Received: 10 May 2019 | Revised: 2 June 2020 | Accepted: 11 June 2020

DOI: 10.1002/sim.8681

RESEARCH ARTICLE

Statistics in Medicine WILEY

A hidden Markov model for population-level cervical cancer screening data

Braden C. Soper[1] | **Mari Nygård**[2] | **Ghaleb Abdulla**[1] | **Rui Meng**[3] | **Jan F. Nygård**[4]

[1]Computing Directorate, Lawrence Livermore National Laboratory, Livermore, California, USA

[2]Research Department, Cancer Registry of Norway, Oslo, Norway

[3]Department of Statistics, University of California, Santa Cruz, California, USA

[4]Registry Informatics Department, Cancer Registry of Norway, Oslo, Norway

Correspondence
Braden C. Soper, Computing Directorate, Lawrence Livermore National Laboratory, 7000 East Ave., Livermore, CA 94550, USA.
Email: soper3@llnl.gov

Funding information
Lawrence Livermore National Laboratory

The Cancer Registry of Norway has been administrating a national cervical cancer screening program since 1992 by coordinating triennial cytology exam screenings for the female population between 25 and 69 years of age. Up to 80% of cancers are prevented through mass screening, but this comes at the expense of considerable screening activity and leads to overtreatment of clinically asymptomatic precancers. In this article, we present a continuous-time, time-inhomogeneous hidden Markov model which was developed to understand the screening process and cervical cancer carcinogenesis in detail. By leveraging 1.7 million individual's multivariate time-series of medical exams performed over a 25-year period, we simultaneously estimate all model parameters. We show that an age-dependent model reflects the Norwegian screening program by comparing empirical survival curves from observed registry data and data simulated from the proposed model. The model can be generalized to include more detailed individual-level covariates as well as new types of screening exams. By utilizing individual screening histories and covariate data, the proposed model shows potential for improving strategies for cancer screening programs by personalizing recommended screening intervals.

KEYWORDS

cancer screening, cervical cancer, hidden Markov model, personalized screening, population-level data, precision medicine, real-world evidence

1 | INTRODUCTION

Cervical cancer continues to be a major public health challenge with approximately 570 000 new cases and 311 000 deaths each year.[1] Organized screening programs have been shown to be an effective policy in reducing morbidity and mortality. For example, in Nordic countries it has been estimated that through regular mass screening of the female population, up to 80% of cervical cancer cases have been prevented.[2]

The frequency of screenings plays a critical role in the effectiveness of these programs. Screening too frequently is costly, burdensome to the population, and can lead to the overtreatment of clinically insignificant precancers. Screening too infrequently risks missing precancers which require treatment. The recommended time between a normal and a subsequent screening exam is known as the screening interval, and is typically determined by experts after a thorough review of the available evidence, typically focusing on reducing the risk of developing cervical cancer in the years following a normal examination.[3] Across cytology-based screening programs the recommended screening interval ranges from 1 to 5 years.[3,4]

Abbreviations: HMM, hidden Markov model; HPV, human papillomavirus; MMS, multistate Markov model.

An alternative to homogeneous, mass screening programs is a more personalized screening program that takes individuals' data into account when developing screening interval recommendations. Such personalized screenings would allow for a safe decrease in the screening frequency of women at low-risk and a prudent increase in the screening frequency of women with elevated risk. Such a system has the potential to reduce costs, both in terms of healthcare dollars and patient well-being. In the case of cervical cancer screening, such a system will be of great utility when human papillomavirus (HPV)-vaccinated women enter the screening program.[5]

A necessary step toward personalized screening programs is the development of reliable predictive models that take into account both disease dynamics and individual patient data. Multistate Markov models (MMSs) and hidden Markov models (HMMs) are two well-known dynamic, probabilistic graphical models[6] with a long history in biomedical applications. Their generality and flexibility make them attractive models for biomedical panel data, with both MMSs and HMMs having been applied to a wide variety of biomedical domains including cancer. A nonexhaustive list of examples include liver cancer,[7] colorectal cancer,[8] bladder cancer,[9] cancer recurrence,[10,11] and Leukemia.[12] Direct applications to cancer screening include breast,[13-15] colorectal,[16] and cervical.[17-22]

In most of the studies applying Markov models to cervical cancer, including Myers et al[18] and Malagón et al,[19] model parameters are not learned from real-world individual-level panel-data. Instead the Markov models are used as stochastic computational models that study disease progression and regression within a simulated population. Model parameters are either instantiated with values from the literature or are calibrated against various global population statistics from the literature. Furthermore, they do not adequately address parameter uncertainty.[22] In other applications of Markov models to cervical cancer, the focus is on the evaluation of different screening tests or economic impacts.[20-22] To the best of our knowledge, the only direct application of HMMs to cervical cancer screening data is by Kirby and Spiegelhalter,[17] which modeled cervical cancer screening as a two-state HMM in continuous-time.

In this article, we present a continuous-time, time-inhomogeneous HMM which models the multistage morphological changes leading to cervical cancer. We perform full model inference using population-level data, estimating age-dependent transition intensities and screening exam error probabilities. In order to handle the relatively large data size and model complexity, we parallelize the inference over thousands of processors utilizing the high-performance computing (HPC) capabilities of Lawrence Livermore National Laboratory (LLNL). This allowed us to jointly model population-level data, performing maximum likelihood inference with hundreds of thousands of individual time-series from the Norwegian cervical cancer screening program. The model is validated by comparing Kaplan-Meier estimators based on a hold-out set of screening data and synthetic data generated from the learned model.

2 | DATA

The Cancer Registry of Norway has run a national cervical cancer screening program since October 1991, collecting all screening and diagnostic results.[23,24] Though screening guidelines exist, screening is at the discretion of the individual. As a result, the number of screening records and the time between screenings vary considerably between women.

Several features of the dataset make it unique, the first being its size. Norway's cervical cancer screening program records cervical exams from the entire target population. Thus, the dataset is a true population-level dataset, containing over 1.7 million women with more than 10 million exams recorded from 1992 until the end of 2015. In the current dataset all women above 16 years of age with more than one exam have been included. The maximum number of exams observed was 53, while the median number of exams was 6. Three types of exams are used in the screening program: cytology, histology, and HPV. The data was randomly sampled and divided into a training set of 300 000 women for parameter estimation, and a hold out set of 300 000 women for model validation.

Second, the data has been deidentified and slightly obscured for additional anonymization, the details of which can be found in Ursin et al.[25] One modification made through this process is that all exam dates are coarsened to the month level. Thus, we do not have access to the ordering or timing of multiple screening exams occurring within the same calendar month. As a consequence any model for such data must be able to deal with multiple tests, in both type and quantity, performed at each observation. Though this feature of the dataset is introduced artificially, multiple simultaneous tests are common in medical practice and should be incorporated into any robust model of biomedical diagnostics and/or screening.

Finally, because the data is from a population-level screening program and cancer is a relatively rare disease, the resulting dataset is highly skewed toward disease-free observations. As a result certain exam results are particularly rare.

While this can be a common problem in biomedical data, particularly in oncology, the problem is compounded by the stated goal of building more personalized predictive models.

3 | MODEL

Cervical cancer is an infrequent end-stage of minor cellular abnormalities caused by an HPV infection.[26] These abnormalities progress from minor changes, through more definitely premalignant changes, to localized invasive cancer. If left untreated this can lead to metastatic disease and ultimately death. Being able to detect cancer in its early development, followed by prompt, appropriate treatment is the key element which justifies cancer screening programs.[27,28]

At its most crude approximation, the state of the cervix can be labeled as normal or abnormal, according to signs indicative of cancer or cancer development. This is the modeling approach taken in Kirby and Spiegelhalter.[17] At the other extreme one could consider a continuum of states between healthy and cancerous. Because our ultimate objective is to construct a model that will help improve screening procedures through more personalized screening guidelines, we have chosen the number of states to correspond to actionable clinical diagnoses based on an individual's risk for being diagnosed with cancer in the future. We label these states *normal*, *low-risk*, *high-risk*, and *cancer*. Normal indicates no need for additional exams before the next screening. Low-risk indicates the need for intensified medical follow-up, but without treatment. High-risk indicates a state which requires immediate treatment because the probability for further cancer progress is high. To capture competing events such as death and emigration, we consider a fifth *censoring/death* state. For notational simplicity we will label the states normal, low-risk, high-risk, cancer, and censoring/death by $N0,L1,H2,C3$, and $D4$, respectively.

We assume that cervical cancer progression and regression are subject to the Markov property. While this may be a strong assumption, it is a reasonable first-order approximation to disease dynamics since future states will certainly depend on the current state. We furthermore assume that states $N0$, $L1$, $H2$, and $C3$ can only transition from and to adjacent states. For example, it is possible to transition from state $L1$ to states $N0$ and $H2$, but not $C3$ directly. It is possible to transition from any state into state $D4$, which is absorbing.

Note that it is not possible to directly observe the states $N0,L1,H2,C3$, and $D4$. Instead, exams which may be subject to error are used as indirect observations of these states. Cytology, histology, and HPV tests are given as a means of determining an individual's risk of having a precancerous lesions or cervical cancer. We can interpret this screening process as an attempt to classify a woman's cervix into one of the four possible states. It is in this context that we interpret the cancer screening process as a HMM.

The direct application of HMMs to the Norwegian screening data is not straight forward due to the data obfuscation process. Modifications to standard continuous-time HMMs (to be discussed below) are necessary to handle such data, putting our model in the more general class of continuous-time dynamic Bayesian networks. We note that even though the screening data is given at the month level as a result of the data obfuscation, we still model the underlying dynamics in continuous-time. This is partly done for convenience in handling the irregularly sampled data, but it is also more computationally stable. Computing the likelihood of a discrete-time HMM at the month level is numerically unstable due to the need to raise the probability transition matrix to a large power. Furthermore it can be argued that cervical cancer progresses on a longer time scale than months,[29] thus it is a reasonable approximation to use a continuous-time model with discrete-time data.

Finally, we also account for surgical treatment. Surgery removes the observed abnormal tissue from the cervix and we therefore model this as a direct transition to the normal state. This is clearly an approximation, as it is possible that a treatment does not remove all abnormalities. Women with late stage cancer and women who receive types of treatment other than surgery are kept in the $C3$ stage and can only transition into $D4$. A more general approach would be to introduce a probability distribution on the underlying states an individual can enter after a treatment. Due to the high success rates of treating precancerous lesions and early stage cancer of the cervix, this is a reasonable assumption to make.[30] A graphical model of this Markov chain is shown in Figure 1. The next two subsections present the model in more technical detail.

3.1 | Markov model for cervical cancer

Let there be N women in the screening population. Denote by $S^i(t)$ the underlying state of woman $i \in \{1,2, \dots ,N\}$ at time $t \geq 0$. We assume that for each i, $S^i(t)$ follows a continuous-time, Markov jump process on state space $\{N0,L1,H2,C3,D4\}$.

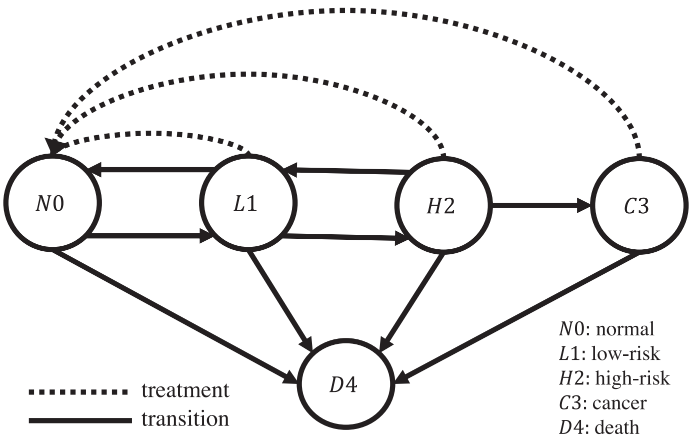

FIGURE 1 Continuous-time Markov model of cervical cancer development with consecutive states of normal, low-risk, high-risk, cancer, and death. The solid lines represent possible transitions in the Markov chain while dashed lines represent instantaneous resets due to surgical treatment

To further simplify notation we take the state space to be $\{0,1,2,3,4\}$, where the cervical cancer risk states map on to their associated numeric values. We denote this state space by $\mathcal{S}$.

