A Hierarchical Inhomogeneous Hidden Markov Model for Cancer Screening Modeling

Rui Meng*
Department of Statistics
University of California
Santa Cruz, CA 95060
rmeng1@ucsc.edu

Braden Soper
Lawrence Livermore National Laboratory
Livermore, CA
soper3@llnl.gov

Jan F. Nygard
Cancer Registry of Norway
Oslo, Norway
jan.nygard@kreftregisteret.no

Mari Nygard
Cancer Registry of Norway
Oslo, Norway
mari.nygard@kreftregisteret.no

Herbert Lee
Department of Statistics
University of California
Santa Cruz, CA 95060
herbie@ucsc.edu

Abstract

Continuous-Time Hidden Markov Models are an attractive approach for modeling clinical disease progression data because they easily handle both irregularly sampled and noisy data. Most applications in this context consider time-homogeneous models due to their relative computational simplicity. However, the time homogeneous assumption is too strong to accurately model the natural history of many diseases. Moreover, the population at risk is not homogeneous either, since disease exposure and susceptibility can vary considerably.

In this paper we propose a Hierarchical Continuous-Time Inhomogeneous Hidden Markov Model to explain heterogeneity in both time and the population. We present an efficient, scalable EM algorithm for inference in which we introduce prior distributions to deal with imbalanced data. Moreover, automatic differentiation is used to speed up optimization in the M-step. Finally, The model and inference method are applied to a true population-level, cervical cancer screening dataset from the Cancer Registry of Norway. Data used in the analyses will be available on request from the Cancer Registry of Norway, given legal basis according to the GDPR.

Experiments illustrate that our model outperforms some recent state-of-the-art recurrent neural network models on prediction accuracy, also significantly performs better than a standard Continuous-Time Hidden Markov Model based on Kaplan-Meier estimators.

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

Preprint. Under review.

1 Introduction

Population-based screening programs for identifying undiagnosed individuals have a long history in improving public health. Examples include screening programs for cancer (e.g. cervical, breast, colon), tuberculosis and fetal abnormalities.

While the primary objective of such programs is to identify and treat undiagnosed individuals, these cancer screening programs and the population-level, longitudinal datasets associated with them, present many opportunities for the data-driven, computational sciences. In conjunction with modern analytic and computational techniques, such data has the potential to yield novel insights into the natural history of diseases as well as improving the effectiveness of the screening programmes.

Hidden Markov Models (HMM) are a standard choice for disease progression modeling for at least three reasons. First, the underlying disease is represented as an unobserved, latent Markov process. Second, noisy measurements of the disease states are efficiently incorporated as conditional probability distributions in the emission mechanism. Third, any modeling assumptions for a particular application are easily incorporated into the transition probability matrix and emission mechanism.

However, standard HMMs assume that measurements are regularly sampled at discrete intervals which is often not the case in disease screening programs. Measurements are often irregularly sampled because patients come in for screenings at irregular intervals, even if regular screening tests are recommended. To deal with irregular sampling Continuous-Time Hidden Markov Models (CTHMM) are used often used since they easily handle samples taken at arbitrary time intervals. CTHMM have been proposed in many applications such as networks [42], medicine [7], seismology [30] and finance [21]. Liu summarizes current inference methods for CTHMMs in [28] and proposes efficient EM-based learning approaches. However, these methods only apply to time-homogeneous models.

Because the natural history of many diseases depends heavily on the age of the individual, the time-homogeneous assumption is not valid. For this reason, time-inhomogeneous HMMs are more appropriate. Although such models have many appealing theoretical properties according to the Kolmogorov equations [44], parameter inference is intractable in most non-trivial cases. For this reason, many inference studies of continuous-time, time-inhomogeneous HMMs (CTIHMMs) in the medical domain depend on inefficient microsimulations [36, 34, 8]

Finally, we point out that many previous HMM models of disease progression assume that the observations come from a homogeneous population. In large populations, this will typically not be the case. For example, in population-level screening data a large proportion of individuals have benign test results while only a small proportion have abnormal test results. This suggests that it is crucial to consider population heterogeneity when modeling disease progression at the population level.

To deal with these difficulties we introduce piece-wise constant transition intensity functions which allows for tractable parameter inference yet is considerably more flexible in terms of time-inhomogeneity. We then propose using a latent clustering method (i.e. frailty model) to capture unobserved population heterogeneity in terms of disease exposure and susceptibility. Specifically, we propose a new hierarchical hidden Markov model for disease progression in which patients are categorized into classes based on risk levels. With a standard EM algorithm inference is still expensive. Therefore, we propose an efficient and scalable EM algorithm combining both soft and hard assignment in the E-step and an auto-differentiation based Limited-memory BFGS optimization method in M-step.

We apply this model to cervical cancer screening data from the Cancer Registry of Norway. This is a true population-level dataset with over 1.7 million women and more than 10 million screening results. Based on the cervical cancer screening data, our model is illustrated to have a better predictive accuracy compared with state of the art models, base on AUC (Area Under The Curve) under binary classification framework. Moreover, our model is significantly better than a simple hidden Markov model by comparing model-generated Kaplan-Meier curves with observed Kaplan-Meier curves.

2 Related Work

Longitudinal observation data are widely existed, especially in healthcare area. [35] proposed multiple self-controlled case series to model the multiple drug exposure based on conditional poisson regression. [2] extend it from discrete time to continuous time using Hawkes process modeling. Moreover, [23] propose baseline regularization to leverage the diverse health profiles for adverse drug events. It is also a generalized linear model extended from [22].

As for screening test, health status are of interest. It is crucial to consider a latent model of health status. Hidden Markov model is state of that art approach. Most of hidden Markov model variants consider the discrete time [14, 5, 38]. Continuous time hidden Markov models can handle data at any time stamp [11] and therefore are suitable for irregularly-sampled longitudinal data [3, 27, 41]. Furthermore, [28] summarize and discuss learning approach for continuous time hidden Markov model and proposes efficient EM-based learning approaches. Since screening process significantly depends on patient's age, our model is based on CTIHMM, which is discussed in [36, 34, 8].

As a sequence model, recurrent neural network is able to deal with variable-length of time series and capture the temporal correlation. Various variants of RNN are proposed to better balance the memory and new features. [16, 9] propose the concept of Long Short-Term Memory (LSTM) architecture. Several variants are utilized in handwritting recognition [12] , language modeling [32], video data [45]. [10] propose gated recurrent neural network (GRU) as another state of the art RNN model.

3 Model

We model disease progression based on longitudinal, clinical screening results. First, since different patients should have different disease risks, we assume all patients are assigned to different clusters associated with distinct disease exposure risk. In particular we assume all patients come from two risk categories: High disease exposure risk and low disease exposure risk. A hierarchical model is used to assign patients to suitable categories based on the patients' latent frailty [43]. Prior information is incorporated into the hierarchical structure to get more accurate estimates when few screening test results are available.

Second, we assume distinct Markov models for disease progression in the different latent categories. The intuition is that different subsets of the population have different exposure and susceptibility risks, thus the disease progression should be modeled differently in each subpopulation. Thus, we utilize different transition mechanisms for individual models, but all models share the same state space. The screening process is assumed to be imperfect and subject to noise. Thus screening data is assumed to come from a hidden Markov model. The emission mechanism is assumed to be the same across the different Markov models.

Finally, to handle the irregular sampling of screening test data and the heterogeneity of the Markov transition function, a CTIHMM is used.

The resulting model is a hierarchical, continuous-time, time-inhomogeneous HMM, which we call a hierarchical inhomogeneous HMM (HIHMM). We apply the model to a population-level cervical cancer screening dataset. To model the underlying cervical cancer disease process and screening process, we generalize the CTIHMM presented in [37]. The progression of cervical cancer can roughly be described by progression through four main states: normal epithelium, low-grade dysplasia, high-grade dysplasia, and cancer. Human Papilloma Virus is the causal agent which initiates and drives the progress through the stages, increasing the chance of a woman with an HPV infection to progress to cervical cancer. Low-grade dysplasia might result from conditions other than HPV, while high-grade dysplasia is caused by HPV.

There are three different types of screening tests to detect the underlying disease states: cytology, histology, and a test for detecting human papilloma virus (HPV). Results of cytology or histology have four levels, denoted as $0, 1, 2$ and 3. corresponding to the underlying disease states, while the results of the HPV test have two levels, positive or negative. Infection with the HPV virus is known to initiate and sustain the diseases process, thus a positive HPV test increases the likelihood of an abnormal disease state. During the screening process some women may be treated. As such our model assumes that if a woman gets treated then the next immediate state will be normal. Considering the heterogeneity of the population, we assume two Hidden Markov Models $\mathcal{M}_0$ and $\mathcal{M}_1$, with structures

Table 1: Notation

N	Number of individuals
Z	Number of latent frailty states
K	Number of test categories
M	Number of disease states
$\boldsymbol{M}$	Number of disease states for different models
$\boldsymbol{L}$	Number of test levels
$\boldsymbol{z}$	Index of model assignment
$\boldsymbol{S}$	Disease states
$\boldsymbol{E}$	Number of screening tests
$\boldsymbol{G}$	Screening test results

shown in Figure 1, to describe the low-risk and high-risk populations respectively. The low-risk model assumes only three states (normal, low-grade and death) while high-risk model assumes five states (normal, low-grade, high-grade, cancer and death). Because cancer is a relatively rare disease, the majority of screened individuals will have normal test results. Thus this frailty HMM is designed to better fit such zero-inflated data.

