A prevalence-incidence-clearance model for interval-censored screening and surveillance data in a population with an elevated disease risk at baseline

Running title: Prevalence-incidence-clearance model for screening and surveillance data

Authors and affiliations:
Kelsi R Kroon[1,2 §], Johannes A Bogaards[1,2], Chris JLM Meijer[3,4], Johannes Berkhof[1,2]

[1] Amsterdam UMC, location Vrije Universiteit Amsterdam, Epidemiology and Data Science, De Boelelaan 1117, Amsterdam, the Netherlands
[2] Amsterdam Public Health Research Institute, Methodology, Amsterdam, the Netherlands
[3] Amsterdam UMC, location Vrije Universiteit Amsterdam, Pathology, De Boelelaan 1117, Amsterdam, the Netherlands
[4] Cancer Center Amsterdam, Imaging and Biomarkers, Amsterdam, the Netherlands

§ Corresponding Author:
Kelsi R Kroon, k.kroon@amsterdamumc.nl
Amsterdam UMC, location Vrije Universiteit Amsterdam, Epidemiology and Data Science, De Boelelaan 1117, 1081 HV Amsterdam, the Netherlands

Abstract

Accurate risk assessment is essential for screening and surveillance programs, but this is complicated when considering baseline conditions linked to an increased risk of disease that may decline over time (e.g., certain infections and viral-induced disease). In longitudinal screening and surveillance studies, individuals may have prevalent disease at baseline or develop it during follow-up, either from their baseline condition ("early" event) or from a new condition ("late" event). Additionally, data are interval-censored between visits, making the exact time of disease onset unknown.

We propose a prevalence-incidence-clearance model for interval-censored data to estimate cumulative disease risk based on individual risk factors, with the motivating example of human papillomavirus (HPV) infections, which may progress to high-grade cervical lesions and cancer (CIN2+). Early events are modelled with an exponential competing risks framework, where HPV infections either progress to CIN2+ or to a (latent) "clearance" state. Late events are modelled by adding a background risk. Parameters are estimated with an expectation-maximisation algorithm with weakly informative Cauchy priors. The algorithm was validated through simulation studies and applied to screening and post-treatment surveillance data from the Netherlands. Our model accurately predicts cumulative CIN2+ risk in HPV-positive women and fits the observed cumulative incidence curve better than existing methods. Furthermore, it provides easily interpretable parameters and its baseline hazard can be checked for lack of fit. This is especially important when applying the model to facilitate decision-making for national programs.

Keywords: cumulative risk estimation; cure models; mixture model; interval censoring; screening data; EM algorithm

1. Introduction

In public health and medical research, prediction models play a crucial role in identifying individuals who are at highest risk of developing a disease, although some may only have a temporarily heightened risk due to a baseline condition. A motivating example is human papillomavirus (HPV) infection, which causes almost all cervical cancers worldwide[1] and is also associated with several other cancers and warts[2]. After an initial HPV infection, individuals are at risk of developing a lesion until the infection clears. Ideally, models for predicting the risk of a lesion feature easily interpretable key components that can be generalized to different populations and guide decision-making regarding the benefits and harms of healthcare interventions such as cervical cancer screening and surveillance.

Continuing with the example of cervical cancer, screening programs have been evolving from traditional 'one-size-fits-all' approaches to more targeted methods informed by viral and host risk factors in order to optimally balance the benefit and harms. To develop risk-based cervical screening programs, it is necessary to accurately predict the cumulative risk of cervical intraepithelial neoplasia grade 2 or worse (CIN2+) following an HPV infection. In a screening setting, women are observed at multiple time-points during follow-up and there are two potential scenarios for detection of CIN2+. Women may have underlying *prevalent* CIN2+ at baseline, or CIN2+ might develop in-between two of the visits (*interval-censored incident CIN2+*). Incident CIN2+ can develop either from a baseline infection ("early" event) or from a newly acquired or reactivated infection during follow-up ("late" event). The baseline infection may also clear during follow-up so that the cumulative risk of CIN2+ is likely to level off during follow-up. In addition, clearance of an infection and start of a new infection may happen in parallel because there are multiple HPV genotypes that can result in multiple infections.

Non-parametric and parametric models have widely been used for estimating the cumulative risk but they have significant limitations when applied to our setting. Non-parametric cumulative incidence models[3–6] assume that there are no prevalent disease cases. This can be accounted for by defining a very small interval with a left bound of zero[7]. However, a non-parametric model also does not allow for separate modelling of covariate effects on prevalent and incidence disease, which is crucial for

interpreting cumulative disease risks. Parametric models allow for more flexibility with respect to the modelling of the effects of covariates. A Weibull prevalence-incidence model has been proposed[7] which assumes a logistic regression model for prevalent disease and a Weibull regression model for incident disease. A Weibull model assumes that the hazard is proportionally related to the covariates, which is not likely to hold when the disease risk is only temporarily increased. Another option is the family of parametric cure models[8–11], where a proportion of the population has an increased risk of disease described by a relative survival distribution compared to the distribution in the general population. However, these models do not distinguish between incident and prevalent disease, assume that data are right-censored instead of interval-censored and assume that the disease risk returns to the average of the general population which may not be realistic.