A general continuous-time, Markov jump process on a finite state-space $\mathcal{S}$ can be characterized by an initial state probability distribution π over $\mathcal{S}$ and a probability transition function $P(y,t|x,s)$, which gives the probability of the process being in state $y \in \mathcal{S}$ at time $t > 0$, given that the process was in state $x \in \mathcal{S}$ at time $s < t$. Denote by $P(t|s)$ the matrix whose i,j entry is $P(j,t|i,s)$. According to the Kolmogorov equations, the matrix $P(t|s)$ is determined by an *intensity matrix* $\Lambda(u)$ with i,j entry $\lambda_{ij}(u)$ defined for all $u \in [s,t]$. Specifically, given such an intensity matrix $\Lambda(u)$, the probability transition matrix $P(t|s)$ is the unique solution to the matrix differential equation $\frac{\partial}{\partial t}P(t|s) = P(t|s) \cdot \Lambda(t)$, with initial condition $P(s|s) = \mathbf{I}_n$, the identity matrix.

If $\Lambda(t) = \Lambda$ is constant in t, the solution to the differential equation is the matrix exponential $P(t|s) = e^{(t-s)\Lambda}$. In this case the resulting Markov chain only depends on the time difference $t - s$ and is said to be time-homogeneous. Depending on the underlying Markov model, this matrix may or may not have a closed form representation. Regardless, efficient numerical schemes exist to compute this matrix.[31]

A time-homogeneous assumption is often made as a means of simplifying the analysis and inference of the resulting Markov model. However, cervical cancer is not a time-homogeneous process. Certain strains of the HPV are known to be the primary causes of cervical cancer, and certain conditions accelerate progression from low-grade lesions caused by HPV to cancer.[32-35] Exposure to HPV is typically through sexual contact, thus risk of HPV infection is highly correlated with age. Moreover, the dependence on age is nonmonotonic, as the exposure to HPV increases from late teens to mid-twenties and then decreases as women get older, with a possible slight increase in risk in early middle-age. Figure 2 shows the empirical distribution of women's ages at the time of a positive HPV test alongside the empirical distribution of women's ages at the time of a screening. Note that women are typically only given an HPV test after an abnormal cytology exam in the time period of this study. Nonetheless it is clear that HPV infections are not uniformly distributed across the screening population. The discrepancy in distributions suggests that HPV incidence is strongly correlated with age.

Because of the high correlation between age and HPV exposure, we use age as a proxy for HPV exposure risk and model the transition intensity matrix $\Lambda(t)$ as a function of age. We thus model cervical cancer as a time-inhomogeneous Markov jump process. Because the dependence on age may be nonmonotonic, we allowed for nonlinearities in the functional dependence on age.

A flexible and efficient method is to model the transition intensities λ_{ij} as piecewise constant functions of a woman's age. Assume that woman i is observed at ages $a^i_1, a^i_2, \dots, a^i_{n_i}$ where n_i is the number of exams for woman i. Recall that even though the recommended screening interval is highly regular, the actual real-world screening exams are highly irregular. Thus, the time between observations for a single woman is not uniformly spaced ($a^i_{j+1} - a^i_j \neq a^i_{k+1} - a^i_k$ for all $j,k \in \{1,2,\dots,n_i\}$) and the number of exams between women varies ($n_i \neq n_j$ for $i \neq j$). Let $a_{\min} = \min_{i,j}\{a^i_j\}$ and $a_{\max} = \max_{i,j}\{a^i_j\}$. Then for all women i and all exams j we have $a_{\min} \leq a^i_j \leq a_{\max}$. We define a partition of the interval $[a_{\min}, a_{\max}]$ by $q + 1$ change-points, $\tau_0, \tau_1, \tau_2, \dots, \tau_q$, satisfying $\tau_0 = 0$ and $a_{\min} = \tau_1 < \tau_2 < \dots < \tau_{q-1} < \tau_q = a_{\max}$. Then for woman i at observation j, for each $r \in \{1,2,\dots,q\}$ we define the time-dependent indicator variable $x^r_{ij} = 1$ if $\tau_{r-1} < a^i_j \leq \tau_r$ and $x^r_{ij} = 0$ otherwise. For each $r \in \{1,2,\dots,q\}$ and $x, y \in \mathcal{S}$ we introduce parameters β^r_{xy}. The intensity parameters λ_{xy} for woman i are then $\lambda_{xy}(a^i_j) = \sum_{r=1}^{q} \beta^r_{xy} x^r_{ij}(a^i_j)$. For notational convenience we denote the vector of ages at observations for woman i by

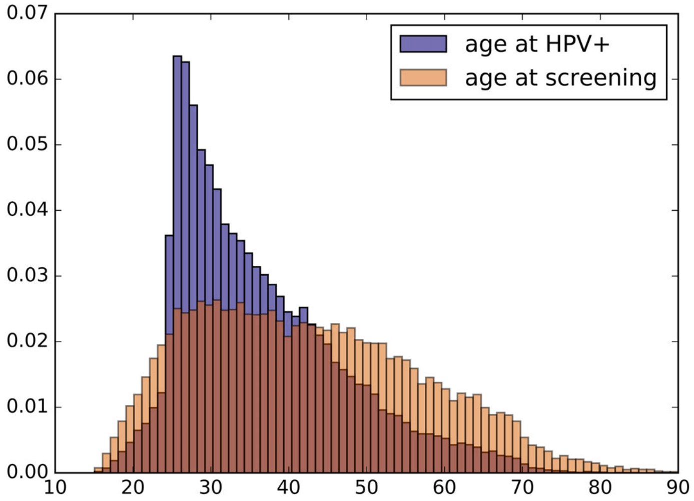

FIGURE 2 Histogram of ages of women at time of positive HPV test and age at screening. HPV, human papillomavirus [Color figure can be viewed at wileyonlinelibrary.com]

$\boldsymbol{a}^i$ and the set of all β_{xy}^r by $\boldsymbol{\beta}$. Considering the transition intensity matrix Λ as a function of age a_j^i and parameters $\boldsymbol{\beta}$ we write $\Lambda(a_j^i, \boldsymbol{\beta})$ as the matrix with xy entry $\lambda_{xy}(a_j^i)$.

The initial state distribution π can be thought of as a prior distribution on a woman's state the first time she comes in for a screening. Because of the time-inhomogeneous assumption, we expect the initial state distribution π to also be a function of the individual's age. For each $s \in \mathcal{S}$ denote by π_s the probability that a woman is in state s at the time of an initial screening. We then assume that the initial state distribution depends on the age of the woman at her initial screening, namely, $\pi_s = \pi_s(a_1^i)$. Note that we set $\pi_4 = 0$, the alternative being meaningless. Then for each $r \in \{1,2, \dots ,q\}$ and each $s \in \mathcal{S}$ we introduce parameters c_s^r so that the initial state probabilities π_s for woman i become $\pi_s^i(a_1^i) = \frac{\sum_{r=1}^q e^{c_s^r} x_{i1}^r(a_1^i)}{\sum_{\ell=0}^3 \sum_{r=1}^q e^{c_\ell^r} x_{i1}^r(a_1^i)}$. For identifiability reasons we set $c_0^r = 0$ for all r. The set of all c_s^r is denoted by c

3.2 | HMM for screening

With the above Markov model of cervical cancer progression and regression, we now proceed to define the screening process in terms of a modified HMM. The dataset consists of N variable length sequences of results from various medical exams. Let the number of exams for individual i be n_i. If individual i is screened at times $t_1^i < t_2^i < \dots < t_{n_i}^i$, we denote the underlying hidden state of individual i at time t_j^i by $S_j^i = S^i(t_j^i)$. The entire sequence of hidden states for individual i is given by $\mathbf{S}^i = (S_1^i, S_2^i, \dots , S_{n_i}^i)$.

At each screening visit up to $K > 0$ types of medical exams may be given to a woman. Let r_k denote the number of possible outcomes for diagnostic exam $k \in \{1,2, \dots ,K\}$. It is possible that exam k is performed multiple times at one screening visit. For example, at screening j individual i may be given a cytological exam more than once. We denote the number of exams of type k given to woman i at screening j by T_{jk}^i. Then given T_{jk}^i we let $\mathbf{R}_{jk}^i$ be the vector of length r_k whose elements are the counts of outcomes from all exams of type k for woman i at screening j. Thus, if $R_{jk\ell}^i$ is the ℓth element of the vector $\mathbf{R}_{jk}^i$, then $\sum_{\ell=1}^{r_k} R_{jk\ell}^i = T_{jk}^i$. We furthermore denote the vector of exam counts given to individual i at visit j by $\mathbf{T}_j^i = (T_{j1}^i, T_{j2}^i, \dots, T_{jK}^i)$. Similarly we denote the sequence of vectors corresponding to exam results for individual i at visit j by $\mathbf{R}_j^i = (\mathbf{R}_{j1}^i, \mathbf{R}_{j2}^i, \dots , \mathbf{R}_{jK}^i)$. Finally, we denote the entire sequence of observations for individual i by $\mathbf{T}^i = (\mathbf{T}_1^i, \dots , \mathbf{T}_{n_i}^i)$ and $\mathbf{R}^i = (\mathbf{R}_1^i, \dots , \mathbf{R}_{n_i}^i)$.

While the observed exam results R_{jk}^i are typically the data of most interest in standard HMMs, we will also leverage the information in the panel of diagnostics ordered at a given screening. In population-level registries descriptions as to why a certain set of screening exams were given are often missing. Absent such information we must assume there are latent features present that the attending clinician had access to that we do not. These latent features may be highly correlated with the underlying state of the screened woman. For example, if the screened woman is exhibiting symptoms, this may lead the doctor to order a panel of tests that she would otherwise not order had the woman not exhibited symptoms. Conversely, if the woman is not exhibiting observable symptoms, the doctor is likely to order fewer and less invasive screening exams. In either case the number and type of screening exam performed at a particular screening time is informative of the underlying disease state. This is of particular importance in the Norwegian cervical cancer screening data because of the data obfuscation that resulted in screenings only being identified at the granularity of months. Multiple tests during a single screening, while not common in the data, are highly informative, since these are typically cases that required immediate follow-up. We thus consider the variables T^i_{jk} as observable data along with the R^i_{jk}.

In the Norwegian cervical cancer dataset we have $K = 3$: cytology, histology, and HPV. For notational simplicity we associate these three exams with the labels 1,2,3, respectively. Note that it is possible to choose $K > 3$ as there are numerous distinct HPV tests which could be modeled as distinct exam types. For model parsimony we consider all HPV tests to be of one type. Since HPV tests are binary (positive/negative) we have $r_3 = 2$.

Cytology and histology have many more possible results than HPV tests giving us r_1,$r_2 > 2$. For model parsimony, we reduce the number of possible results for cytology and histology by categorizing the results in accordance with the latent risk level they are most indicative of. Thus, both cytology and histology can take values in $\{N0,L1,H2,C3\}$ giving us $r_1 = r_2 = 4$. In this way we can think of the clinical diagnosis process as a classification problem, where the clinician tries to classify each woman to actionable states $\{N0,L1,H2,C3\}$, which is subject to misclassification.

With the above notation and definitions we can write the hierarchical model for woman i as follows.

$$\begin{array}{ll} S^i_1|a^i_1,\boldsymbol{c} & \sim \text{Categorical}(\pi^i(a^i_1,\boldsymbol{c})) \\ \mathbf{S}^i|S^i_1,a^i_1,\boldsymbol{\beta} & \sim \text{Markov Process}(\Lambda(\cdot,\boldsymbol{\beta}),S^i_1,a^i_1) \\ T^i_{jk}|S^i_j & \sim \text{Poisson}(\eta^k_{S^i_j}) \\ R^i_{jk}|T^i_{jk},S^i_j,\boldsymbol{\theta}^k_{S^i_j} & \sim \text{Multinomial}(T^i_{jk},\boldsymbol{\theta}^k_{S^i_j}). \end{array}$$

4 MODEL PARAMETER ESTIMATION

Standard methods for performing maximum likelihood estimation in HMMs are based on the expectation-maximization (EM) algorithm.[36] One example is the Baum-Welch (BW) algorithm[37] which utilizes a forward-backward algorithm to compute posterior marginal distributions of all hidden states for the E-step of an EM algorithm. By modifying this forward-backward algorithm to allow for instantaneous transitions to the normal state after surgical treatment, we are able to compute the posterior marginal distributions for all hidden states in the E-step. The modification involves decomposing an individual's likelihood function by treatments (if there are any). Each component likelihood can then be treated as a conditionally independent sequence of observations.