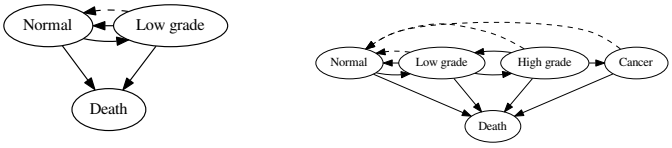

Figure 1: Transition structure of model $\mathcal{M}_0$ and $\mathcal{M}_1$. Solid lines denote the intensity transition while dashed lines denote that any state comes back to the normal state once treatment is completed.

The hierarchical structure of the HIHMM allows for an arbitrary number of latent frailty states, provided relevant Markov models can be ascribed to the disease progression associated with each. The rest of this section describes this general HIHMM in detail.

3.1 Variables

Suppose there are N individuals in the screening population. Let individual n have T_n screening visits at ages $a_1, \ldots, a_{T_n}$. We assume Z categories are considered in the hierarchical frailty structure. We introduce the following variables for our model:

$$\begin{aligned} \text{Frailty State (hidden):} \quad & z_n \in \{1, \ldots, Z\} \\ \text{Disease States (hidden):} \quad & S_{n,t} \in \{1, \ldots, M_{z_n}\} \\ \text{Number of screening tests (observable):} \quad & E_{n,t,k} \in \mathbb{N} \\ \text{Screening test results (observable):} \quad & G_{n,t,k} \in \mathbb{N}^{L_k} \end{aligned}$$

The underlying disease state of individual n is assumed to evolve according to a continuous-time, time-inhomogeneous Markov process, but we only observe screening results at specific time stamps with corresponding ages $a_1, \ldots a_{T_n}$. For individual n, $\boldsymbol{z}_n$ is an indicator variable associated with their latent frailty class, and $\boldsymbol{S}_{n,t}$ refers to the latent disease state on the tth screening visit. Each screening visit includes $\boldsymbol{E}_{n,t,k}$ tests of the kth test type, and the corresponding results are denoted as $\boldsymbol{G}_{n,t,k}$. For notational convenience, we assume $M_z = M$ for $z = 1, \ldots Z$.

3.2 Model of Disease Progression

This underlying Markovian disease process is parameterized by an $M \times M$ transition intensity matrix Q with the ijth element q_{ij} satisfying $q_{ij} \geq 0$ for $i \neq j$ and $q_{ii} = -\sum_{i \neq j} q_{ij}$. The time spent in state i is exponentially distributed with rate $-q_{ii}$. Given that a transition occurs from state i, the probability of transitioning to state j is $\frac{q_{ij}}{q_i}$ where $q_i = \sum_{i \neq j} q_{ij}$.

When performing inference of Markov processes, the matrix Q plays a critical role. We briefly discuss the likelihood functions associated with continuous-time HMMs. In the following sections assume there are M states and one individual has $T+1$ observations $\boldsymbol{O} = (O_0, O_1, \ldots, O_T)$. Furthermore, assume that the underlying transition time are $\tilde{\boldsymbol{t}} = (\tilde{t}_0, \tilde{t}_1, \ldots, \tilde{t}_{\tilde{T}-1})$, where $\tilde{t}_0$ is the initial time.

3.2.1 Continuous-time, Homogeneous Hidden Markov Model

In continuous-time, homogeneous Markov processes, the transition intensity matrix is invariant: $Q(t) \equiv Q$. For simplicity of expression, we assume the initial state is known, $p(S(\tilde{t}_0)) = 1$. Then complete likelihood (CL) for the observed continuous-time, homogeneous hidden Markov model is given by

$$\begin{aligned} \text{CL} &= \prod_{i=0}^{T'-1} (q_{S(t'_i),S(t'_{i+1})}/q_{S(t'_i)}) q_{S(t'_i)} e^{-q_{S(t'_i)}\tilde{\Delta}_i} \prod_{j=0}^{T} p(O_j|S(t_j)) \\ &= \prod_{i=1}^{M} \left(e^{-q_i\tau_i} \prod_{j\neq i} q_{ij}^{n_{ij}} \right) \prod_{j=0}^{T} p(O_j|S(t_j)) \,, \end{aligned} \tag{1}$$

where $\tilde{\Delta}_i = \tilde{t}_{i+1} - \tilde{t}_i$ and n_{ij} denotes the number of times the state changes from state i to state j during the whole process and τ_i denotes the duration that the process is in state i.

In disease screening data the process is observed at the irregular timestamps $\boldsymbol{t} = (t_0, t_1, \ldots, t_{T-1})$, not the underlying transition timestamps $\tilde{\boldsymbol{t}}$. One common approach for inference in this case is to marginalize out the underlying latent disease states. The marginalized complete likelihood (MCL) is given by

$$\text{MCL} = \prod_{i=0}^{T-1} P(\triangle'_i)_{S(t_i),S(t_{i+1})} \prod_{j=0}^{T} p(O_j|S(t_j)) \,,$$

where $\triangle_i = t_{i+1} - t_i$ and $P(\triangle_i) = e^{Q\triangle_i}$ is the transition probability matrix from time t_i to time t_{i+1}.

3.2.2 Continuous-time, Inhomogeneous Hidden Markov Model

Continuous-time, Inhomogeneous Markov processes drop the invariance assumption of the transition intensity matrix Q in the homogeneous case, by assuming that every $q_{ij}(t)$ is a function of t. For the observed continuous-time, inhomogeneous hidden Markov model, deriving the corresponding complete likelihood is intractable, because the time spent in state i does not follow an exponential distribution. To perform inference on CTIHMMs, an alternative approach is to consider the corresponding MCL. The only difference in MCLs between the inhomogeneous model and the homogeneous model is the computation of the transition matrix $P([t_i, t_{i+1}])$ from time t_i to time t_{i+1} for $i = 0, \ldots, T-1$. For the inhomogeneous model, $P([t_i, t_{i+1}]) = e^{\int_{t_i}^{t_{i+1}} Q(t)dt}$.

In general, the matrix exponential $e^{\int_{t_i}^{t_{i+1}} Q(t)dt}$ has no analytical expression. Even though we can take advantage of numerical computational methods, its computation is prohibitively expensive. To ease this computational burden, we propose a piecewise constant transition intensity matrix Q in which each element q_{ij} is a piecewise constant function of time. Specifically, we partition time into I disjoint intervals covering the range of observable time. We then have a set of disjoint partitions $\mathcal{A} = \{A_i\}_{i=1}^{I}$. Each transition intensity function q_{ij} is then a piecewise constant function via the defined partition $\mathcal{A}$, denoted by $q_{ij}(t) = \sum_{k=1}^{I} q_{ijk} \mathbf{1}_{A_k}(t)$, where $\mathbf{1}_A(t)$ is a Kronecker delta function ($\mathbf{1}_A(t) = 1$ if $t \in A$ otherwise $\mathbf{1}_A(t) = 0$) and $q_{ij\ell} \geq 0$.

Under this simplified scenario, the corresponding MCL has a concise and tractable form. Locally, the continuous-time inhomogeneous Markov process is treated as a combination of several continuous-time homogeneous Markov processes. Specifically, given a partition $\mathcal{A}$, the transition probability matrix Q from time t_i to t_{i+1} is simplified as follows. Without loss of generality, we assume $t_i \in A_j = (a_j, a_{j+1})$ and $t_{i+1} \in A_k = (a_k, a_{k+1})$, where $j < k$. Then the transition matrix from t_i to t_{i+1} is directly expressed as $P(t_i, t_{i+1}) = e^{\sum_{\ell=j}^{k} s_\ell Q_\ell}$, where $s_\ell = a_{\ell+1} - a_\ell$, except for the two ends $s_j = a_{j+1} - t_i$ and $s_k = t_{i+1} - a_k$.

3.3 Hierarchical Inhomogeneous Hidden Markov Model

The populations in which screening programs operate are naturally heterogeneous in many ways. For example, disease exposure and susceptibility can vary considerably based on age, gender, genetics or location. When the explanatory variables for such population heterogeneity are not known or missing from the data, frailty models are a common methodology in epidemiological modeling [43]. For this reason, we propose a hierarchical structure to explain any possible population heterogeneity related to disease exposure risk.

In the context of the hierarchical continuous-time inhomogeneous hidden Markov model, we denote the set of all observations in sequence n by $\boldsymbol{O}_n = [\boldsymbol{E}_n, \boldsymbol{G}_n]$ and suppose there are Z different CTIHMMs under the hierarchical structure. Let $\boldsymbol{\psi}$ denote all parameters and let $\boldsymbol{\psi}_z$ denote all parameters for model $\mathcal{M}_z$. Then the hierarchical model is given by

$$\begin{aligned} \boldsymbol{O}_n &\sim \mathcal{M}_{z_n}(\boldsymbol{\psi}_{z_n}, \boldsymbol{\theta}_n) \\ z_n &\sim \mathrm{Cat}(\boldsymbol{p})\,, \end{aligned}$$

where $\boldsymbol{\theta}_n$ denote all covariates for the nth individual. Also, an informative prior of the model indicator z_n is a categorical distribution with parameters $\boldsymbol{p}$, used to provide expert knowledge of the model assignment. This prior contributes to reasonable model inference, especially when modeling screening data which is highly imbalanced in terms of latent class membership. Figure 1 shows the case where $Z = 2$ and the index z_n has a Bernoulli prior with parameter p, i.e. $z_n \sim \mathrm{Ber}(p)$.