We address the problems identified in existing methods by adding two elements to the prevalence-incidence model to create a prevalence-incidence-clearance (PIC) model, motivated by the example of HPV infection and CIN2+. Specifically, we describe the non-proportional hazard of incident disease using separate models for early and late events. For early events, we use a competing risks framework for incident CIN2+ cases, where infections can either progress to CIN2+ according to an exponential time distribution or regress to a latent clearance state. This means that the cumulative incidence of early events does not go to 1 but converges to an asymptote below 1. Late events are modelled by adding a background CIN2+ risk to the model, which refers to late-occurring events that are not caused by the HPV infection at baseline. A particularly important property of our model is that the model for incident disease does not depend on the time since onset of the baseline condition. This property enables us to generalize the results from a specific screening cohort to a general screening population that is screened repeatedly. To examine whether the model property is violated by the data, we present a score test comparing the exponential distribution of time to incident disease against a Gamma alternative.

The article is structured as follows. In Section 2, we describe the PIC model and some of its features. In Section 3, we describe an expectation-maximisation (EM) algorithm for estimating the model parameters with interval-censored screening data. The EM algorithm has been implemented in an R package (*PICmodel)*. In Section 4, we

present simulation studies to verify that the algorithm can accurately recover the true parameter values and a score test to check the exponential distribution for incident disease. In Section 5, we apply the model to estimate the cumulative risk of CIN2+ in two different data sets. The first data set consists of a screening population of women with an HPV-positive test result at baseline[12] and the second data set consists of a patient cohort of women treated for CIN2+ at baseline who are at risk of recurrent disease[13]. Finally, in Section 6 we conclude and discuss our findings.

2. Model

Due to the nature of screening programs and disease surveillance, the exact time of onset of disease is not observed. Instead, there is only an interval of the last time an individual with the baseline condition was observed without the disease, and the first time that the disease was detected. If the disease did not develop during follow-up, this interval will be right-censored. Further, some individuals are immediately referred at baseline, therefore it can be determined whether the disease is present at baseline. However, some individuals are not referred at baseline but the disease is detected at their first visit, in which case it is unknown whether the disease was present at baseline.

For individuals $i = 1, \dots, n$ who have an elevated risk due to a baseline condition, let $(L_i, R_i]$ be the interval when the disease develops, where L_i is the last time the ith individual was observed without disease and R_i is the first time that disease is detected for the ith individual. We allow $R_i = \infty$ to include individuals who do not develop disease during follow-up. Let z_i be an indicator variable for prevalent disease at baseline (where $z_i = 1$ means that the disease was present at baseline), which, as described above, is unknown for some individuals. Let $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ be disjoint vectors of parameters, and let $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ be subject-specific vectors of covariates for prevalence and progression respectively. Let $\pi(\boldsymbol{x}_1;\ \boldsymbol{\theta}_1)$ be the probability of having the disease at baseline (prevalent disease). Then the probability that the disease has developed before time t, i.e. the time to disease, denoted by T, lies in $[0, t)$, is a mixture distribution given by

$$P(T \leq t; \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \lambda_2, \beta) = \pi(\boldsymbol{x}_1; \boldsymbol{\theta}_1) + (1 - \pi(\boldsymbol{x}_1; \boldsymbol{\theta}_1)) F(t; \boldsymbol{x}_2, \boldsymbol{\theta}_2, \lambda_2, \beta) \quad (1)$$

where $F(t; \boldsymbol{x}_2, \boldsymbol{\theta}_2, \lambda_2, \beta)$ is the cumulative risk of incident disease. Among incident cases there are two independent causes, disease can either develop from the baseline

condition (an “early” event), or it can develop due to a new condition during follow-up or reactivation of the existing condition (a “late” event). We do not know the cause of an event during follow-up so we model a mixture of early and late events. Early events are modelled with a competing events framework, where infections either progress to the disease or to a latent clearance state, both of which are assumed to follow an exponential distribution with constant rate process parameters $\lambda_1(\boldsymbol{x_2}, \boldsymbol{\theta_2})$ and λ_2, respectively. Then the cumulative risk of developing “early” disease at time t is given by

$$G(t; \boldsymbol{x}_2, \boldsymbol{\theta}_2, \lambda_2) = \frac{\lambda_1(\boldsymbol{x_2}, \boldsymbol{\theta_2})}{\lambda_1(\boldsymbol{x_2}, \boldsymbol{\theta_2}) + \lambda_2}\left(1 - \exp(-(\lambda_1(\boldsymbol{x_2}, \boldsymbol{\theta_2}) + \lambda_2)t)\right), \tag{2}$$