The details of this modification are as follows. Let **S**,**T**, and **R** denote the collection of variables over all individuals, namely, $\mathbf{S} = \{\mathbf{S}^i\}_{i=1,\dots,N}$, $\mathbf{T} = \{\mathbf{T}^i\}_{i=1,\dots,N}$, and $\mathbf{R} = \{\mathbf{R}^i\}_{i=1,\dots,N}$. Furthermore let $\boldsymbol{\phi} = (\mathbf{c},\boldsymbol{\beta},\boldsymbol{\eta},\boldsymbol{\theta})$ be the vector of all nonlatent parameters. Then the complete likelihood decomposes as follows.

$$\begin{aligned} P(\mathbf{S},\mathbf{T},\mathbf{R}|\boldsymbol{\phi}) &= \prod_{i=1}^{N} P(\mathbf{S}^i,\mathbf{T}^i,\mathbf{R}^i|\boldsymbol{\phi}) = \prod_{i=1}^{N} P(\mathbf{T}^i,\mathbf{R}^i|\mathbf{S}^i,\boldsymbol{\phi})P(\mathbf{S}^i|\boldsymbol{\phi}) \\ &= \prod_{i=1}^{N} P(\mathbf{T}^i_1,\mathbf{R}^i_1|\mathbf{S}^i_1,\boldsymbol{\phi})P(\mathbf{S}^i_1|\boldsymbol{\phi}) \prod_{j=2}^{n_i} P(\mathbf{T}^i_j,\mathbf{R}^i_j|\mathbf{S}^i_j,\boldsymbol{\phi})P(\mathbf{S}^i_j|\mathbf{S}^i_{j-1}\boldsymbol{\phi}). \end{aligned}$$

Suppose individual i is treated at visit j^* where $1 < j^* < n_i$. Then we can decompose the ith likelihood term as

$$\begin{aligned} P(\mathbf{S}^i,\mathbf{T}^i,\mathbf{R}^i|\boldsymbol{\phi}) = P(\mathbf{T}^i_1,\mathbf{R}^i_1|\mathbf{S}^i_1,\boldsymbol{\phi})P(\mathbf{S}^i_1|\boldsymbol{\phi}) \prod_{j=2}^{j^*} P(\mathbf{T}^i_j,\mathbf{R}^i_j|\mathbf{S}^i_j,\boldsymbol{\phi})P(\mathbf{S}^i_j|\mathbf{S}^i_{j-1}\boldsymbol{\phi}) \times \dots \\ P(\mathbf{T}^i_{j^*+1},\mathbf{R}^i_{j^*+1}|\mathbf{S}^i_{j^*+1},\boldsymbol{\phi})P(\mathbf{S}^i_{j^*+1}|\mathbf{S}^i_{j^*}\boldsymbol{\phi}) \prod_{j'=j^*+2}^{n_i} P(\mathbf{T}^i_{j'},\mathbf{R}^i_{j'}|\mathbf{S}^i_{j'},\boldsymbol{\phi})P(\mathbf{S}^i_{j'}|\mathbf{S}^i_{j'-1}\boldsymbol{\phi}). \end{aligned}$$

Because it is assumed that treatment is successful at visit j^* we require that $\lim_{\epsilon\downarrow 0} S^i(t^i_{j^*} + \epsilon) = 0$ with probability 1. Moreover, because of the intervention, the Markov property no longer holds between the state of the individual

TABLE 1 Maximum likelihood estimates, $\hat{\theta}^1_s$, of cytology exam result probabilities conditioned on hidden state s

	Cytology result			
State	**N0**	**L1**	**H2**	**C3**
Normal	0.99889 (0.99875, 0.99901)	0.00002 (0.00002, 0.00003)	0.00105 (0.00095, 0.00118)	0.00003 (0.00003, 0.00004)
Low-risk	0.00010 (0.00009, 0.00010)	0.99962 (0.99957, 0.99966)	0.00020 (0.00017, 0.00024)	0.00008 (0.00008, 0.00009)
High-risk	0.10839 (0.09530, 0.12302)	0.03187 (0.02819, 0.03602)	0.85139 (0.83275, 0.86827)	0.00836 (0.00717, 0.00973)
Cancer	0.02677 (0.02538, 0.02823)	0.01307 (0.01285, 0.01330)	0.01212 (0.01201, 0.01224)	0.94804 (0.94629, 0.94971)

TABLE 2 Maximum likelihood estimates, $\hat{\theta}^2_s$, of histology exam result probabilities conditioned on hidden state s

	Histology result			
State	**N0**	**L1**	**H2**	**C3**
Normal	0.97717 (0.97244, 0.98109)	0.01527 (0.01257, 0.01854)	0.00756 (0.00630, 0.00908)	0.00000 (0.00000, 0.00000)
Low-risk	0.50417 (0.49107, 0.51726)	0.48770 (0.47387, 0.50140)	0.00625 (0.00477, 0.00819)	0.00189 (0.00179, 0.00198)
High-risk	0.13404 (0.12749, 0.14088)	0.05534 (0.05228, 0.05857)	0.79083 (0.78094, 0.80039)	0.01979 (0.01860, 0.02105)
Cancer	0.01398 (0.01370, 0.01427)	0.00798 (0.00792, 0.00804)	0.02100 (0.02006, 0.02198)	0.95704 (0.95575, 0.95828)

TABLE 3 Maximum likelihood estimates, $\hat{\theta}^3_s$, of HPV exam result probabilities conditioned on hidden state s

	HPV result	
State	**–**	**+**
Normal	0.91882 (0.89870, 0.93523)	0.08118 (0.06477, 0.10130)
Low-risk	0.48836 (0.45586, 0.52096)	0.51164 (0.47904, 0.54414)
High-risk	0.08290 (0.07100, 0.09657)	0.91710 (0.90343, 0.92900)
Cancer	0.01003 (0.01001, 0.01005)	0.98997 (0.98995, 0.98999)

Abbreviation: HPV, human papillomavirus.

immediately before visit j^* and the state of the individual at visit $j^* + 1$. With these observations we can then replace the term $P(\mathbf{S}^i_{j^*+1}|\mathbf{S}^i_{j^*}\boldsymbol{\phi})$ with $P(\mathbf{S}^i_{j^*+1}|\mathbf{S}^i_{j^*} = 0, \boldsymbol{\phi})$. We have thus decomposed the ith likelihood term $P(\mathbf{S}^i, \mathbf{T}^i, \mathbf{R}^i|\boldsymbol{\phi})$ into two terms which can each be treated as distinct HMM sequences. Applying the forward-backward algorithm to each sequence as if they were conditionally independent HMM sequences allows us to use a standard BW algorithm for maximum likelihood estimation. Note that the above procedure easily generalizes to individuals with more than one treatment.

We used a hard-assignment E-step.[6] Instead of averaging the likelihoods over the posterior marginal distributions of hidden states, a hard-assignment E-step uses the maximum a posteriori hidden state to complete the data for the subsequent M-step. This can lead to faster convergence of the EM algorithm. The M-step is then a standard maximization problem over all other model parameters.

Evaluation of the likelihood becomes expensive for large datasets. This is primarily due to the computation of the transition probability matrix $P(t|s)$ at each screening time. Because the EM algorithm can require a large number of likelihood evaluations, we parallelized the inference process across thousands of processors, utilizing the HPC capabilities of LLNL. The E-step can be parallelized without modification. The M-step was parallelized through the computation of the log-likelihood function. Maximum likelihood estimates presented here are based on 300 000 individual sequences of observations.

The data was distributed among compute clusters with 100 individual sequences per cluster. Through a bootstrap technique[38] we resampled the clusters with replacement to obtain ten bootstrap samples of 300 000 individual sequences. For each of the ten resampled populations we computed maximum likelihood estimates using our modified EM algorithm. From the resulting bootstrapped maximum likelihood estimates we estimated the mean and variance of the maximum likelihood estimator. Again, due to computational costs this approximation was parallelized across many compute

TABLE 4 Maximum likelihood estimates, $\hat{\eta}_s^k$, of Poisson intensities for number of type k tests conditioned on hidden state s

State	Cytology	Histology	HPV
Normal	0.98774 (0.98698, 0.98850)	0.01437 (0.01414, 0.01461)	0.00581 (0.00557, 0.00606)
Low-risk	0.89938 (0.89144, 0.90738)	0.04262 (0.04087, 0.04445)	0.10304 (0.09517, 0.11156)
High-risk	0.49359 (0.49008, 0.49712)	0.59287 (0.58634, 0.59946)	0.05835 (0.05277, 0.06453)
Cancer	0.51046 (0.44966, 0.57947)	0.79600 (0.73522, 0.86181)	0.02481 (0.02332, 0.02640)

Abbreviation: HPV, human papillomavirus.

TABLE 5 Maximum likelihood estimates, $\hat{\pi}$, of the probabilities of being in a particular state at the time of the first screening

Age range	Normal	Low-risk	High-risk	Cancer
(16,19)	0.92992 (0.92615, 0.93350)	0.06703 (0.06374, 0.07048)	0.00283 (0.00251, 0.00318)	0.00022 (0.00022, 0.00023)
(20,24)	0.92882 (0.92567, 0.93181)	0.06273 (0.06053, 0.06500)	0.00833 (0.00747, 0.00928)	0.00013 (0.00012, 0.00013)
(25,29)	0.93408 (0.93169, 0.93639)	0.04945 (0.04785, 0.05109)	0.01632 (0.01555, 0.01713)	0.00016 (0.00015, 0.00016)
(30,34)	0.94918 (0.94570, 0.95243)	0.03554 (0.03365, 0.03754)	0.01506 (0.01363, 0.01664)	0.00021 (0.00021, 0.00022)
(35,39)	0.95263 (0.94935, 0.95566)	0.03250 (0.03098, 0.03408)	0.01464 (0.01305, 0.01641)	0.00024 (0.00024, 0.00025)
(40,49)	0.95551 (0.95264, 0.95821)	0.03303 (0.03126, 0.03491)	0.01131 (0.01034, 0.01237)	0.00015 (0.00014, 0.00015)
(50,59)	0.96314 (0.96009, 0.96593)	0.02797 (0.02625, 0.02981)	0.00860 (0.00750, 0.00987)	0.00028 (0.00028, 0.00029)
60+	0.96047 (0.95710, 0.96356)	0.02747 (0.02575, 0.02930)	0.01168 (0.01026, 0.01328)	0.00038 (0.00038, 0.00039)

nodes. By the asymptotic normality of maximum likelihood estimators, the variance estimates were used to obtain 95% confidence intervals (CI).

Due to the large number of hidden variables in the model, maximum likelihood estimation is a highly non-convex optimization problem. This leads to an inference procedure that is very sensitive to initial conditions as a local optimization procedure was used in the M-step, namely, a limited-memory BFGS method.[39] Informative initial conditions were chosen for the observation error probabilities $\boldsymbol{\theta}_s^k$. Specifically, initial conditions were chosen for the $\boldsymbol{\theta}_s^k$ such that all tests are initialized with high accuracy. Letting $\boldsymbol{\theta}_s^k[\ell]$ be the ℓth element of the parameter vector $\boldsymbol{\theta}_s^k$, with $\ell \in \{0,1,2,3\}$, our initial conditions were such that $\boldsymbol{\theta}_s^k[s] > 0.9$ and $\sum_{\ell \neq s}\boldsymbol{\theta}_s^k[\ell] < 0.1$. Informative initial conditions were also used for very rare events. In particular, since it is very unlikely for a woman to be diagnosed with a high-risk lesion or cancer on a first exam, the initial values for π reflected this. In addition, the censor/death transition intensities were given much smaller initial values than all other transition intensities. The parameter $\theta_0^2[3]$ was fixed at zero, the justification being that a woman truly in a normal state will never be diagnosed with cancer by a histology exam.

