://doi.org/10.1016/s0197-2456(00)00098-2.
18. Andriole GL, Crawford ED, Grubb RL, et al. Prostate cancer screening in the randomized prostate, lung, colorectal, and ovarian cancer screening trial: mortality results after 13 years of follow-up. *J Natl Cancer Inst*. 2012;104(2):125-132. https://doi.org/10.1093/jnici/djr500.
19. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Stat Med*. 2005;24(11):1713-1723. https://doi.org/10.1002/sim.2059.
20. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Stat Med*. 2007;26(11):2389-2430. https://doi.org/10.1002/sim.2712.
21. Pinsky PF, Black A, Kramer BS, Miller A, Prorok PC, Berg C. Assessing contamination and compliance in the prostate component of the prostate, lung, colorectal, and ovarian (PLCO) cancer screening trial. *Clin Trials*. 2010;7(4):303-311. https://doi.org/10.1177/1740774510374091.
22. Pinsky P, Miller A, Kramer B, et al. Evidence of a healthy volunteer effect in the prostate, lung, colorectal, and ovarian cancer screening trial. *Am J Epidemiol*. 2007;165(8):874-881. https://doi.org/10.1093/aje/kwk075.
23. Spix C, Berthold F, Hero B, Michaelis J, Schilling F. Correction factors for self-selection when evaluating screening programmes. *J Med Screen*. 2016;23(1):44-49.
24. Keiding N, Andersen P, eds. *Survival and Event History Analysis*. Hoboken, NJ: Wiley; 2006.
25. Wu WYY, Törnberg S, Elfström KM, Liu X, Nyström L, Jonsson H. Overdiagnosis in the population-based organized breast cancer screening program estimated by a non-homogeneous multi-state model: a cohort study using individual data with long-term follow-up. *Breast Cancer Res*. 2018;20(1):153. https://doi.org/10.1186/s13058-018-1082-z.

**How to cite this article:** Bhatt R, van den Hout A, Pashayan N. A multistate survival model of the natural history of cancer using data from screened and unscreened population. *Statistics in Medicine*. 2021;40:3791–3807. https://doi.org/10.1002/sim.8998
APPENDIX A. MODEL HAZARDS AND MISCLASSIFICATION

Figure A1 is the hazard rates for modeled parameters from the PLCO data.

Figure A2 is the probability of misclassification and sensitivity by age in the PLCO trial: The vertical lines indicate the minimum and maximum screening ages in the PLCO trial. The vertical dashed lines specify the age range of screening in the PLCO trial. Estimates of probability of misclassification and sensitivity outside these age ranges should be read with caution.

FIGURE A1  The hazard rates for modeled parameters from the PLCO data. A, Hazard function from State 1; B, Hazard function from State 2

FIGURE A2  Probability of misclassification and sensitivity by age in the PLCO trial: The vertical lines indicate the minimum and maximum screening ages in the PLCO trial
APPENDIX B. SIMULATION STUDY RESULTS OF DIFFERING SAMPLES

A simulation study of Model C; Gompertz hazards on transitions of 1 to 2, 1 to 4 and 2 to 4, and an exponential hazard on the transition from 2 to 3. Model C includes age-varying misclassification.