4 Inference

Two approaches are accessible for the maximum likelihood estimation of CTHMMs. One is to maximize the complete likelihood (CL) and the other is to maximize the marginal complete likelihood (MCL). Given the CL expression (1) and current estimate of the transition intensity matrix $\tilde{Q}$, the expected complete log-likelihood (ECLL) is

$$\begin{aligned} \mathrm{ECLL} &= E\left(\sum_{i=1}^{M}\left(-q_i\tau_i + \sum_{j\neq i} n_{ij}\log(q_{ij})\right)\right) + E\left(\sum_{j=0}^{T}\log p(O_j|S(t_j))|\tilde{Q}, \boldsymbol{O}\right) \\ &= \sum_{i=1}^{M}\left(-q_i E(\tau_i|\tilde{Q}, \boldsymbol{O}) + \sum_{j\neq i} E(n_{ij}|\tilde{Q}, \boldsymbol{O})\log(q_{ij})\right) + \sum_{j=0}^{T} E\left(\log p(O_j|S(t_j))|\tilde{Q}, \boldsymbol{O}\right)\,, \end{aligned}$$

where $E(\log p(O_j|S(t_j))|\tilde{Q}, \boldsymbol{O}) = E(E\left(\log p(O_j|S(t_j))|S(t_j)\right)|\tilde{Q}, \boldsymbol{O})$. Therefore, the computation focuses on three summarized statistics $\{E(\tau_i|\tilde{Q}, \boldsymbol{O})\}$, $\{E(n_{ij}|\tilde{Q}, \boldsymbol{O})\}$ and $\{p(S(t_j)|\tilde{Q}, \boldsymbol{O})\}$. By maximizing ECLL with respect Q, the estimate of Q has a closed-form expression as

$$\hat{q}_{ij} = \frac{E(n_{ij}|\tilde{Q}, \boldsymbol{o})}{E(\tau_i|\tilde{Q}, \boldsymbol{o})}\,. \tag{2}$$

This analytic formula is proposed in [6]. Base on (2), Metzner [33] proposed a scalable algorithm on CTHMMs, named Enhanced MLE-method. This approach is feasible for the non-equidistant time-lags scenario, but Q is assumed to be diagonalizable, and $\{E(\tau_i|\tilde{Q}, \boldsymbol{o})\}$ and $\{E(n_{ij}|\tilde{Q}, \boldsymbol{o})\}$ should have a closed-form expression. Since Q is not always diagonalizable in practice, [28] proposed the *Examp* method based on a classical method of Van Loan [29]. On the other hand, Hobolth introduced a *uniformization* approach to reduce the computational burden in [15].

From the other perspective of the MCL, given a current estimate of the transition matrix $\tilde{Q}$, the expected marginal complete log-likelihood is

$$\mathrm{EMCLL} \;=\; E\left(\sum_{i=0}^{T-1}\log(P(\Delta t_{t_i})_{s(t_i)s(t_{i+1})})\right) + E\left(\sum_{j=0}^{T}\log p(O_j|s(t_j))|\tilde{Q}, \boldsymbol{O}\right)\,. \tag{3}$$

Through maximizing (3), it is difficult to get a closed form for $\hat{q}_{ij}$, because the computation of the matrix exponential is complicated. Therefore, numerical optimization approaches are considered to approximate $\hat{q}_{ij}$.

The CTIMP is a generalization of the CTHMP and the computation is more challenging. Even in our piecewise constant case, the complete likelihood is difficult to express. Hence, we propose a scalable EM-based approach, focusing on maximizing its MCL, instead of the CL.

4.1 Scalable EM for HIHMMs

We propose a scalable EM-based method for maximum likelihood estimation of HIHMMs. For each observation sequence $\boldsymbol{O}_n$ there is an associated latent model index z_n and latent state sequence $\boldsymbol{S}_n$. Standard EM approaches maximize the expectation of MCL with respect to the posterior distribution of (z_n and $\boldsymbol{S}_n$). But the computation of this posterior distribution is expensive. We propose a more efficient EM-based approach. First, we decompose the joint posterior distribution by $p(z_n, \boldsymbol{S}_n|-) = p(z_n|-)p(\boldsymbol{S}_n|z_n, -)$. Then we combine both soft assignment and hard assignment in the E-step. A soft assignment approach is used in standard EM algorithms, while the hard assignment approach approximates the posterior distribution in the E-step as a degenerated delta distribution. One well-known application of the hard assignment EM algorithm is the K-means algorithm [17]. Both soft assignment and hard assignment approaches are studied in [18]. Considering their advantages, we use soft assignment for $p(z_n|-)$ and hard assignment for $p(\boldsymbol{S}_n|z_n, -)$, because the computation of $p(\boldsymbol{S}_n|z_n, -)$ is prohibitively expensive. For simplicity, we ignore covariates $\boldsymbol{\theta}_n$ in the notation. The recursive procedures are given as follows.

- Given previous estimates $\boldsymbol{\psi}^{(t-1)}$, compute the conditional posterior distribution of z_n:

$$\begin{aligned} p(z_n|\boldsymbol{O}_n, \boldsymbol{\psi}^{(t-1)}) &\propto \mathrm{Cat}(z_n|\boldsymbol{p})p(\boldsymbol{O}_n|z_n, \boldsymbol{\psi}_{z_n}^{(t-1)}) \\ &\sim \mathrm{Cat}(\tilde{\boldsymbol{p}}_n)\,, \end{aligned} \tag{4}$$

  where $\tilde{p}_{nk} = \frac{p_k p(\boldsymbol{O}_n|z_n=k, \boldsymbol{\psi}_k^{(t-1)})}{\sum_{z=1}^{Z} p_z p(\boldsymbol{O}_n|z_n=z, \boldsymbol{\psi}_z^{(t-1)})}$ for $k = 1, \ldots, Z$ and $p(\boldsymbol{O}_n|z, \boldsymbol{\psi}_z)$ is accessible through the filter-forward backward-sample algorithm (FFBS), which is a sequential Monte Carlo approach first proposed in [20].

- Update the optimal state sequence $\boldsymbol{S}_n$ given corresponding observations $\boldsymbol{O}_n$ and model indicator z using the Viterbi algorithm:

$$\boldsymbol{S}_{nz}^{(t)} = \mathrm{Viterbi}(\boldsymbol{O}_n, \boldsymbol{\psi}_z)\,. \tag{5}$$

- Maximize the EMCLL with respect to $\boldsymbol{\psi}$ by

$$\begin{aligned} \boldsymbol{\psi}^{(t)} &= \arg\max_{\boldsymbol{\psi}} \sum_{n=1}^{N} E_{z_n, \boldsymbol{S}_n}(\ell(\boldsymbol{\psi}|\boldsymbol{O}_n, z_n, \boldsymbol{S}_n)|\boldsymbol{O}_n, \boldsymbol{\psi}^{(t-1)}) \\ &= \arg\max_{\boldsymbol{\psi}} \sum_{n=1}^{N} \sum_{z=1}^{Z} \sum_{\boldsymbol{S}_n} p(z_n = z|\boldsymbol{O}_n, \boldsymbol{\psi}^{(t-1)}) p(\boldsymbol{S}_n|z_n = z, \boldsymbol{O}_n, \boldsymbol{\psi}^{(t-1)}) \ell(\boldsymbol{\psi}|\boldsymbol{O}_n, z, \boldsymbol{S}_n) \\ &= \arg\max_{\boldsymbol{\psi}} \sum_{n=1}^{N} \sum_{z=1}^{Z} p(z_n = z|\boldsymbol{O}_n, \boldsymbol{\psi}^{(t-1)}) \ell(\boldsymbol{\psi}|\boldsymbol{O}_n, z, \boldsymbol{S}_{nz}^{(t)}) \\ &= \arg\max_{\boldsymbol{\psi}} \sum_{n=1}^{N} \sum_{z=0}^{Z} p(z_n = z|\boldsymbol{O}_n, \boldsymbol{\psi}^{(t-1)}) \left(\log p_z + \log p(\boldsymbol{S}_n|z, \boldsymbol{\psi}) + \log p(\boldsymbol{O}_n|\boldsymbol{S}_n, \boldsymbol{\psi})\right)\,. \end{aligned} \tag{6}$$

4.2 Scalable Computation

Population screening datasets can contain millions of records. In this case direct inference may be prohibitively expensive and more scalable approaches are necessary. We scale our EM algorithm by parallelizing the inference across observations using $\tilde{N}$ clusters, $\{C_n\}_{n=1}^{\tilde{N}}$, in three parts. We first compute the conditional posterior distribution of z_n in each cluster using (4). The time complexity for each cluster is $O(|C_n|ZM^2T)$ [19]. We then compute the optimal state sequences in each cluster C_n using (5) with the same time complexity $O(|C_n|ZM^2T)$ [1]. Finally, we compute the gradients

in each cluster then reduce all local gradients to global gradients for the optimization in the M-step. In detail the EMCLL is rewritten as

$$\begin{aligned} & \sum_{n=1}^{N} E_{z_n, \boldsymbol{S}_n}(\ell(\boldsymbol{\psi}|\boldsymbol{O}_n, z_n, \boldsymbol{S}_n)|\boldsymbol{O}_n, \boldsymbol{\psi}^{(t-1)}) \\ = & \sum_{\tilde{n}=1}^{\tilde{N}} \sum_{n \in C_{\tilde{n}}} E_{z_n, \boldsymbol{S}_n}(\ell(\boldsymbol{\psi}|\boldsymbol{O}_n, z_n, \boldsymbol{S}_n)|\boldsymbol{O}_n, \boldsymbol{\psi}^{(t-1)}). \end{aligned}$$