5 | MODEL SELECTION

To choose a suitable partition of ages we used the Akaike information criteria (AIC).[40] Models with different age partitions were fit, with the number of age partitions ranging from a minimum of one to a maximum of 16. While an exhaustive

TABLE 6 Maximum likelihood estimates for age-dependent progression intensities, $\hat{\lambda}_{sr}(t)$

	Progression intensities		
Age range	λ_{01}	λ_{12}	λ_{23}
(16,19)	0.02027 (0.01887, 0.02177)	0.01858 (0.01608, 0.02147)	0.00016 (0.00016, 0.00016)
(20,24)	0.01202 (0.01139, 0.01269)	0.02565 (0.02397, 0.02745)	0.00015 (0.00015, 0.00016)
(25,29)	0.00746 (0.00720, 0.00773)	0.03959 (0.03707, 0.04227)	0.00015 (0.00014, 0.00016)
(30,34)	0.00584 (0.00543, 0.00628)	0.04299 (0.03963, 0.04664)	0.00055 (0.00048, 0.00062)
(35,39)	0.00547 (0.00524, 0.00572)	0.03645 (0.03471, 0.03826)	0.00074 (0.00063, 0.00087)
(40,49)	0.00556 (0.00538, 0.00575)	0.02970 (0.02822, 0.03126)	0.00127 (0.00107, 0.00149)
(50,59)	0.00440 (0.00422, 0.00458)	0.02713 (0.02294, 0.03209)	0.00161 (0.00141, 0.00184)
60+	0.00403 (0.00380, 0.00427)	0.03826 (0.03285, 0.04456)	0.00419 (0.00268, 0.00654)

TABLE 7 Maximum likelihood estimates for age-dependent regression intensities, $\hat{\lambda}_{sr}(t)$

	Regression intensities	
Age range	λ_{10}	λ_{21}
(16,19)	0.17558 (0.16438, 0.18755)	0.24261 (0.21317, 0.27612)
(20,24)	0.15543 (0.14972, 0.16135)	0.11439 (0.10677, 0.12257)
(25,29)	0.14622 (0.14191, 0.15066)	0.07753 (0.07327, 0.08204)
(30,34)	0.15576 (0.14621, 0.16594)	0.07372 (0.06799, 0.07995)
(35,39)	0.15805 (0.15125, 0.16514)	0.06958 (0.06369, 0.07602)
(40,49)	0.17165 (0.16498, 0.17859)	0.08370 (0.07832, 0.08946)
(50,59)	0.19910 (0.19187, 0.20660)	0.10237 (0.08486, 0.12348)
60+	0.24198 (0.22838, 0.25638)	0.06951 (0.05268, 0.09172)

search of such partitions is not possible, several candidate models were selected to obtain a best relative model. The age ranges for candidate partitions were either roughly equal in size or had shorter age ranges for younger women and longer age ranges for older women. The nonuniform partitions were motivated by the higher incidence of HPV in younger women (Figure 2).

The lowest AIC score observed among all candidate models was the model with the following eight age partitions: {16-19, 20-24, 25-29, 30-34, 35-39, 40-49, 50-59, 60+}. We will refer to this model as *the AIC-chosen model*. The highest AIC score observed among all candidate models was the model with only one partition. Note that a partition of size one is simply a time-homogeneous model. We will refer to this model as *the time homogeneous model*. AIC results are shown in Figure 3.

Models with more than a single age partition correspond to time-inhomogeneous models. Because the models are all piecewise constant, the time-homogeneous model can be thought of as a constrained version of any one of the time-inhomogeneous models. To test the fit of the time-inhomogeneous models we performed an asymptotic likelihood-ratio test based on Wilk's theorem.[41] For each candidate time-inhomogeneous model, we performed the asymptotic likelihood-ratio test where under the null hypothesis the time-homogeneous model was true while under the alternative hypothesis the candidate time-inhomogeneous model was true. The computed *P*-values for all tests were less than machine precision (approximately 10^{-16}). This is strong evidence that the time-inhomogeneous models fit the data better than the time-homogeneous model. This result, along with the AIC scores, suggests that the development of cervical cancer has a strong dependence on age.

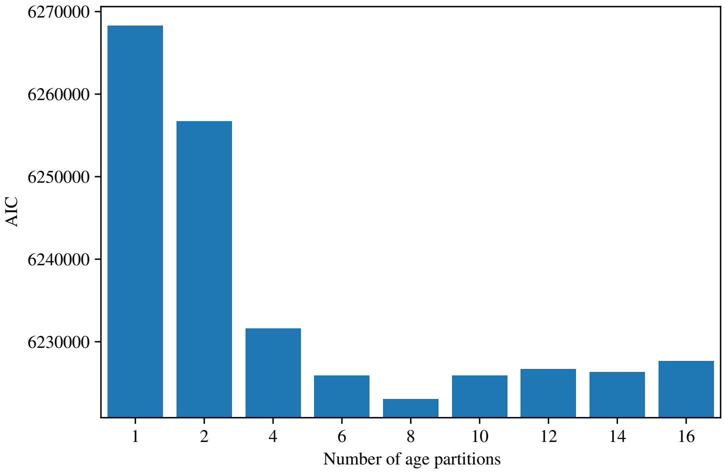

FIGURE 3 AIC values for candidate models arranged with the number of age partitions increasing along the x-axis. Lower AIC values indicate better model fit. Models with more age partitions fit the data better than models with few age partitions, suggesting time-inhomogeneous models are more appropriate. Note that the worst model fit according to AIC is the model with a single age partition, which corresponds to the time-homogeneous model. AIC, Akaike information criteria [Color figure can be viewed at wileyonlinelibrary.com]

6 | MODEL VALIDATION AND REFINEMENT

To validate the models' fit to the observed data, we used Kaplan-Meier estimators as suggested in Titman and Sharples.[42] The Kaplan-Meier estimators were used to diagnose, at a more fine-grained level, which model parameters were not fitting well to the data. To do this we simulate two synthetic screening populations from the time-homogeneous model and the AIC-chosen time-inhomogeneous model, respectively. Because the individual screening times are assumed to be random, the model infers nothing about them and does not provide a way to intermittently sample the underlying Markov processes. In order to randomly screen this synthetic population we draw screening intervals (times between two subsequent exams) from the age-structured empirical distribution of screening intervals. Given these screening times, the types and numbers of screening exams, along with test results, can be simulated from the learned model. Following this procedure for all synthetic Markov chains gives us a synthetic population-level dataset.

In order to simulate trajectories from a Markov chain, the distribution on sojourn times is needed. This is trivial for time-homogeneous Markov models, but more involved in the time-inhomogeneous case. The distribution function for sojourn times from Markov processes with piecewise constant intensities is given below. Let $T_s(a)$ be the amount of time the chain spends in state s given an initial age of a. Given the continuous-time Markov chain on state space $\mathcal{S}$, the quantity $T_s(a) > 0$ is a random variable and the mean sojourn time is $E[T_s(a)]$. The following proposition gives an explicit solution for the cumulative distribution function of the random variable $T_s(a)$. The following notation is used in the statement and proof of the proposition. Fix a and t such that $\tau_0 \leq a < a + t \leq \tau_q$. Let $k, \ell \in \{0,1,2, \dots ,q\}$ be the indices satisfying $\tau_k \leq a < \tau_{k+1}$ and $\tau_\ell \leq a + t < \tau_{\ell+1}$. Let $n = \ell - k$, the number of additional age partitions covered by the time lapse t when starting at age a. For any $s \in \mathcal{S}$ define the values κ_i^s for $i = 0,1, \dots ,n$ as follows:

$$\kappa_0^s = -t \sum_{r \neq s} \lambda_{sr}(\tau_\ell)$$

$$\kappa_1^s = -(\tau_{k+1} - a) \sum_{r \neq s} \lambda_{sr}(\tau_k) - (a - \tau_\ell) \sum_{r \neq s} \lambda_{sr}(\tau_\ell)$$

$$\kappa_m^s = -(\tau_{k+m} - \tau_{k+m-1}) \sum_{r \neq s} \lambda_{sr}(\tau_{k+m-1}) \text{ for } m = 2, 3, \dots , n.$$

Proposition 1. *For any $a > 0$, $s \in \mathcal{S}$, the cumulative distribution function for the sojourn time $T_s(a)$ evaluated at time $t > 0$ is given by $P(T_s(a) < t) = 1 - e^{\sum_{i=0}^{n} \kappa_i^s}$.*

Proof. The case $q = 1$ coincides with a time-homogeneous Markov process. In this case $n = 0$ and the result is obtained from standard results for sojourn times in homogeneous Markov processes. We give a proof for the case that $q > 1$ and $n > 1$. The case $n \leq 1$ follows the same reasoning.

Working with the inverse cumulative distribution function and expanding based on conditional events we have

$$P(T_s(a) > t) = P(T_s(a) > t | T_s(a) > \tau_\ell - a) P(T_s(a) > \tau_{k+1} - a) \prod_{j=0}^{n-2} P(T_s(a) > \tau_{\ell-j} - a | T_s(a) > \tau_{\ell-j-1} - a). \tag{1}$$

By the Markov property, for any ages $x, y \in [\tau_0, \tau_q]$ satisfying $a < x < y$ we have

$$P(T_s(a) > y - a | T_s(a) > x - a) = 1 - P(T_s(x) < y - x).$$

Moreover, if $\tau_i \leq x < y \leq \tau_{i+1}$ for some i, the Markov process is time homogeneous on the time interval $y - x$. It follows that $P(T_s(x) < y - x) = 1 - e^{-(y-x)\sum_{r \neq s} \lambda_{sr}(x)}$. Applying this result to each term in (1), reindexing the terms, and then using the definition of the κ_i^s above gives us $P(T_s(a) > t) = e^{\kappa_0^s + \kappa_1^s + \sum_{i=2}^{n} \kappa_i^s}$, from which the result follows. ■

Proposition 1 allows us to generate random samples from a time-inhomogeneous Markov process with piecewise constant intensities. The following corollary gives the simulation algorithm. The result follows from inverse transform sampling.

Corollary 1. *The following simulates an exit time $T_s^*(a)$ from state $s \in \mathcal{S}$ starting at age $a > 0$ in the above Markov model.*

1. *Draw a random variable $u \sim \text{Uniform(0,1)}$.*
2. *Find the k such that $\tau_k \leq a < \tau_{k+1}$.*
3. *Find the ℓ such that $P(T_s(a) < \tau_\ell - a) < u < P(T_s(a) < \tau_{\ell+1} - a)$.*
4. *Then $T_s^*(a)$ is a random exit time from state s when starting at age a, where*

$$T_s^*(a) = \frac{\sum_{i=1}^{\ell-k} \kappa_i^s - \log(1-u)}{\sum_{r \neq s} \lambda_{sr}(\tau_\ell)}.$$

Proof. By inverse transform sampling we draw u from a uniform distribution and find the t that solves $u = P(T_s(a) < t)$. Let ℓ be the integer satisfying $P(T_s(a) < \tau_\ell - a) < u < P(T_s(a) < \tau_{\ell+1} - a)$. From Proposition 1 we know there is an $n(a,t)$ and corresponding $\kappa_i^s(a, t)$ that depend on a,t in such a way that $P(T_s(a) < t) = 1 - e^{\sum_{i=0}^{n(a,t)} \kappa_i^s(a,t)}$. From this we have $\log(1 - u) = \sum_{i=0}^{n(a,t)} \kappa_i^s(a, t)$. Note that $n(a,t) = n(a,s)$ for any $s \in [\tau_\ell - a, t]$. Similarly for $i = 1,2, \dots ,n(a,t)$, $\kappa_i^s(a, t) = \kappa_i^s(a, t')$ for any $t' \in [\tau_\ell - a, t]$. In particular $n(a, t) = n(a, \tau_\ell - a)$ and $\kappa_i^s(a, t) = \kappa_i^s(a, \tau_\ell - a)$ for $i = 1,2, \dots ,n(a,t)$. We can then write

$$\log(1-u) = \kappa_0^s(a,t) + \sum_{i=1}^{n(a,t)} \kappa_i^s(a,t) = -t \sum_{r \neq s} \lambda_{sr}(\tau_\ell) + \sum_{i=1}^{n(a,\tau_\ell - a)} \kappa_i^s(a, \tau_\ell - a).$$

Solving for t gives us $t = \frac{\sum_{i=1}^{n(a,\tau_\ell - a)} \kappa_i^s(a,\tau_\ell) - \log(1-u)}{\sum_{r \neq s} \lambda_{sr}(\tau_\ell)}$. ■

6.1 | Comparing synthetic data and observed data by Kaplan-Meier curves

Given the two simulated datasets and a hold-out dataset of equal size (300 000 women) not used to train the model, we computed Kaplan-Meier curves as described in Titman and Sharples.[42] Note that these are not true Kaplan-Meier estimators. Because the true underlying states are hidden, we cannot compute Kaplan-Meier estimators from the observed data. Instead these curves are heuristic tools for the visual inspection of model fit. The events for which we computed Kaplan-Meier curves are normal to low-risk ($N0 \rightarrow L1$), normal to high-risk ($N0 \rightarrow H2$), normal to cancer ($N0 \rightarrow C3$), and low-risk to normal ($L1 \rightarrow N0$). Individual Kaplan-Meier estimators were computed for each age cohort corresponding to the AIC-chosen time-inhomogeneous model. These Kaplan-Meier plots can be found in the supplementary material.