| Parameters | True value | Mean   | Rel. bias (mean) | Median   | Rel. bias (median) | Average SE | Emp. SE | Number of subjects |
|------------|------------|--------|------------------|----------|--------------------|------------|--------|-------------------|
| $\lambda_{12}$ | -4.00 | -4.10 | -0.02 | -4.03 | -0.01 | 0.94 | 0.68 | 300 |
| $\lambda_{14} = \lambda_{24}$ | -4.50 | -4.48 | 0.00 | -4.47 | 0.01 | 0.30 | 0.28 | 300 |
| $\lambda_{23}$ | -1.50 | -1.61 | -0.07 | -1.62 | -0.08 | 0.96 | 0.95 | 300 |
| $\beta_{12}$ | 0.05 | 0.04 | 0.29 | 0.04 | 0.16 | 0.06 | 0.05 | 300 |
| $\beta_{14} = \beta_{24}$ | 0.10 | 0.10 | -0.04 | 0.10 | -0.03 | 0.04 | 0.04 | 300 |
| $\nu_m$ | -0.50 | -2.56 | -4.13 | -0.75 | -0.50 | 458.99 | 17.86 | 300 |
| $\gamma_m$ | -0.20 | -0.06 | 0.68 | -0.22 | -0.09 | 43.18 | 2.31 | 300 |
| $\lambda_{12}$ | -4.00 | -4.02 | -0.01 | -4.01 | -0.00 | 0.29 | 0.24 | 600 |
| $\lambda_{14} = \lambda_{24}$ | -4.50 | -4.45 | 0.01 | -4.44 | 0.01 | 0.21 | 0.22 | 600 |
| $\lambda_{23}$ | -1.50 | -1.57 | -0.04 | -1.53 | -0.02 | 0.86 | 0.69 | 600 |
| $\beta_{12}$ | 0.05 | 0.04 | 0.13 | 0.04 | 0.10 | 0.03 | 0.02 | 600 |
| $\beta_{14} = \beta_{24}$ | 0.10 | 0.10 | -0.04 | 0.10 | -0.03 | 0.03 | 0.03 | 600 |
| $\nu_m$ | -0.50 | -1.01 | -1.02 | -0.43 | -0.13 | 4.12 | 2.59 | 600 |
| $\gamma_m$ | -0.20 | -0.21 | -0.06 | -0.21 | -0.04 | 0.86 | 0.41 | 600 |
| $\lambda_{12}$ | -4.00 | -3.95 | 0.01 | -3.95 | 0.01 | 0.74 | 0.54 | 900 |
| $\lambda_{14} = \lambda_{24}$ | -4.50 | -4.45 | 0.01 | -4.45 | 0.01 | 0.17 | 0.18 | 900 |
| $\lambda_{23}$ | -1.50 | -1.58 | -0.06 | -1.53 | -0.02 | 0.63 | 0.75 | 900 |
| $\beta_{12}$ | 0.05 | 0.05 | 0.08 | 0.05 | 0.04 | 0.04 | 0.03 | 900 |
| $\beta_{14} = \beta_{24}$ | 0.10 | 0.10 | -0.02 | 0.10 | -0.02 | 0.02 | 0.02 | 900 |
| $\nu_m$ | -0.50 | -0.71 | -0.42 | -0.60 | -0.2 | 2.45 | 1.84 | 900 |
| $\gamma_m$ | -0.20 | -0.20 | 0.02 | -0.20 | -0.02 | 0.47 | 0.25 | 900 |
| $\lambda_{12}$ | -4.00 | -4.00 | -0.00 | -4.00 | -0.00 | 0.19 | 0.18 | 1200 |
| $\lambda_{14} = \lambda_{24}$ | -4.50 | -4.44 | 0.01 | -4.43 | 0.01 | 0.14 | 0.15 | 1200 |
| $\lambda_{23}$ | -1.50 | -1.57 | -0.04 | -1.61 | -0.07 | 0.59 | 0.57 | 1200 |
| $\beta_{12}$ | 0.05 | 0.04 | 0.10 | 0.05 | 0.09 | 0.02 | 0.02 | 1200 |
| $\beta_{14} = \beta_{24}$ | 0.10 | 0.10 | -0.03 | 0.10 | -0.03 | 0.02 | 0.02 | 1200 |
| $\nu_m$ | -0.50 | -0.62 | -0.24 | -0.48 | 0.05 | 1.57 | 1.42 | 1200 |
| $\gamma_m$ | -0.20 | -0.21 | -0.07 | -0.20 | -0.01 | 0.10 | 0.09 | 1200 |