Taking gradient on both sizes, we have

$$\begin{aligned} & \frac{\partial}{\partial \boldsymbol{\psi}} \sum_{n=1}^{N} E_{z_n, \boldsymbol{s}_n}(\ell(\boldsymbol{\psi}|\boldsymbol{O}_n, z_n, \boldsymbol{S}_n)|\boldsymbol{O}_n, \boldsymbol{\psi}^{(t-1)}) \\ = & \sum_{\tilde{n}=1}^{\tilde{N}} \frac{\partial}{\partial \boldsymbol{\psi}} \sum_{n \in C_{\tilde{n}}} E_{z_n, \boldsymbol{S}_n}(\ell(\boldsymbol{\psi}|\boldsymbol{O}_n, z_n, \boldsymbol{S}_n)|\boldsymbol{O}_n, \boldsymbol{\psi}^{(t-1)}). \end{aligned}$$

Automatic differentiation (AD) [4] is utilized to compute the gradients in each cluster. Summing over all clusters, we get the gradient of the EMCLL. Using this gradient we adapt the Limited-Memory BFGS [26] algorithm to estimate $\boldsymbol{\psi}$. The complexity of each cluster in optimization is intractable, but it increase inference speed around $\tilde{N}$ times.

5 Experimental results

The HIHMM is demonstrated on a true population-level cervical cancel screening test dataset from the Cancer Registry of Norway. Data used in the analyses will be available on request from the Cancer Registry of Norway, given legal basis according to the GDPR. We describe the full model and inference method as follows.

5.1 Data and Model Description

In Norway each citizen has a unique personal identification number (PIN) which is registered in a nation-wide, computerized central population registry together with information on vital status, migration, and current address on the individual level. The Norwegian Co-ordinated Cervical Cancer Screening Program (NCCSP) was initiated in 1992 and the Cancer Registry of Norway (CRN) is responsible for the management of the program. All laboratories which analyze exams related to cervical cancer screening in Norway are legally obliged to report the result to the CRN. Data for cervical cancer cases was obtained from the nationwide population-based cancer registry of Norway, established in 1953. Registration of cancer cases and screening events are assessed to be close to 100 % accurate and complete [24, 25]. All relevant information about screening tests and individual vital status were linked by the PIN.

After linkage, the data file was anonymized by applying a multi-step approach: i) All PIN identifiers were replaced with the study specific allocation number (AN) and the key between PIN and AN was deleted; ii) We replaced all dates to the 15th of the month to reduce the risk of re-identification of individuals while preserving the most salient features of the sequence of clinical exams; iii) the dataset was then distorted by adding a non-zero random integer between -4 and 4 to the month in every date [40]. The final dataset available for analysis contains data from 1992 until the end of 2016, with a total of 1,7 million women and 10 million screening test.

The underlying disease process and screening tests were described in section 2 and figure 1. Since only two CTIHMMs are utilized, the hierarchical structure is simplified to

$$\begin{aligned} \boldsymbol{O}_n & \sim \mathcal{M}_{z_n}(\boldsymbol{\psi}_{z_n}, \boldsymbol{\theta}_n) \\ z_n & \sim \mathrm{Ber}(p)\,, \end{aligned}$$

for $n = 1, \ldots, N$. For notation convenience, we ignore the covariates $\boldsymbol{\theta}$ in the reminder of paper. For each individual n, the corresponding posterior distribution of the model index z_n is given by

$$p(z_n|\boldsymbol{O}_n, \boldsymbol{\psi}) \sim \mathrm{Ber}(\tilde{p}_n),$$

where $\tilde{p}_n = \frac{pp(\boldsymbol{O}_n|z_n=1,\boldsymbol{\psi})}{(1-p)p(\boldsymbol{O}_n|z_n=0,\boldsymbol{\psi})+pp(\boldsymbol{O}_n|z_n=1,\boldsymbol{\psi})}$.

The dataset contains censored observations at the last time stamp t_c. Censored observations are denoted by O_c which indicates whether the woman is dead or alive at time t_c. For ease of notation, we ignore the subscript of woman n and then model the initial state S_{z1} under model $\mathcal{M}_z$ as

$$\begin{aligned} S_{z1}|a_1 & \sim \mathrm{Cat}(\boldsymbol{\pi}_z(\mathcal{A}, a_1)) \\ \boldsymbol{\pi}_{zi} & \sim \mathrm{Dir}(\boldsymbol{\alpha}_{zi})\,, \end{aligned}$$

where a_1 denotes the age at the time of the first screening test, $\mathcal{A}$ is a disjoint partition of observable ages, $\boldsymbol{\pi}_z(\mathcal{A}, a) = \boldsymbol{\pi}_{zi}$ if and only if $a \in \mathcal{A}_i$, and $\boldsymbol{\alpha}_{z\ell} \in \mathbb{R}^{+M_z}$.

The observations $\boldsymbol{O}$ have two levels: the number of screening tests $\boldsymbol{E}$ and the results of screening tests $\boldsymbol{G}$. Assume k is the index of test categories and again ignore the subscript n and time index t. Then given state s,

$$\begin{aligned} E_k &\sim \text{Poisson}(\eta_{sk})\,, \\ \boldsymbol{G}_k|E_k &\sim \text{Multinomial}(E_k, \tilde{\boldsymbol{\pi}}_{sk})\,, \\ \tilde{\boldsymbol{\pi}}_{sk} &\sim \text{Dir}(\tilde{\boldsymbol{\alpha}}_{sk})\,, \end{aligned}$$

where E_k denotes the number of screening tests of screening test k. $\boldsymbol{G}_k = \{G_{k0}, \ldots, G_{kL_k}\}$ denotes the numbers of screening test results for all categories of screening test k . Here, $\tilde{\boldsymbol{\alpha}}_{sk} \in \mathbb{R}^{+L_k}$ and L_k denotes the number of different results in the kth screening test category. To ease notation, we assume all observation sequences have the same number of observations denoted by T, but it should be understood that this is not a necessary condition. Then, the censored observation, which is binary (dead or alive) is modeled as

$$p(O_c|S_T) = \begin{cases} P(t_T, t_c)_{S_T,\text{death}} & \text{if } O_c = \text{death} \\ 1 - P(t_T, t_c)_{S_T,\text{death}} & \text{if } O_c \neq \text{death}\,. \end{cases}$$

5.2 Inference with Treatment Information

Because of treatments, the inference has three modifications for FFBS, Viterbi and EMCLL. Without loss of generality we suppose one woman has m treatments indexed by $\{r_1, \ldots, r_m\}$. Then the observation sequence $\boldsymbol{O}$ is partitioned as

$$\{\boldsymbol{O}_1, \ldots, \boldsymbol{O}_{r_1}\}, \ldots, \{\boldsymbol{O}_{r_m}, \ldots \boldsymbol{O}_T, \boldsymbol{O}_c\}.$$

Throughout the FFBS, the marginal likelihood is decomposed as

$$\begin{aligned} p(\boldsymbol{O}|z, \boldsymbol{\psi}) = p(\boldsymbol{O}_1, \ldots, \boldsymbol{O}_{r_1}|z, \boldsymbol{\psi}) \prod_{j=1}^{m-1} p(\boldsymbol{O}_{r_j+1}, \ldots, \boldsymbol{O}_{r_{j+1}}|S_{r_j} = 0, z, \boldsymbol{\psi}) \\ p(\boldsymbol{O}_{r_m+1}, \ldots \boldsymbol{O}_T, \boldsymbol{O}_c|S_{r_m} = 0, z, \boldsymbol{\psi}). \end{aligned} \tag{7}$$

Each component of (7) is tractable using FFBS [20].

A similar decomposition is implemented in the Viterbi algorithm to find the most likely sequence of hidden states.

$$\begin{aligned} (S_1, \ldots, S_{r_1}) &= \text{Viterbi}(\boldsymbol{O}_1, \ldots, \boldsymbol{O}_{r_1}, \boldsymbol{\psi})\,, \\ (S_{r_j+1}, \ldots, S_{r_{j+1}}) &= \text{Viterbi}(\boldsymbol{O}_{r_j+1}, \ldots, \boldsymbol{O}_{r_{j+1}}, \boldsymbol{\psi}|S_{r_j} = 0) \\ & \qquad j = 1, \ldots, m-1\,, \\ (S_{r_m}, \ldots, S_T) &= \text{Viterbi}(\boldsymbol{O}_{r_m+1}, \ldots, \boldsymbol{O}_T, \boldsymbol{O}_c, \boldsymbol{\psi}|S_{r_m} = 0)\,. \end{aligned} \tag{8}$$

According to (6), to compute EMCLL we only need to compute $p(\boldsymbol{S}|z, \boldsymbol{\psi})$. Using a similar decomposition we arrive at

$$p(\boldsymbol{S}|z, \boldsymbol{\psi}) = p(S_1|z, \boldsymbol{\psi}) \prod_{i \in \{r_j\}} p(S_{i+1}|S_i = 0) \prod_{i \notin \{r_j\}, i \neq 1} p(S_{i+1}|S_i).$$

5.3 Hyper-parameter settings

This section describes the hyper-parameter settings of this application. We choose the age partition $\mathcal{A}$ as $[0, 23)$, $[23, 30)$, $[30, 60)$ and $[60, \infty)$ based on information of HPV results. The details are shown in the appendix.