6.2 | Improving the model fit by creating a mixed model

While AIC scores and likelihood-ratio tests provide evidence for goodness-of-fit and criteria for model selection, they may not be sufficient tools for all applications. In our case, because the data is imbalanced toward normal and low-risk exam results, the likelihood is dominated by these data. This means that when maximizing the likelihood, the normal, and low-risk exam results will overwhelmingly dominate the inference process. For convex likelihood functions such imbalanced data is not a problem. Because there is a unique global maximum, with enough iterations the maximum likelihood estimate will be reached (to any desired accuracy). But when the likelihood is non-convex, this is not the case. In latent variable models, such as HMMs, the likelihood function is non-convex. Therefore, maximum likelihood estimation is a non-convex optimization problem with multiple local optima. Therefore, the EM algorithm is not guaranteed to find the global optimum but only a local optimum.

This combination of a highly non-convex likelihood and highly imbalanced data can lead to poor fitting of parameters associated with rarely observed events. For example, because the data is so imbalanced, the normal and low-risk exam results can dominate the inference process early on, taking the parameter estimates to a region in parameter space with a local optima that does not include the "best" parameters related to the high-grade and cancer exam results. The optimization process essentially gets stuck in such regions and is prevented from improving on the parameters associated with rare events. In such cases it is possible that parameters associated with high grade and cancer states will be harder to learn. This is particularly problematic in medical applications where the sparse data are often the most informative and of the most interest.

Evidence of such complexities were present in our application. In most cases, the AIC-chosen Kaplan-Meier estimators showed better overall fit to the empirical Kaplan-Meier estimators compared with the time-homogeneous Kaplan-Meier estimators. However, despite the strong evidence favoring time-inhomogeneous models, there were cases where the time-homogeneous Kaplan-Meier estimators had a better fit. Specifically, this occurred in several of the normal to cancer ($N0 \rightarrow C3$) transitions. This is likely due to difficulties with the maximization of the likelihood resulting from highly imbalanced data and a non-convex likelihood. One common method for dealing with non-convex optimization problems is using multiple random initializations of the parameters. This methodology is of limited use when the optimization procedure is very expensive, as in our case.

Instead we simply replaced certain transition intensities in the AIC-chosen model with the corresponding transition intensities from the time-homogeneous model. The intensity λ_{23} governs the transitions $H2 \rightarrow C3$, thus we focused on this set of parameters. In the AIC-chosen model, the transition intensity λ_{23} is defined by eight different parameters, while in the time-homogeneous model it is defined by only one parameter. By setting the eight parameters equal to that of the single parameter from the time-homogeneous model, we constructed a "mixed" model that is time-inhomogeneous, but in a nonuniform way. Namely, one of the parameters, λ_{23}, is age independent.

The marginal likelihood and AIC for the resulting mixed model were both improved from that of the AIC-chosen model. This was evidence that difficulties in the optimization procedure were responsible for the poor fit. We next generated synthetic data from this mixed model and computed Kaplan-Meier estimators to compare to our previous results.

TABLE 8 Maximum likelihood estimates for age-dependent transition intensities, $\hat{\lambda}_{sr}(t)$

	Death/censor intensities			
Age range	$\boldsymbol{\lambda_{04}}$	$\boldsymbol{\lambda_{14}}$	$\boldsymbol{\lambda_{24}}$	$\boldsymbol{\lambda_{34}}$
(16,19)	0.00002 (0.00002, 0.00002)	0.00014 (0.00014, 0.00015)	0.00222 (0.00217, 0.00228)	0.01817 (0.01817, 0.01817)
(20,24)	0.00006 (0.00005, 0.00007)	0.00016 (0.00015, 0.00017)	0.00082 (0.00073, 0.00091)	0.03124 (0.03057, 0.03192)
(25,29)	0.00012 (0.00011, 0.00013)	0.00019 (0.00018, 0.00021)	0.00163 (0.00143, 0.00185)	0.03464 (0.03422, 0.03507)
(30,34)	0.00012 (0.00011, 0.00014)	0.00016 (0.00016, 0.00016)	0.00273 (0.00233, 0.00321)	0.04211 (0.03940, 0.04502)
(35,39)	0.00010 (0.00009, 0.00011)	0.00015 (0.00014, 0.00015)	0.00398 (0.00361, 0.00440)	0.04110 (0.03845, 0.04394)
(40,49)	0.00010 (0.00009, 0.00011)	0.00014 (0.00013, 0.00014)	0.00518 (0.00447, 0.00602)	0.03170 (0.02872, 0.03498)
(50,59)	0.00020 (0.00019, 0.00020)	0.00028 (0.00027, 0.00029)	0.00618 (0.00540, 0.00708)	0.02772 (0.02381, 0.03228)
60+	0.00105 (0.00103, 0.00107)	0.00115 (0.00112, 0.00119)	0.02124 (0.01924, 0.02345)	0.02386 (0.02091, 0.02723)

An improvement in the fit for the older age cohorts in the $N0 \rightarrow C3$ transitions was observed. However, the fit was not improved for the younger age cohorts. We then went back to some alternate models we had previously discarded when doing model selection and generated Kaplan-Meier plots for all of them. Looking through the results we saw a particular model that had a very good fit for the younger age cohorts, but not the older age cohorts, for the $N0 \rightarrow C3$ Kaplan-Meier curves. This model divided ages into the three cohorts {16-29,30-69,70 + }. We again exchanged the intensity parameters from this model into the current mixed model, but only for the partitions including ages 16 to 29, in the parameter λ_{23}. The resulting Kaplan-Meier estimators for these initial mixed models can be found in the supplementary material.

The transition intensity λ_{23} was now age dependent, but to a lesser extent than the original AIC-chosen time-inhomogeneous model. This latest mixed model again showed improvements in the likelihood, AIC and Kaplan-Meier fits. However, the resulting parameters were no longer guaranteed to be maximum likelihood estimates. We then used this set of parameters as the initialization of a new EM algorithm to obtain a new set of maximum likelihood estimates. The resulting model is referred to as the *mixed time-inhomogeneous model*. The resulting maximum likelihood estimates along with their associated 95% CIs can be found in Tables 1 to 8. The model generated Kaplan-Meier estimates can be found in Figures 4 to 7. We briefly discuss these results below.

7 | RESULTS FROM THE MIXED MODEL

Tables 1 to 3 show the maximum likelihood estimates, $\hat{\theta}_s^k$, of the probabilities of exam results conditioned on hidden state. Concordance between underlying hidden states and diagnostic categories was high for cytology, ranging between 85.14% (95% CI 83.28-86.83) and 99.96% (95% CI 99.96-99.97) (Table 1). For histology, the concordance was high for the normal and cancer state, 97.71% (95% CI 97.24-98.11) and 95.70% (95% CI 95.57-95.82), respectively (Table 2). However, only 48.77% (95% CI 47.77-50.14) of observed histology diagnoses indicative of low-risk state were in concordance with the model's predicted unobserved low-risk state. With 91.71% (95% CI 90.34-92.90) for high-risk, and 98.997% (95% CI 98.995-98.999) for cancer state (Table 3), a positive HPV test had a higher concordance with these hidden states than either cytology or histology.

Table 4 shows maximum likelihood estimates, $\hat{\eta}_s^k$, of Poisson intensities for the number of tests conditioned on the underlying hidden state. Higher values indicate a higher probability of having the given test, conditioned on the true hidden state. The probability of having a cytology exam decreases as the disease state increases, while the probability of having a histology exam increases as the disease state increases. The probability of having an HPV test is highest for abnormal disease states and lowest for the normal state.

Table 5 shows the maximum likelihood estimates, $\hat{\pi}$, of the probabilities of being in a particular state at the time of the first screening. Regardless of age the probability of being in the normal state at first screening, which ranged between 92.99% (95% CI 92.62-93.35) and 96.31% (95% CI 96.01, 96.59), was the most common outcome with higher probabilities with increasing age. On the other hand, the probability of being in the low-risk state decreased with age, being highest in the youngest age group at 6.703% (95% CI 6.374-7.048) and lowest in the oldest age-group at 2.747% (95% CI 2.575-2.930). The probability of being in the high-risk state at the first screening was highest for women between 25 and 40 years of age, ranging between 1.464% (95% CI 1.305-1.641) and 1.632% (95% CI 1.555-1.713). The cancer state was rare and evenly distributed between age-groups, ranging between 0.013% (95% CI 0.012-0.013) and 0.038% (95% CI 0.038-0.039).

The transition intensity for progressing from the underlying state normal to low-risk was highest for women below 20 years of age at 0.020 (95% CI 0.0189-0.0218) (Table 6). However, the transition intensity from low-risk to high-risk was lowest for this same age group at 0.0186 (95% CI 0.0161-0.0215). The transition intensity from low-risk to high-risk peaked in the age range of 30 to 34 years at 0.0430 (95% CI 0.0396-0.0466), while the transition intensity from high-risk to cancer peaked among women above 60-years-old at 0.004 (95% CI 0.002-0.006).

The transition intensities for regressing from low-risk to normal are among the highest observed across all age ranges and are the highest among women under 20 or above 50 years of age (Table 7). The transition intensity for regressing from high-risk to low-risk was found to be highest among women below 20 years of age and lowest among women over 60 years of age.

Figures 4 to 7 show the Kaplan-Meier curves for normal to low-risk ($N0 \rightarrow L1$, Figure 4), normal to high-risk ($N0 \rightarrow H2$, Figure 5), normal to cancer ($N0 \rightarrow C3$, Figure 6), and low-risk to normal ($L1 \rightarrow N0$, Figure 7) for the observed hold-out data, time-homogeneous model and the mixed time-inhomogeneous model. Individual Kaplan-Meier estimators are computed for each age cohort corresponding to the AIC-chosen time-inhomogeneous model(Figure 3). Note that

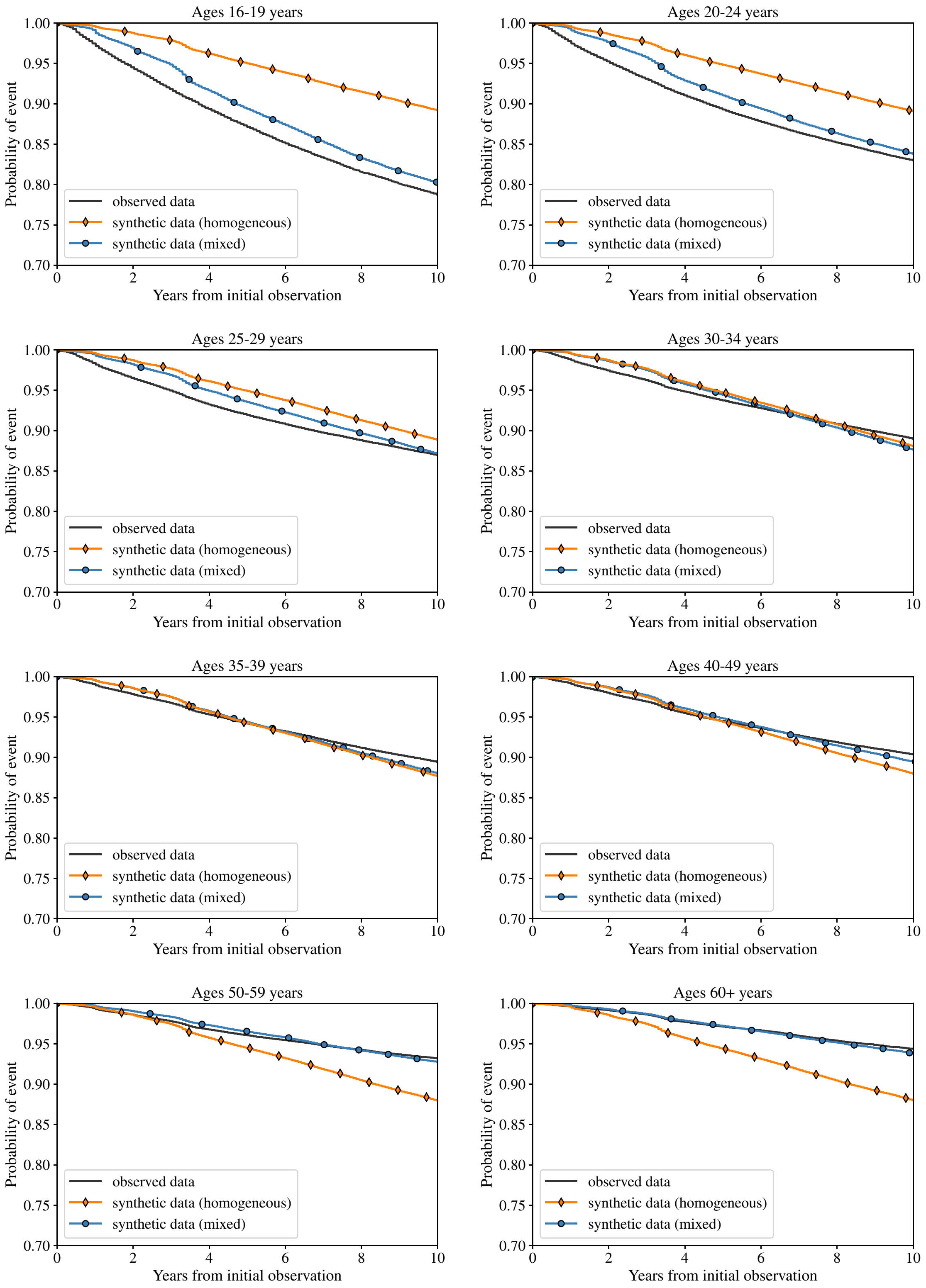

FIGURE 4 Kaplan-Meier estimators with an event defined by an observed normal exam result followed by an observed low-risk exam result. Three datasets were used: One hold out test set from the observed data (black), and two synthetic datasets generated from the time-homogeneous model (orange diamonds) and the mixed time-inhomogeneous model (blue circles) [Color figure can be viewed at wileyonlinelibrary.com]

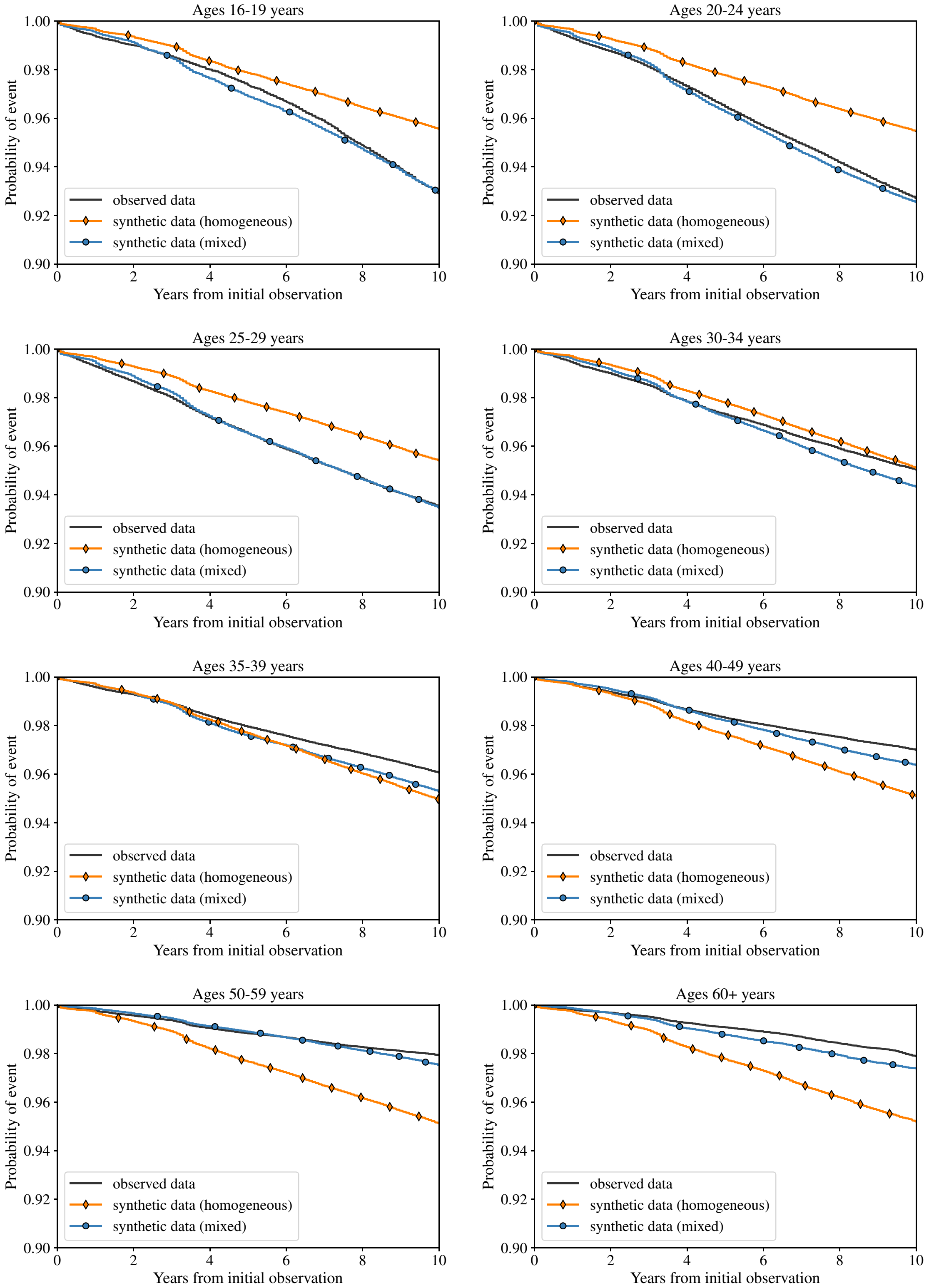

FIGURE 5 Kaplan-Meier estimators with an event defined by an observed normal exam result followed by an observed high-risk exam result. Three datasets were used: One hold out test set from the observed data (black), and two synthetic datasets generated from the time-homogeneous model (orange diamonds) and the mixed time-inhomogeneous model (blue circles) [Color figure can be viewed at wileyonlinelibrary.com]

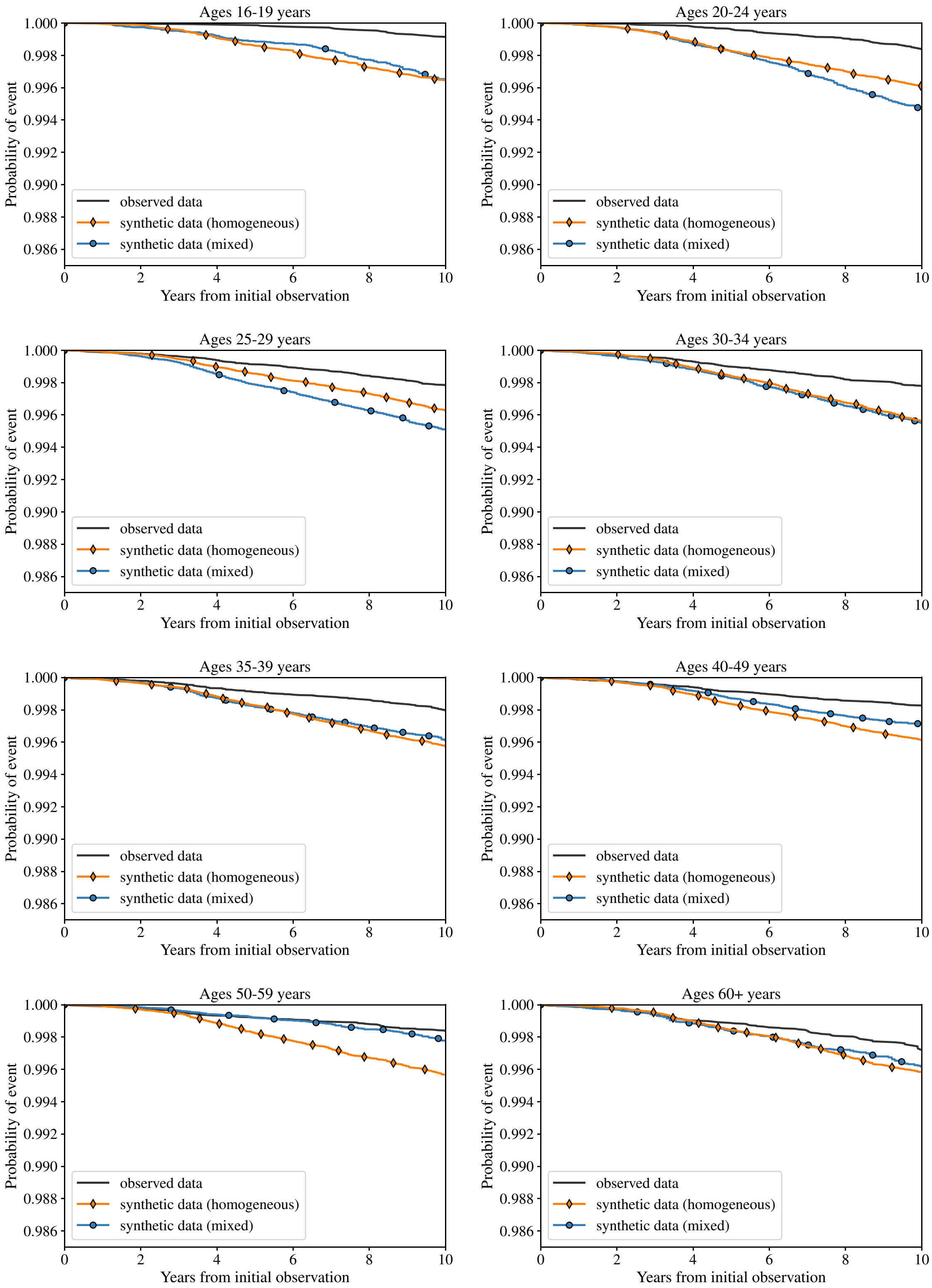

FIGURE 6 Kaplan-Meier estimators with an event defined by an observed normal exam result followed by an observed cancer exam result. Three datasets were used: One hold out test set from the observed data (black), and two synthetic datasets generated from the time-homogeneous model (orange diamonds) and the mixed time-inhomogeneous model (blue circles) [Color figure can be viewed at wileyonlinelibrary.com]