5.4 Model Comparison

We randomly select 80000 patients' records for training and select other 20000 patients' records for testing. The goal is to predict the last visiting status and the status is defined based on the observations at the last visit. Specifically, if a patient has at least one result whose level is greater than 1, then the status is defined as high risk denoted as 1. Otherwise, the status is defined as low risk denoted as 0. Thus, the problem is defined as a binary classification problem.

The prediction procedure is defined as

- Train HIHMM to obtain model parameter estimates $\hat{\boldsymbol{\psi}}$.
- Given new patient historical records $\boldsymbol{O}^* = (\boldsymbol{O}_1^*, \ldots, \boldsymbol{O}_{T-1}^*)$, compute the predictive distribution of model index

$$\begin{aligned} p(z^*|\boldsymbol{O}^*, \hat{\boldsymbol{\psi}}) &\propto \mathrm{Cat}(z^*|\boldsymbol{p}) p(\boldsymbol{O}^*|z^*, \hat{\boldsymbol{\psi}}) \\ &\sim \mathrm{Cat}(\boldsymbol{p}^*) \end{aligned} \tag{9}$$

- Given model index z, compute predictive distribution of state at the last second visiting $p(S_{T-1}^*|z, \boldsymbol{O}^*, \hat{\boldsymbol{\psi}})$ derived from FFBS.
- Compute predictive distribution of state of the last visiting is

$$\begin{aligned} p(S_T^*|\boldsymbol{O}^*, \hat{\boldsymbol{\psi}}) &= \sum_z p(z^* = z|\boldsymbol{O}^*, \hat{\boldsymbol{\psi}}) p(S_T^*|z, \boldsymbol{O}^*, \hat{\boldsymbol{\psi}}) \\ &= \sum_z p_z^* p(\boldsymbol{S}_T^*|z, \boldsymbol{O}^*, \hat{\boldsymbol{\psi}}) \end{aligned} \tag{10}$$

  where

$$p(\boldsymbol{S}_T^*|z, \boldsymbol{O}^*, \hat{\boldsymbol{\psi}}) = \sum_s p(S_T^*|S_{T-1}^* = s) p(S_{T-1}^* = s|z, \boldsymbol{O}^*, \hat{\boldsymbol{\psi}}) \tag{11}$$

- Estimate the predictive distribution of screening test results:

$$p(\boldsymbol{G}_T^*|\boldsymbol{O}^*, \boldsymbol{E}_T^*, \hat{\boldsymbol{\psi}}) = \sum_s p(S_T^* = s|\boldsymbol{O}^*, \hat{\boldsymbol{\psi}}) p(\boldsymbol{G}_T^*|S_T^* = s, \boldsymbol{E}_T^*, \hat{\boldsymbol{\psi}})$$

- Estimate the predictive distribution of the last status:

$$G^* \sim \mathrm{Ber}\,(p^*) \tag{12}$$

  where $p^* = p\left(\sum_{i=0}^{1} \sum_{j=2}^{3} \boldsymbol{G}_T^*[i,j] \geq = 1|\boldsymbol{O}^*, \boldsymbol{E}_T^*, \hat{\boldsymbol{\psi}}\right)$.

- Estimate the last status by $\hat{G^*} = \begin{cases} 1 & p^* \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$.

As for state of the art recurrent neural network models, each patient's record is modeled as one time-series and the features at each visit includes patient's age, patient's screening result and patient's treatment indicator. Specifically, screening result of patient n at tth visiting is $\vec{\boldsymbol{G}_{n,t}}$. And treatment indicator is a binary number, it is equal to 1 if and only if the patient has accepted treatment. LSTM [9], stacked LSTM [13] and GRU [10] are implemented for model comparison. As for stacked LSTM, two LSTMs are stacked. We summarize prediction results in Table 2. It shows our model outperforms state of the art methods in term of Area Under The Curve(AUC). And HIHMM with prior setting $p = 0.2$ obtains the best prediction performance.

5.5 Model Validation

We set model index prior $p = 0.2$ based on model comparison results and expert opinion. And we present two types of results on population-level data. First we present the MLEs for all model parameters along with bootstrapped standard deviations. Second we perform model validation using Kaplan-Meier estimators as suggested in [39].

Table 2: Model prediction for the status of the last visit in terms of Accuracy (ACC), Area Under The Curve(AUC), True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN).

Method	Prior	ACC	AUC	TP	FP	TN	FN
LSTM		0.9920	0.85	43	147	19779	13
stacked LSTM		0.9914	0.86	39	151	19788	22
GRU		0.9922	0.85	59	131	19786	24
HIHMM	0.2	0.9916	0.9098	73	117	19759	51
HIHMM	0.3	0.9916	0.9095	56	134	19776	34
HIHMM	0.4	0.9918	0.8997	53	137	19782	28
HIHMM	0.5	0.9917	0.8995	53	137	19781	29

5.5.1 Parameter Estimation

We randomly divide all data into clusters such that each cluster has 100 individual observation sequences. Using a bootstrap technique, we randomly select 2400 clusters with replacements for model inference. We independently repeat the same inference on different selections 5 times. The mean and standard deviation of all parameter estimates are given in the following tables.

Table 3 shows more than 75% of women have a posterior probability of belonging to the high-risk model $\mathcal{M}_1$ that is less than the prior probability $p = 0.2$. This suggests that more than 75% are likely to be in the low-risk disease exposure category according to their screening test results. Moreover, it also justifies the expert knowledge that around 20% belong to the high-risk disease exposure category. Whereas, the 90% credible interval of the posterior hyper-parameter $\tilde{p}$ is $(0.1825, 0.2358)$, which is still close to the prior hyper-parameter $p = 0.2$. This implies that the observations do not affect the posterior of the model indexes significantly and choosing a reasonable prior is important. This is likely an artifact of the dataset, which is highly skewed towards normal test results. More balanced datasets may exhibit less sensitivity to prior specification.

Table 3: Maximum likelihood estimates of quantiles of the posterior probability for Model 1.

quantiles	0.05	0.25	0.5	0.75	0.95
$\tilde{p}$	0.1825(0.0004)	0.1936(0.0001)	0.1966(0.0002)	0.1988(0.0001)	0.2358(0.0049)

In Table 4, the estimates of diagnostic test results match the definition of states. The more advanced a patient's disease state, the more likely she is to get an abnormal screening result. The small standard deviations suggest that our data is sufficient to get precise estimates of the emission parameters.

Results from Table 5 show the number of screening tests for women in different states. Due to the fact that the expectation of a Poisson distribution is exactly the Poisson intensity parameter, the Poisson intensities shows that individuals at normal state are more likely to be assigned to a cytology screening test. Individuals in abnormal disease states (low-grade, high-grade and cancer) are more likely to be given histology and HPV tests. This result matches what is expected in clinical practice in that women will be assigned more precise screening tests as they present more severe symptoms.

Table 6 shows the initial information of the population categorized by the specified age partition for model $\mathcal{M}_0$ and model $\mathcal{M}_1$. From a model specification perspective, women in the low-risk model, $\mathcal{M}_0$, are assumed to be more likely to stay at a normal state than women in the high-risk model $\mathcal{M}_1$, regardless of their ages. On the other hand, in the high-risk model, $\mathcal{M}_1$, women in the age interval $(23, 30)$ are mostly like to belong to a high risk state at the initial screening test.

Table 7 displays the estimates of transition intensities in the two models. It shows patients are more likely to transition from the low grade state to the normal state, whether or not they are in model $\mathcal{M}_0$ or model $\mathcal{M}_1$. On the other hand, λ_{34} has significantly higher standard deviations than other intensity parameters because of the scarcity of data for individuals with cancer who died during the period in which the data was collected.

5.5.2 Model Validation

For model validation we randomly select 2400 clusters of data in which each cluster has 100 individual sequences of observations. We implement both the HIHMM and the CTIHMM for the

Table 4: Maximum likelihood estimates of diagnostic test result probabilities conditioned on hidden state.

cytology				
state	0	1	2	3
normal	1.0000(0.0000)	0.0000(0.0000)	0.0000(0.0000)	0.0000(0.0000)
low grade	0.0331(0.0011)	0.8021(0.0020)	0.1631(0.0019)	0.0016(0.0001)
high grade	0.0614(0.0068)	0.0054(0.0010)	0.9198(0.0059)	0.0133(0.0012)
cancer	0.0619(0.0125)	0.0567(0.0084)	0.6254(0.0158)	0.2560(0.0097)

histology				
state	0	1	2	3
normal	0.9879(0.0004)	0.0108(0.0004)	0.0007(0.0001)	0.0006(0.0001)
low grade	0.2621(0.0023)	0.1561(0.0006)	0.5810(0.0030)	00009(0.0002)
high grade	0.0345(0.0018)	0.0068(0.0015)	0.9573(0.0026)	0.0014(0.0001)
cancer	0.0263(0.0062)	0.0133(0.0015)	0.0796(0.0014)	0.8808(0.0066)

HPV		
state	-	+
normal	0.9966(0.0010)	0.0034(0.0010)
low grade	0.3720(0.0048)	0.6280(0.0048)
high grade	0.0357(0.0054)	0.9643(0.0054)
cancer	0.0194(0.0019)	0.9806(0.0019)

Table 5: Maximum likelihood estimates of Poisson intensities for the number of tests conditioned on true state.

state	cytology	histology	HPV
normal	0.9880(0.0002)	0.0140(0.0001)	0.0053(0.0001)
low grade	0.7912(0.0008)	0.1856(0.0018)	0.1003(0.0016)
high grade	0.4595(0.0024)	0.6493(0.0018)	0.0290(0.0017)
cancer	0.5091(0.0191)	0.8627(0.0322)	0.0278(0.0023)

Table 6: Maximum likelihood estimates of the probability of being a particular state at the time of the first screening.

age range	16-23	23-30	30-60	60-
Model 0				
normal	0.9315(0.0010)	0.9415(0.0017)	0.9614(0.0007)	0.9643(0.0009)
low grade	0.0685(0.0010)	0.0585(0.0017)	0.0386(0.0007)	0.0357(0.0009)
Model 1				
normal	0.9187(0.0015)	0.9040(0.0011)	0.9287(0.0006)	0.9308(0.0020)
low grade	0.0761(0.0013)	0.0677(0.0022)	0.0438(0.0011)	0.0354(0.0016)
high grade	0.0041(0.0003)	0.0271(0.015)	0.0262(0.0007)	0.0263(0.0018)
cancer	0.0011(0.0001)	0.0013(0.0002)	0.0013(0.0002)	0.0075(0.0006)

same dataset. We follow the method proposed in [39] that utilizes Kaplan-Meier estimators to validate continuous-time HMMs. Kaplan-Meier estimators are defined according to the definition of a failure, or time-to-event. In multi-state models different failures can be defined depending on which features of the model and data are of interest. Here we define failure as the first observation of a high-risk or cancer test result directly following an initial normal or low-grade test result. Accurately predicting this time-to-event is of practical importance because clinical intervention is only possible in the high-grade state. Treating patients at this stage is critical to preventing precancerous lesions from progressing to cervical cancer.

Table 7: Maximum likelihood estimates for age dependent transition intensities.

age range	16-23	23-30	30-60	60-
Model 0				
λ_{01}	0.1718(0.0061)	0.0809(0.0017)	0.0546(0.0006)	0.0439(0.0015)
λ_{02}	0.0005(0.0000)	0.0018(0.0000)	0.0019(0.0000)	0.0147(0.0001)
λ_{10}	1.7064(0.0327)	1.2637(0.0292)	0.4893(0.0118)	2.2169(0.0641)
λ_{12}	0.0024(0.0002)	0.0021(0.0001)	0.0011(0.0001)	0.0122(0.0013)
Model 1				
λ_{01}	0.1938(0.0065)	0.1191(0.0032)	0.0730(0.0011)	0.0536(0.0020)
λ_{04}	0.0014(0.0001)	0.0015(0.0001)	0.0015(0.0001)	0.0121(0.0002)
λ_{10}	1.6854(0.0295)	1.3541(0.0358)	1.6331(0.0131)	2.3063(0.1288)
λ_{12}	0.0815(0.0086)	0.2276(0.0015)	0.1867(0.0030)	0.2395(0.0090)
λ_{14}	0.0058(0.0003)	0.0048(0.0002)	0.0032(0.0003)	0.0150(0.0006)
λ_{21}	0.3585(0.0593)	0.0780(0.0067)	0.0663(0.0052)	0.2245(0.0288)
λ_{23}	0.0720(0.0095)	0.0307(0.0017)	0.1012(0.0060)	0.5166(0.0366)
λ_{24}	0.0150(0.0004)	0.0069(0.0007)	0.0034(0.0003)	0.0166(0.0006)
λ_{34}	1.0366(0.0611)	2.5642(0.5100)	2.6911(0.1251)	1.6805(0.3432)

The empirical Kaplan-Meier estimator is defined as

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right),$$

where t_i a time when at least one failure is observed, d_i is the number of failures that occurred at time t_i, and n_i is the number of individuals known to have survive up to time t_i. We randomly choose 24000 records to generate an empirical Kaplan-Meier estimator according to our definition of a failure. Furthermore, we generate Kaplan-Meier estimates by simulating 100 sequences from both the CTIHMM and HIHMM and repeat this 100 times. Figure 2 shows the empirical Kaplan-Meier curve in black, simulated Kaplan-Meier curves from the CTIHMM in blue, and simulated Kaplan-Meier curves from the HIHMM. As for the simulated Kaplan-Meier curves, solid lines denote the median curve and dashed lines denote the 95% credible intervals based on the 100 replications. The results show that the empirical Kaplan-Meier curve is always near the median and within the 95% credible intervals generated by the HIHMM. This is not the case with the CTIHMM. In this sense the HIHMM outperforms the CTIHMM in an important clinical metric.

On the other hand, the HIHMM has a relatively high Kaplan-Meier estimate at time 0 because the informative prior $p = 0.2$ is relatively small. This has the effect of driving simulated patients to more likely be in the low-risk model $\mathcal{M}$, at the initial time. Moreover, these patients are more likely to stay at the normal state for longer. However, the trend of the median curve from the HIHMM more closely tracks that of the empirical Kaplan-Meier curve, compared with the trend of the median curve from the CTIHMM. This suggests that the HIHMM models disease progression better than the CTIHMM. The Kaplan-Meier curves simulated from the CTIHMM are always underestimated.

6 Discussion and Conclusion

One of the possible applications of the HIHMM in the context of population-based screening programs is risk stratification of the population. The latent random variable z_n is an indicator of belonging to a frail class in the population. Given the learned model parameters $\boldsymbol{\psi}$ it is possible to compute the posterior probability of belonging to the frailty class for individual women. In other words, given an observed sequence of test results $\boldsymbol{O}_n$ and model parameters $\boldsymbol{\psi}$, the posterior predictive distribution $p(z_n|\psi, \boldsymbol{O}_n)$ is of interest. This parameter gives a measure of the likelihood of an individual to be at risk of developing cervical cancer conditioned on their observed test results. Such information could be used to more efficiently screen a population by avoiding the over screening of women at low-risk and the under screening of women at high-risk.

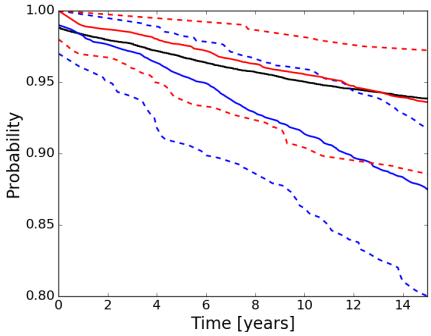

Figure 2: Empirical Kaplan-Meier curve (black) and simulated Kaplan-Meier curves, which are summarized using the 95% credible interval (dashed lines) and the median (solid lines), from the CTIHMM (blue) and HIHMM (red).

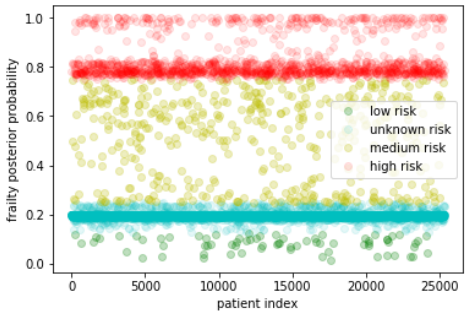

Figure 3: Posterior probabilities of belonging to the frailty class for each individual from a test set. Risk stratification is possible by thresholding the probabilities. Threshold probabilities in this example are $(0, 0.125, 0.25, 0.75, 1)$. Color indicates falling between two probability thresholds.

Examples of these posterior probabilities are shown in Fig 3. For illustration purposes, we have chosen risk thresholds of $\{0.125, 0.25, 0.75\}$ with the following interpretation.

$$
\begin{aligned}
0 \leq p(z_n|\boldsymbol{\psi}, \boldsymbol{O}_n) < 0.125 &\implies \text{low-risk} \\
0.125 \leq p(z_n|\boldsymbol{\psi}, \boldsymbol{O}_n) < 0.25 &\implies \text{unknown risk} \\
0.25 \leq p(z_n|\boldsymbol{\psi}, \boldsymbol{O}_n) < 0.75 &\implies \text{medium-risk} \\
0.75 < p(z_n|\boldsymbol{\psi}, \boldsymbol{O}_n) \leq 1 &\implies \text{high-risk}
\end{aligned}
$$

Two main clusters are apparent in the data corresponding to unknown risk and high risk. The unknown risk cluster is those patients close to the prior probability of 20%. These patients lack sufficient observations to make an informed decision about their risk profile. This suggests these patients should be followed up with the standard screening protocol. The high risk cluster is those patients which are more likely to be in a high-grade state. This suggests these patients may require immediate follow up. The two smaller clusters of low risk and medium risk are comprised of patients that may require decreased or increase screening frequencies, respectively, relative to the standard screening protocol.