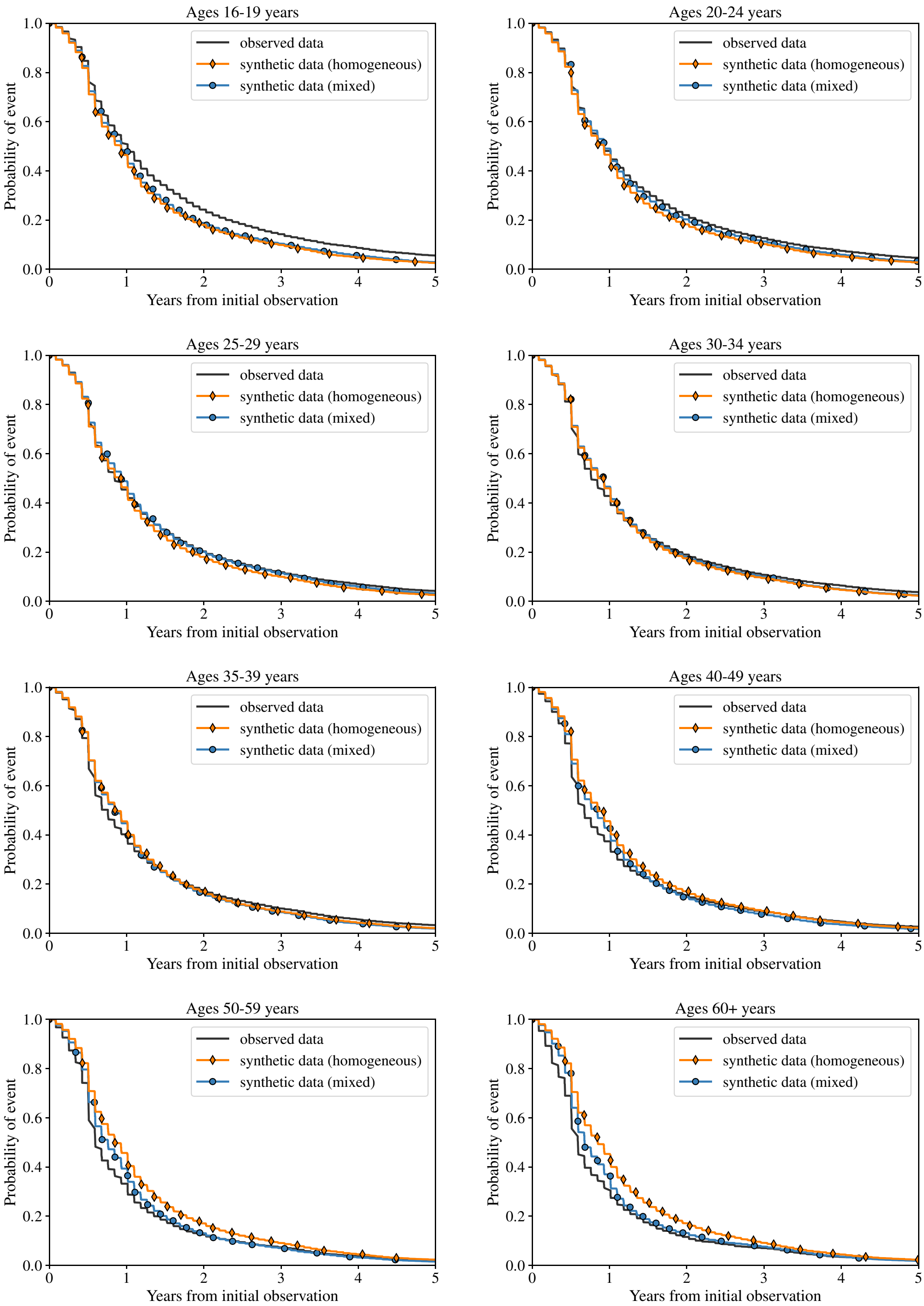

FIGURE 7 Kaplan-Meier estimators with an event defined by an observed low-risk exam result followed by an observed normal exam result. Three datasets were used: One hold out test set from the observed data (black), and two synthetic datasets generated from the time-homogeneous model (orange diamonds) and the mixed time-inhomogeneous model (blue circles) [Color figure can be viewed at wileyonlinelibrary.com]

the scale on the y-axis is different in each series of figures, and some of the scales are quite small in order to better illustrate the differences in model fit.

In the Kaplan-Meier curves for progression from normal to low-risk ($N0 \rightarrow L1$), the time-homogeneous model underestimates the failures among younger women and overestimates failures among older women, and clearly has a worse fit than the mixed time-inhomogeneous model (Figure 4). The mixed time-inhomogeneous model slightly overestimates the failures for the youngest cohort, but has an overall very good fit with the observed data for up to 7 years after the initial observation. In Figure 5 the Kaplan-Meier curves for a normal to high-risk ($N0 \rightarrow H2$) observation is shown. Similar to Figure 4, the time-homogeneous model has a poorer fit than the time-inhomogeneous model. Although the time-inhomogeneous model has a relatively good fit across all age groups, it consistently slightly overestimates the failure rate. In the Kaplan-Meier curves for normal to cancer ($N0 \rightarrow C3$), both models slightly overestimated the risk of being diagnosed with cancer for all age groups (Figure 6). The mixed time-inhomogeneous model fit the data as well as or better than the time-homogeneous model in all age groups other than the ages between 20 and 29, in which case it overestimated the risk slightly more.

For all Kaplan-Meier curves for progressive events (observations from less severe exam results to more severe exam results), the probabilities of the events occurring is small with less than 8% of the population having experienced the event within 10 years from the initial observation. In Figure 7, the Kaplan-Meier curves for the regression from low-risk to normal ($L1 \rightarrow N0$) is shown. The overall fit is good for both the time-homogeneous and time-inhomogeneous models, suggesting that regression is not dependent on age. A regression event is much less rare than the progression events. Five years after the initial observation almost the entire population has had a normal test result.

8 | DISCUSSION

Several studies have been published presenting Markov models representing the natural history of cervical cancer.[18,19] While these studies present similar Markov models for disease history as ours, they do not perform statistical inference over individually observed real-world screening data. Instead, the Markov models are used as stochastic computational models which are calibrated against global population statistics. To the best of our knowledge this article is the first to use actual population-level cancer screening data to learn all model parameters of a HMM at scale.

The HMM presented in this article was developed based on screening technology applied in Norway from 1992 as part of a national cervical cancer screening program. HPV testing was incorporated in 2005 to facilitate the follow-up of equivocal cytology results.[43] The model was constructed based on the understanding of the screening process as an attempt to classify each woman into one of four states on the following cervical cancer risk scale: (i) a normal state requiring no near-term clinical action; (ii) a low-risk state requiring near-term follow-up tests; (iii) a high-risk state requiring immediate treatment; and finally, (iv) a cancer state considered a failure of the screening program. We consider these four states as hidden, indirectly observed through various medical exams, namely, cytological, histological, and HPV exams. Using the number and results of each of the three exams as observed data in a HMM allows inference on the true risk state of the woman being screened.

The fact that cervical cancer risk is associated with HPV exposure and sexual behavior suggests the need to develop time-inhomogeneous models. The transition intensities were modeled as piecewise constant functions of the woman's age. To determine a suitable partition of ages we used AIC, which allowed the selection of a relatively optimal partition of eight partitions. Stratifying the population according to this partition required a relatively large number of parameters. Access to the entire Norwegian screening population provided sufficient data needed for inference and ensured that all possible exam results were observed across all age groups. This large data requirement, combined with the model's complexity, presented technical challenges to obtaining maximum likelihood estimates and standard errors. Parallel computation approaches were necessary to perform full model inference.

We have shown through the use of Kaplan-Meier estimators that the time-inhomogeneous models fit the data better than the time-homogeneous model. This supports the proposition that the development of cervical cancer is a highly age-dependent process. However, the transition between the more severe states, specifically from $H2 \rightarrow C3$, did not show the same degree of age-dependence based on the Kaplan-Meier estimators. By combining parameter estimates from both a time-homogeneous model and two time-inhomogeneous models, we initialized a final run of the EM algorithm to obtain a mixed time-inhomogeneous model. This final mixed time-inhomogeneous model was chosen as the best overall fit to the data. It had the largest marginal likelihood, smallest AIC score and best overall Kaplan-Meier fits.

For each screening exam, cytology, histology, and HPV-testing, the model estimated the probabilities of observable results conditioned on the underlying unobserved state. It is repeatedly described in the literature that a single cytology exam has only 50% to 60% sensitivity to diagnose a precancer, and that the sensitivity increases to 82% if cytology was used to follow-up slightly abnormal results.[44,45] Similarly, it is well documented in large randomized clinical trials that HPV tests are more sensitive than cytological exams for detecting precancers and cancers in screening,[46-48] and international screening guidelines now recommend HPV test for women below 30 years of age.[49] Our estimated model parameters reflect comparable results by showing a higher conditional probability of correctly identifying a high-risk state for HPV of 94%, compared with cytology's 83.5%.

Estimation of the these conditional probabilities show very small misclassification errors for cytological exams. The vast majority of the data consists of cytological exams, and thus the likelihood function will be heavily influenced by this data. The maximum likelihood estimates will then be biased toward low errors in the cytological exams. More detailed information on screening exam accuracy can be built into the model through the use of bounded parameter spaces or informative prior distributions.

Although it is well known that the histology exam is prone to sampling errors, it was somewhat surprising to observe that for each defined state, histology exams failed to demonstrate the highest accuracy. Of note, histology is an invasive procedure, is performed less often as compared with Cytology and HPV, and only upon clinical indication. Of note, both HPV and Histology concurred for about 50% of the model predicted low-risk state, compared with 99% of cytology.[50,51] One explanation for this result is that a biopsy might miss the area with the low-grade dysplasia, while the cytological exam will sample a more representative set of cervical cells.[52]

While it is generally accepted that cancer develops through multiple stages,[53,54] it is difficult in empirical studies to estimate the transmission probabilities between different disease states. Our model provides insight to the transitions between stages through the Markov transition intensities. The low-risk state is considered a transitional state, the immediate sequelae of new HPV infections.[55] According to the model it was most frequent at first screening in younger women. It is well in line with the observation that HPV infections decline with age, and new infections typically do not progress to precancers in older women.[56] Furthermore, the model reflected the known age-distribution of high-risk states, which was the most common state at first screenings for women aged 25 to 40. They had the highest progression intensity between low-risk and high-risk states, which is in line with what is known about the epidemiology of precancerous lesions in Norway.[57]

The model estimated higher regression intensities from high-risk to low-risk for women below 20 years of age, while age specific differences were less expressed for regression from low-risk to normal. Comparing progression and regression intensities for all ages, the regression intensity from low-risk to normal was always higher than progression from normal to low-risk. Similarly, the regression intensity from high-risk to low-risk was always higher than progression from low-risk to high-risk state. This is in line with the empirical observations.[55] It has been estimated that 20% to 30% of precancers will regress, while our data suggest higher precancer regression intensities for women below 20 years of age (0.2426), with similar intensities for women between 25 and 60 years of age, and a lower intensity for those 60 years and older (0.0695).

In the current article, we focus on a woman's age as a proxy for HPV infection, but other individual covariates can be incorporated similarly when available. Having access to more complete HPV data could possibly improve the predictive performance of such a model. There are at least two ways our model could be extended to utilize complete HPV data. The first is to simply use the HPV data as we do here, as a screening exam that is informative of the underlying state of the woman. Having more observed HPV results does not change the model, it simply provides more information at each screening visit. This should further improve inference for the underlying hidden states, thus providing better inference on the disease dynamics.

A second approach would be to expand the state space of the Markov model to include HPV positive and negative versions of the current Markov state space. For example, we could consider normal-HPV-negative, normal-HPV-positive, low-risk-HPV-negative, low-risk-HPV-positive, and so on. The topology of the Markov chain could then be extended to reflect the interaction of HPV and cervical cancer, similar to the approaches taken in Myers et al[18] and Malagón et al.[19]

Learning a one-size-fits-all model from the entire screening population can have real consequences. For example, from Figures 4 to 5 we see that the time-homogeneous model coincides with the time-inhomogeneous model, and fits the data best, in the age ranges that are near the mode of the empirical age distribution (Figure 2). In other words, the most populace cohorts dominate model inference. As a result, when we stratify the population by age to examine model fit we see that the time-homogenous model overestimates the transition rates for younger women and underestimates the transition rates for older women for the progressive transitions ($H0 \rightarrow L1$ and $L1 \rightarrow H2$). Current screening guidelines, with uniform fixed screening interval for all participants, may lead to under-screening of high-risk younger women and overscreening of low-risk older women. The model presented in this study shows potential for improving strategies for controlling cancer by personalizing screening interval recommendations by utilizing both an individual's age and screening history.

ACKNOWLEDGEMENTS

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-JRNL-773737.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available upon request from the Cancer Registry of Norway pursuant the legal requirements mandated by the European GDPR, Article 6 and 9. The data are not publicly available due to privacy and ethical restrictions.