In summary, this paper has made the following contributions:

- We make CTIHMM inference for population-level datasets possible by using piece-wise constant intensity functions and deriving a scalable inference algorithm.
- We put a hierarchical structure over the CTIHMM to explain population heterogeneity in terms of frailty, resulting in our HIHMM.
- We utilize prior distributions in the model to achieve more accurate estimates when data is scarce.

- We perform full model inference and prediction on subset of cancer screening dataset and show that our model outperforms other state of the art recurrent neural network models on the prediction task.
- We perform full model inference on a population-level cancer screening dataset and show that population heterogeneity improves model performance in terms of Kaplan-Meier estimators.
- We illustrate how the model may be used to better inform public health professionals by providing a risk stratification mechanism.

References

[1] Arturs Backurs and Christos Tzamos. Improving Viterbi is Hard: Better Runtimes Imply Faster Clique Algorithms. *arXiv e-prints*, art. arXiv:1607.04229, Jul 2016.

[2] Yujia Bao, Zhaobin Kuang, Peggy Peissig, David Page, and Rebecca Willett. Hawkes process modeling of adverse drug reactions with longitudinal observational data. In Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 177–190, Boston, Massachusetts, 18–19 Aug 2017. PMLR. URL `http://proceedings.mlr.press/v68/bao17a.html`.

[3] N Bartolomeo, P. Trerotoli, and G. Serio. Progression of liver cirrhosis to hcc: an application of hidden markov model. *BMC Med Research Method*, 11, 2011.

[4] Atilim Gunes Baydin, Barak Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Siskind. Automatic differentiation in machine learning: A survey. *Journal of Machine Learning Research*, 18:1–43, 04 2018.

[5] Matthew J. Beal, Zoubin Ghahramani, and Carl E. Rasmussen. The infinite hidden markov model. In *Machine Learning*, pages 29–245. MIT Press, 2002.

[6] Mogens Bladt and Michael Sørensen. Statistical inference for discretely observed markov jump processes. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(3):395–410, 2005. ISSN 13697412, 14679868. URL `http://www.jstor.org/stable/3647667`.

[7] Alexandre Bureau, Stephen Shiboski, and James P. Hughes. Applications of continuous time hidden markov models to the study of misclassified disease outcomes. *Statistics in Medicine*, 22(3):441–462, 2003. doi: 10.1002/sim.1270. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.1270`.

[8] K Canfell, R Barnabas, Patnick J, and Beral V. The predicted effect of changes in cervical screening practice in the uk: results from a modelling study. *British journal of cancer*, 91(3): 530–536, 2004.

[9] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv e-prints*, art. arXiv:1406.1078, Jun 2014.

[10] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv e-prints*, art. arXiv:1412.3555, Dec 2014.

[11] D. R. Cox and H. D. Miller. *The Theory of Stochastic Processes*. 1965.

[12] P. Doetsch, M. Kozielski, and H. Ney. Fast and robust training of recurrent neural networks for offline handwriting recognition. In *2014 14th International Conference on Frontiers in Handwriting Recognition*, pages 279–284, Sep. 2014. doi: 10.1109/ICFHR.2014.54.

[13] Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. Transition-Based Dependency Parsing with Stack Long Short-Term Memory. *arXiv e-prints*, art. arXiv:1505.08075, May 2015.

[14] Jurgen V. Gael, Yee W. Teh, and Zoubin Ghahramani. The infinite factorial hidden markov model. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1697–1704. Curran Associates, Inc., 2009. URL `http://papers.nips.cc/paper/3518-the-infinite-factorial-hidden-markov-model.pdf`.

[15] Asger Hobolth and Jens Ledet Jensen. Summary statistics for endpoint-conditioned continuous-time markov chains. *J. Appl. Probab.*, 48(4):911–924, 12 2011. doi: 10.1239/jap/1324046009. URL `https://doi.org/10.1239/jap/1324046009`.

[16] Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. *Neural Computation*, 1997.

[17] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. An efficient k-means clustering algorithm: analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):881–892, July 2002. ISSN 0162-8828. doi: 10.1109/TPAMI.2002.1017616.

[18] Michael Kearns, Yishay Mansour, and Andrew Y. Ng. An information-theoretic analysis of hard and soft assignment methods for clustering. In *UAI*, pages 282–293. Morgan Kaufmann, 1997.

[19] Wael Khreich, Eric Granger, Ali Miri, and Robert Sabourin. On the memory complexity of the forward–backward algorithm. *Pattern Recognition Letters*, 31(2):91 – 99, 2010. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2009.09.023. URL http://www.sciencedirect.com/science/article/pii/S0167865509002578.

[20] Genshiro Kitagawa. Non-gaussian state-space modeling of nonstationary time series. *Journal of the American Statistical Association*, 82(400):1032–1041, 1987. ISSN 01621459. URL http://www.jstor.org/stable/2289375.

[21] Vikram Krishnamurthy, Elisabeth Leoff, and Jörn Sass. Filterbased stochastic volatility in continuous-time hidden markov models. *Econometrics and Statistics*, 6:1 – 21, 2018. ISSN 2452-3062. doi: https://doi.org/10.1016/j.ecosta.2016.10.007. URL http://www.sciencedirect.com/science/article/pii/S2452306216300144. STATISTICS OF EXTREMES AND APPLICATIONS.

[22] Zhaobin Kuang, James Thomson, Michael Caldwell, Peggy Peissig, Ron Stewart, and David Page. Baseline regularization for computational drug repositioning with longitudinal observational data. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 2521–2528. AAAI Press, 2016. ISBN 978-1-57735-770-4. URL http://dl.acm.org/citation.cfm?id=3060832.3060974.

[23] Zhaobin Kuang, Peggy Peissig, Vitor Santos Costa, Richard Maclin, and David Page. Pharmacovigilance via baseline regularization with large-scale longitudinal observational data. *KDD : proceedings. International Conference on Knowledge Discovery & Data Mining*, 2017. doi: 10.1145/3097983.3097998.

[24] I. K. Larsen, M. Smastuen, T. B. Johannesen, F. Langmark, D. M. Parkin, F. Bray, and B. Moller. Data quality at the cancer registry of norway: an overview of comparability, completeness, validity and timeliness. *Eur J Cancer*, 45(7):1218–31, 2009. ISSN 1879-0852 (Electronic). URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=19091545.

[25] M. K. Leinonen, S. A. Hansen, G. B. Skare, I. B. Skaaret, M. Silva, T. B. Johannesen, and M. Nygard. Low proportion of unreported cervical treatments in the cancer registry of norway between 1998 and 2013. *Acta Oncol*, 57(12):1663–1670, 2018. ISSN 1651-226X (Electronic) 0284-186X (Linking). doi: 10.1080/0284186X.2018.1497296. URL https://www.ncbi.nlm.nih.gov/pubmed/30169991.

[26] Dong C. Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Math. Program.*, 45(1-3):503–528, August 1989. ISSN 0025-5610. doi: 10.1007/BF01589116. URL https://doi.org/10.1007/BF01589116.

[27] Yu-Ying Liu, Hiroshi Ishikawa, Mei Chen, Gadi Wollstein, Schuman Joel S., and James M. Rehg. Longitudinal modeling of glaucoma progression us-ing 2-dimensional continuous-time hidden markov model. *Med Image Comput Assist Interv.*, 2013.

[28] Yu-Ying Liu, Shuang Li, Fuxin Li, Le Song, and James M Rehg. Efficient learning of continuous-time hidden markov models for disease progression. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3600–3608. Curran Associates, Inc., 2015.

[29] C. Van Loan. Computing integrals involving the matrix exponential. *IEEE Transactions on Automatic Control*, 23(3):395–404, June 1978. ISSN 0018-9286. doi: 10.1109/TAC.1978.1101743.

[30] Shaochuan Lu. A continuous-time hmm approach to modeling the magnitude-frequency distribution of earthquakes. *Journal of Applied Statistics*, 44(1):71–88, 2017. doi: 10.1080/02664763.2016.1161736. URL https://doi.org/10.1080/02664763.2016.1161736.

[31] Dougal Maclaurin. *Modeling, Inference and Optimization with Compposable Differentiable Procedures*. PhD dissertation, Harvard University, 2016.

[32] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and Optimizing LSTM Language Models. *arXiv e-prints*, art. arXiv:1708.02182, Aug 2017.

[33] Philipp Metzner, Illia Horenko, and Christof Schütte. Generator estimation of markov jump processes based on incomplete observations nonequidistant in time. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 76:066702, 01 2008.

[34] Evan R. Myers, Douglas C. McCrory, Kavita Nanda, Lori Bastian, and David B. Matchar. Mathematical Model for the Natural History of Human Papillomavirus Infection and Cervical Carcinogenesis. *American Journal of Epidemiology*, 151(12):1158–1171, 06 2000. ISSN 0002-9262. doi: 10.1093/oxfordjournals.aje.a010166. URL https://dx.doi.org/10.1093/oxfordjournals.aje.a010166.

[35] Shawn E. Simpson, David Madigan, Ivan Zorych, Martijn J. Schuemie, Patrick B. Ryan, and Marc A. Suchard. Multiple self-controlled case series for large-scale longitudinal observational databases. *Biometrics*, 69(4):893–902, 2013. doi: 10.1111/biom.12078. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12078.