ORCID

Braden C. Soper https://orcid.org/0000-0002-0308-708X
Mari Nygård https://orcid.org/0000-0002-4100-4855

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68(6):394-424. https://doi.org/10.3322/caac.21492.
2. Vaccarella S, Franceschi S, Engholm G, Lonnberg S, Khan S, Bray F. 50 years of screening in the Nordic countries: quantifying the effects on cervical cancer incidence. *Br J Cancer*. 2014;111(5):965-969. https://doi.org/10.1038/bjc.2014.362.
3. Saslow D, Castle PE, Cox JT, et al. American cancer society guideline for human papillomavirus (HPV) vaccine use to prevent cervical cancer and its precursors. *CA Cancer J Clin*. 2007;57(1):7-28.
4. Saslow D, Solomon D, Lawson HW, et al. American cancer society, American society for colposcopy and cervical pathology, and American society for clinical Pathology screening guidelines for the prevention and early detection of cervical cancer. *CA Cancer J Clin*. 2012;62(3):147-172. https://doi.org/10.3322/caac.21139.
5. Pedersen K, Burger EA, Nygard M, Kristiansen IS, Kim JJ. Adapting cervical cancer screening for women vaccinated against human papillomavirus infections: the value of stratifying guidelines. *Eur J Cancer*. 2018;91:68-75. https://doi.org/10.1016/j.ejca.2017.12.018.
6. Koller D, Friedman N. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. Cambridge, Massachusetts: The MIT Press; 2009.
7. Kay R. A Markov model for analysing cancer markers and disease states in survival studies. *Biometrics*. 1986;42(4):855-865.
8. Conlon ASC, Taylor JMG, Sargent DJ. Multi-state models for colon cancer recurrence and death with a cured fraction. *Stat Med*. 2014;33(10):1750-1766. https://doi.org/10.1002/sim.6056.
9. Santamaría C, García-Mora B, Rubio G, Navarro E. A Markov model for analyzing the evolution of bladder carcinoma. *Math Comput Model*. 2009;50(5-6):726-732. https://doi.org/10.1016/j.mcm.2008.12.019.
10. Dancourt V, Quantin C, Abrahamowicz M, Binquet C, Alioum A, Faivre J. Modeling recurrence in colorectal cancer. *J Clin Epidemiol*. 2004;57(3):243-251. https://doi.org/10.1016/j.jclinepi.2003.07.012.
11. García-Mora B, Santamaría C, Navarro E, Rubio G. Modeling bladder cancer using a Markov process with multiple absorbing states. *Math Comput Model*. 2010;52(7-8):977-982. https://doi.org/10.1016/j.mcm.2010.02.036.
12. Bureau A, Shiboski S, Hughes JP. Applications of continuous time hidden Markov models to the study of misclassified disease outcomes. *Stat Med*. 2003;22:441-462.
13. Chen H. H., Duffy S. W., Tabar Laszlo. A Markov chain method to estimate the tumour progression rate from preclinical to clinical phase, sensitivity and positive predictive value for mammography in breast cancer screening. *J Royal Stat Soc Ser D (Stat)*. 1996;45(3):307–317.
14. Duffy SW, Chen H-H, Tabar L. Estimation of mean sojourn time in breast cancer screening using markov chain model of both entry to and exit from the preclinical detectable phase. *Stat Med*. 1995;14:1531-1543.
15. Uhry Z, Hédelin G, Colonna M, et al. Multi-state Markov models in cancer screening evaluation: a brief review and case study. *Stat Methods Med Res*. 2010;19:463-486.
16. Yen AM, Chen TH, Duffy SW, Chen C-D. Incorporating frailty in a multi-state model: application to disease natural history modelling of adenoma-carcinoma in the large bowel. *Stat Methods Med Res*. 2010;19(5):529-546. https://doi.org/10.1177/0962280209359862.
17. Kirby AJ, Spiegelhalter DJ. Modeling the precursors of cervical cancer. In: Lange N, ed. *Case Studies in Biometry*. New York, NY: Wiley; 1994.
18. Myers ER, McCrory DC, Nanda K, Bastian L, Matchar DB. Mathematical model for the natural history of human papillomavirus infection and cervical carcinogenesis. *Am J Epidemiol*. 2000;151(12):1158-1171. https://doi.org/10.1093/oxfordjournals.aje.a010166.
19. Malagón T, Kulasingam S, Mayrand MH, et al. Age at last screening and remaining lifetime risk of cervical cancer in older, unvaccinated, HPV-negative women: a modelling study. *Lancet Oncol*. 2018;19(12):1569-1578. https://doi.org/10.1016/s1470-2045(18)30536-9.

20. Felix JC, Lacey MJ, Miller JD, Lenhart GM, Spitzer M, Kulkarni R. The clinical and economic benefits of co-testing versus primary HPV testing for cervical cancer screening: a modeling analysis. *J Womens Health (Larchmt)*. 2016;25(6):606-616. https://doi.org/10.1089/jwh.2015.5708.
21. Petry KU, Barth C, Wasem J, Neumann A. A model to evaluate the costs and clinical effectiveness of human papilloma virus screening compared with annual papanicolaou cytology in Germany. *Eur J Obstet Gynecol Reprod Biol*. 2017;212:132-139. https://doi.org/10.1016/j.ejogrb.2017.03.029.
22. Viscondi JYK, Faustino CG, Campolina AG, Itria A, Soarez PC. Simple but not simpler: a systematic review of Markov models for economic evaluation of cervical cancer screening. *Clinics (Sao Paulo)*. 2018;73:e385. https://doi.org/10.6061/clinics/2018/e385.
23. Nygård JF, Skare GB, Thoresen SØ. The cervical cancer screening programme in Norway, 1992-2000: changes in Pap smear coverage and incidence of cervical cancer. *J Med Screen*. 2002;9(2):86-91. http://dx.doi.org/10.1136/jms.9.2.86.
24. Lönnberg S, Hansen BT, Haldorsen T, Campbell S, Schee K, Nygård M. Cervical cancer prevented by screening: Long-term incidence trends by morphology in Norway. *Int J Cancer*. 2015;137(7):1758-1764. https://dx.doi.org/10.1002/ijc.29541.
25. Ursin G, Sen S, Mottu JM, Nygard M. Protecting privacy in large datasets-first we assess the risk; then we fuzzy the data. *Cancer Epidemiol Biomark Prev*. 2017;26(8):1219-1224. https://doi.org/10.1158/1055-9965.EPI-17-0172.
26. Schiffman M, Castle PE, Jeronimo J, Rodriguez AC, Wacholder S. Human papillomavirus and cervical cancer. *Lancet*. 2007;370(9590):890-907. https://doi.org/10.1016/S0140-6736(07)61416-0.
27. Kaiser J. 'Liquid biopsy' for cancer promises early detection. *Science*. 2018;359(6373):259. https://doi.org/10.1126/science.359.6373.259.
28. Vargas AJ, Harris CC. Biomarker development in the precision medicine era: lung cancer as a case study. *Nat Rev Cancer*. 2016;16(8):525-537. https://doi.org/10.1038/nrc.2016.56.
29. Moscicki AB, Schiffman M, Kjaer S, Villa LL. Chapter 5: Updating the natural history of HPV and anogenital cancer. *Vaccine*. 2006;24(Suppl 3):S3/42-S3/51. https://doi.org/10.1016/j.vaccine.2006.06.018.
30. Cohen PA, Jhingran A, Oaknin A, Denny L. Cervical cancer. *Lancet*. 2019;393(10167):169-182. https://doi.org/10.1016/S0140-6736(18)32470-X.
31. Al-Mohy AH, Higham NJ. A new scaling and squaring algorithm for the matrix exponential. *SIAM J Matrix Anal Appl*. 2010;31(3):970-989. https://doi.org/10.1137/09074721x.
32. Mirabello L, Clarke MA, Nelson CW, et al. The intersection of HPV epidemiology, genomics and mechanistic studies of HPV-mediated carcinogenesis. *Viruses*. 2018;10(2):80. https://doi.org/10.3390/v10020080.
33. Schiffman M, Wentzensen N. Human papillomavirus infection and the multistage carcinogenesis of cervical cancer. *Cancer Epidemiol Biomark Prev*. 2013;22(4):553-560. https://doi.org/10.1158/1055-9965.EPI-12-1406.
34. Vintermyr OK, Andersland MS, Bjorge T, et al. Human papillomavirus type specific risk of progression and remission during long-term follow-up of equivocal and low-grade HPV-positive cervical smears. *Int J Cancer*. 2018;143(4):851-860. https://doi.org/10.1002/ijc.31390.
35. Lissouba P, Perre P, Auvert B. Association of genital human papillomavirus infection with HIV acquisition: a systematic review and meta-analysis. *Sex Transm Infect*. 2013;89(5):350-356. https://doi.org/10.1136/sextrans-2011-050346.
36. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J Royal Stat Soc Ser B (Methodol)*. 1977;39(1):1-38.
37. Baum LE, Petrie T, Soules G, Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov Chains. *Ann Math Stat*. 1970;41(1):164-171.
38. Efron B. *Bootstrap Methods: Another Look at the Jackknife*. New York, NY: Springer; 1992:569-593.
39. Byrd RH, Lu P, Nocedal J, Zhu C. A limited memory algorithm for bound constrained optimization. *SIAM J Sci Comput*. 1995;16(5):1190-1208. https://doi.org/10.1137/0916069.
40. Akaike H. *Information Theory and an Extension of the Maximum Likelihood Principle*. New York, NY: Springer; 1973:199-213.
41. Titman AC, Sharples LD. Model diagnostics for multi-state models. *Stat Methods Med Res*. 2009;19(6):621-651. https://doi.org/10.1177/0962280209105541.
42. Titman AC, Sharples LD. A general goodness-of-fit test for Markov and hidden Markov models. *Stat Med*. 2008;27(12):2177-2195. https://doi.org/10.1002/sim.3033.
43. Nygard M, Roysland K, Campbell S, Dillner J. Comparative effectiveness study on human papillomavirus detection methods used in the cervical cancer screening programme. *BMJ Open*. 2014;4(1):e003460. https://doi.org/10.1136/bmjopen-2013-003460.
44. Fahey MT, Irwig L, Macaskill P. Meta-analysis of Pap test accuracy. *Am J Epidemiol*. 1995;141(7):680-689.
45. Nanda K, McCrory DC, Myers ER, et al. Accuracy of the Papanicolaou test in screening for and follow-up of cervical cytologic abnormalities: a systematic review. *Ann Intern Med*. 2000;132(10):810-819.
46. Ronco G, Dillner J, Elfstrom KM, et al. Efficacy of HPV-based screening for prevention of invasive cervical cancer: follow-up of four European randomised controlled trials. *Lancet*. 2014;383(9916):524-532. https://doi.org/10.1016/S0140-6736(13)62218-7.
47. Dillner J, Rebolj M, Birembaut P, et al. Long term predictive values of cytology and human papillomavirus testing in cervical cancer screening: joint European cohort study. *BMJ*. 2008;337:a1754. https://doi.org/10.1136/bmj.a1754.
48. Arbyn M, Buntinx F, Van Ranst M, Paraskevaidis E, Martin-Hirsch P, Dillner J. Virologic versus cytologic triage of women with equivocal Pap smears: a meta-analysis of the accuracy to detect high-grade intraepithelial neoplasia. *J Natl Cancer Inst*. 2004;96(4):280-293.
49. Arbyn M, Anttila A, Jordan J, et al. *European Guidelines for Quality Assurance in Cervical Cancer Screening*. Second Edition—Summary Document. *Ann Oncol*. 2010;21(3):448-458. http://dx.doi.org/10.1093/annonc/mdp471.
50. Carreon JD, Sherman ME, Guillen D, et al. CIN2 is a much less reproducible and less valid diagnosis than CIN3: results from a histological review of population-based cervical samples. *Int J Gynecol Pathol*. 2007;26(4):441-446. https://doi.org/10.1097/pgp.0b013e31805152ab.

51. Dalla PP, Giorgi RP, Collina G, et al. The reproducibility of CIN diagnoses among different pathologists: data from histology reviews from a multileft randomized study. *Am J Clin Pathol*. 2009;132(1):125-132. https://doi.org/10.1309/AJCPBRK7D1YIUWFP.
52. Zuchna C, Hager M, Tringler B, et al. Diagnostic accuracy of guided cervical biopsies: a prospective multileft study comparing the histopathology of simultaneous biopsy and cone specimen. *Am J Obsftet Gynecol*. 2010;203(4):321 e1-321 e6. https://doi.org/10.1016/j.ajog.2010.05.033.
53. Armitage P, Doll R. The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br J Cancer*. 1954;8(1):1-12.
54. Schiffman M, Wentzensen N, Wacholder S, Kinney W, Gage JC, Castle PE. Human papillomavirus testing in the prevention of cervical cancer. *J Natl Cancer Inst*. 2011;103(5):368-383. https://doi.org/10.1093/jnci/djq562.
55. Moscicki AB, Schiffman M, Burchell A, et al. Updating the natural history of human papillomavirus and anogenital cancers. *Vaccine*. 2012;30(Suppl 5):F24-F33. https://doi.org/10.1016/j.vaccine.2012.05.089.
56. Rodriguez AC, Schiffman M, Herrero R, et al. Longitudinal study of human papillomavirus persistence and cervical intraepithelial neoplasia grade 2/3: critical role of duration of infection. *J Natl Cancer Inst*. 2010;102(5):315-324. https://doi.org/10.1093/jnci/djq001.
57. Orumaa M, Leinonen MK, Campbell S, Møller B, Myklebust TA, Nygård M. Recent increase in incidence of cervical precancerous lesions in Norway: Nationwide study from 1992 to 2016. *Int J Cancer*. 2019;145(10):2629-2638. http://dx.doi.org/10.1002/ijc.32195.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Soper BC, Nygård M, Abdulla G, Meng R, Nygård JF. A hidden Markov model for population-level cervical cancer screening data. *Statistics in Medicine*. 2020;1–22. https://doi.org/10.1002/sim.8681