[36] Frank A. Sonnenberg and Beck J. Robert. Markov models in medical decision making: A practical guide. *Med Decis Making*, 13:322–338, 1993.

[37] Braden Soper, Rui Meng, Jan F. Nygard, Mari Nygard, and Herbert Lee. An hpc application to population-level cancer screening data. Abstract from HPC Applications in Precision Medicine, Frankfurt, Germany, June 2018.

[38] Cem Subakan, Johannes Traa, and Paris Smaragdis. Spectral learning of mixture of hidden markov models. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2249–2257. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5518-spectral-learning-of-mixture-of-hidden-markov-models.pdf.

[39] Andrew C. Titman and Linda D. Sharples. A general goodness-of-fit test for Markov and hidden Markov models. *Statistics in Medicine*, 27(12):2177–2195, May 2008. ISSN 02776715, 10970258. doi: 10.1002/sim.3033. URL http://doi.wiley.com/10.1002/sim.3033.

[40] G. Ursin, S. Sen, J. M. Mottu, and M. Nygard. Protecting privacy in large datasets-first we assess the risk; then we fuzzy the data. *Cancer Epidemiol Biomarkers Prev*, 26(8):1219–1224, 2017. ISSN 1538-7755 (Electronic) 1055-9965 (Linking). doi: 10.1158/1055-9965.EPI-17-0172. URL https://www.ncbi.nlm.nih.gov/pubmed/28754793.

[41] Xiang Wang, David Sontag, and Fei Wang. Unsupervised learning of disease progression models. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 85–94, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2956-9. doi: 10.1145/2623330.2623754. URL http://doi.acm.org/10.1145/2623330.2623754.

[42] Wei Wei, Bing Wang, and Don Towsley. Continuous-time hidden markov models for network performance evaluation. *Performance Evaluation*, 49(1):129 – 146, 2002. ISSN 0166-5316. doi: https://doi.org/10.1016/S0166-5316(02)00122-0. URL http://www.sciencedirect.com/science/article/pii/S0166531602001220. Performance 2002.

[43] Amy MF Yen, Tony HH Chen, Stephen W Duffy, and Chih-Dao Chen. Incorporating frailty in a multi-state model: application to disease natural history modelling of adenoma-carcinoma in the large bowel. *Statistical Methods in Medical Research*, 19(5):529–546, 2010. doi: 10.1177/0962280209359862. URL https://doi.org/10.1177/0962280209359862. PMID: 20488838.

[44] A.I. Zeifman and Dean L. Isaacson. On strong ergodicity for nonhomogeneous continuous-time markov chains. *Stochastic Processes and their Applications*, 50(2):263 – 273, 1994. ISSN 0304-4149. doi: https://doi.org/10.1016/0304-4149(94)90123-6. URL `http://www.sciencedirect.com/science/article/pii/0304414994901236`.

[45] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video Summarization with Long Short-term Memory. *arXiv e-prints*, art. arXiv:1605.08110, May 2016.

7 Supplemental Materials for Reproducibility

7.1 Details of Age Segmentation

Since HPV status is one of most import indicator for cervical censor, we segment the age interval based on the empirical density of ages at which patients find positive HPV. The empirical density is estimated based on 100000 patients randomly sampled from the pool using gaussian kernel estimation. Then we fit the density of ages using discontinuous piece-wise linear function with different number of intervals.

Fitting information is summarized in Table 8. We plot optimal sum of square errors under different N in Figure 4. From the figure, it visually shows that $N = 4$ is the optimal number of segmentation based on elbow criteria. Then combining the expert's opinion, we set the corresponding cutting points as $[23, 30, 60]$.

Table 8: Discontinuous piece-wise linear fitting under different number of intervals N. Optimal sum of square errors (SSE) and cutting points (CPs) are given.

N	SSE	CPs
2	3.88e-3	24.8
3	1.78e-3	24.7, 54.7
4	1.23e-3	25.2, 35.6, 60.6
5	0.93e-3	24.5, 26.2, 30.7, 67.2
6	0.78e-3	24.1, 25.2, 32.8, 56.9, 64.4

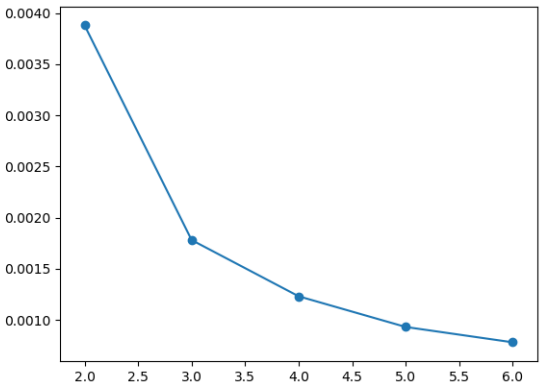

Figure 4: Sum of square errors under different number of segmentation N.

7.2 Details of Parameter Transformation in Limited-Memory BFGS Optimization

In the HIHMM, model parameters $\boldsymbol{\psi}$ are summarized as follows:

- Emission parameters $\tilde{\boldsymbol{\alpha}}_{sk} > 0$ and $\eta_{sk} > 0$ for $s = 1, \dots M$ and $k = 1, \dots, K$.
- Initial state parameters $\boldsymbol{\alpha}_{zi} > 0$ for $z = 1, \dots, Z$ and $i = 1, \dots, |\mathcal{A}|$.
- Unconstrained transition intensity parameters $\boldsymbol{\lambda}_{zi} > 0$ for $z = 1, \dots z$ and $i = 1, \dots, |\mathcal{A}|$.

In the M-step, all model parameters need to be optimized. To take advantage of the L-BFGS approach, we need to transform the constrained optimization problem to an unconstrained optimization problem. To do this we transform all parameters on the log scale, which means we let $\tilde{\boldsymbol{\alpha}}^* = \log \tilde{\boldsymbol{\alpha}}$, $\boldsymbol{\eta}^* = \log \boldsymbol{\eta}$, $\boldsymbol{\alpha}^* = \log \boldsymbol{\alpha}$ and $\boldsymbol{\lambda}^* = \log \boldsymbol{\lambda}$. Then we let $\boldsymbol{\psi}^* = (\tilde{\boldsymbol{\alpha}}^*, \boldsymbol{\eta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\lambda}^*)$. Finally we maximize the EMCLL with respect to $\boldsymbol{\psi}^*$ rather than $\boldsymbol{\psi}$. After the optimization, the optimized $\hat{\boldsymbol{\psi}}$ are accessed by $\hat{\boldsymbol{\psi}} = \exp(\hat{\boldsymbol{\psi}}^*)$.

7.3 Details of Learning Settings

In the proposed learning approach, we set the number of EM iterations at $N_{EM} = 100$, and in the L-BFGS approach we set the number of optimization iterations as $N_{L-BFGS} = 8$. The automatic differentiation is implemented using the autograd package [31] in Python.

7.4 Details of Simulations of Kaplan-Meier Curves

To simulate one Kaplan-Meier curve from a HIHMM, we propose the following procedures:

1 First, reduce all individuals' ages at their first screening test to a set A_1, and categorize individuals' time intervals between two consecutive screening tests into four sets denoted as $\tilde{I}_i$ for $i = 0, \dots 3$. Any time interval (a, b) is categorized into $\tilde{I}_i$, if and only if the largest value of both cytology and histology screening test results at time a is i. Then for each set $\tilde{I}_i$, we map elements of $\tilde{I}_t$ to their corresponding lengths and name the new set as I_i Also, we reduce all posterior probabilities of model indexes into a set $\tilde{P}$.

2 Second, we sample an initial age a_1 from A_1 and sample a posterior probability of model index $\tilde{p}$ from $\tilde{P}$.

3 Then sample model index z by $z \sim \mathrm{Ber}(\tilde{p})$ and sample an initial state $S_1 \sim \mathrm{Cat}(\hat{\boldsymbol{\pi}}_z(\mathcal{A}, a_1))$, where $\hat{\boldsymbol{\pi}}_{zi} = \hat{E}(\boldsymbol{\pi}_{zi}) = \frac{\hat{\boldsymbol{\alpha}}_{zi}}{\sum \hat{\boldsymbol{\alpha}}_{zi}}$.

4 Based on current state S_{t-1} and current age a_{t-1} , sample the screening time interval Δ_{t-1} from $I_{S_{t-1}}$. Then compute transition matrix $P([a_{t-1}, a_t]|\mathcal{A}, \hat{\boldsymbol{\lambda}}_z)$ from age a_{t-1} to $a_t = a_{t-1} + \Delta t - 1$. Then sample S_t by $S_t \sim \mathrm{Cat}(P[a_{t-1}, a_t]_{S_{t-1},:})$.

5 Continue previous sampling processes until the state S_T fails according to the failure definition of Kaplan-Meier estimator. Then compute the failure time by $F = \sum_{t=1}^{T} \Delta_t$.

6 Repeat [2] to [5] M times. We obtain M failure times and then order them as a sequence $\{F_m\}$.

7 Based on the simulated failure times $\{F_m\}$, the simulated curve is $S(t) = 1 - \frac{1}{M} \sum_{m:t \geq F_m} 1$.

Multiple Kaplan-Meier curves are simulated by independently repeating the above procedures.