where $\frac{\lambda_1(\boldsymbol{x_2}, \boldsymbol{\theta_2})}{\lambda_1(\boldsymbol{x_2}, \boldsymbol{\theta_2}) + \lambda_2}$ is the proportion of HPV-positive women at baseline that develop the disease during follow-up. For “late” events, we define the cumulative background risk of developing disease, $H(t; \beta)$, given by:

$$H(t; \beta) = 1 - \exp(-\beta t) \tag{3}$$

where β is the background hazard. Now, we get the following model for incident disease:

$$F\left(t; \boldsymbol{x}_2, \boldsymbol{\theta}_2, \lambda_2, \beta\right) = G(t; \boldsymbol{x}_2, \boldsymbol{\theta}_2, \lambda_2) + [1 - G(t; \boldsymbol{x}_2, \boldsymbol{\theta}_2, \lambda_2)]H(t; \beta). \tag{4}$$

Figure 1 gives a schematic representation of interval-censored data and several possible outcomes observed during screening that are related to the components of the model with the example of cervical cancer screening data.

The model allows for the inclusion of baseline covariates for the prevalence parameter and the rate parameter $\lambda_1(\boldsymbol{x}_2, \boldsymbol{\theta}_2)$. For the proportion of prevalent disease, $\pi(\boldsymbol{x}_1, \boldsymbol{\theta}_1)$, a logistic regression model is used to include covariates:

$$\text{logit}\left(\pi(\boldsymbol{x}_1, \boldsymbol{\theta}_1)\right) = \theta_{1,0} + \theta_{1,1}x_{1,1} + \cdots + \theta_{1,m_1}x_{1,m_1}, \tag{5}$$

where m_1 is the number of covariates included in the logistic model for prevalent disease. For both λ_1 and λ_2, a log transformation is used in which case the covariates for λ_1 have a proportional effect on its hazard:

$$\log(\lambda_1(\boldsymbol{x_2}, \boldsymbol{\theta}_2)) \quad = \theta_{2,0} + \theta_{2,1}x_{2,1} + \cdots + \theta_{2,m_2}x_{2,m_2} \tag{6}$$

where m_2 is the number of covariates used for the rate parameter λ_1. In principle, the parameters λ_2 and β can also be linked to covariates, but this is not considered here as their regression coefficients tend to be less well identified as those of π and λ_1. The main role of λ_2 and β is to describe how the baseline hazard is influenced by clearance

of the baseline condition and the occurrence of "late" events, whereas the relation between cumulative disease risk and covariates is defined through regression equations for π and λ_1.

3. Parameter estimation with the EM algorithm

A variant of the traditional expectation-maximisation (EM) algorithm called the EM gradient algorithm[14] is used to estimate the parameters. The EM gradient algorithm approximately solves the maximisation step of the EM algorithm by using a single Newton step on the expected log-likelihood instead of directly solving the maximisation, and it has been shown to have similar convergence properties as the standard EM algorithm[15].

3.1 Likelihood

To construct the likelihood, the subjects are divided into sets with missing and observed z_i. Subjects with $z_i = 1$ contribute $\log(\pi(\boldsymbol{x}_1; \boldsymbol{\theta}_1))$ to the log-likelihood, and subjects with $z_i = 0$ contribute

$$\log\{(1 - \pi(\boldsymbol{x}_1; \boldsymbol{\theta}_1))[F(R_i; \boldsymbol{x}_2, \boldsymbol{\theta}_2, \lambda_2, \beta\ ) - F(L_i; \boldsymbol{x}_2, \boldsymbol{\theta}_2, \lambda_2, \beta)]\} \tag{7}$$

to the log-likelihood as the difference of the cumulative densities is used to get the likelihood of disease developing between the left and right time points for interval-censored data. When z_i is missing, it cannot be ruled out that the disease was present at baseline, so the left bound is fixed at zero and $F(L_i; \boldsymbol{x}_2, \boldsymbol{\theta}_2, \lambda_2, \beta) = 0$. Therefore, subjects with missing z_i contribute

$$\log\{\pi(\boldsymbol{x}_1; \boldsymbol{\theta}_1) + (1 - \pi(\boldsymbol{x}_1, \boldsymbol{\theta}_1))F(R_i; \boldsymbol{x}_2, \boldsymbol{\theta}_2, \lambda_2, \beta)\} \tag{8}$$

to the log-likelihood. Putting these together, the complete-data log-likelihood is given by

$$\ell_c(\boldsymbol{\theta}) = \sum_{z_i \text{ observed}} \left[z_i \log(\pi(\boldsymbol{x}_1; \boldsymbol{\theta}_1)) + (1 - z_i)\left\{\log(1 - \pi(\boldsymbol{x}_1; \boldsymbol{\theta}_1)) + \log\left(F(R_i; \boldsymbol{x}_2, \boldsymbol{\theta}_2, \lambda_2, \beta) - F(L_i; \boldsymbol{x}_2, \boldsymbol{\theta}_2, \lambda_2, \beta)\right)\right\}\right]$$
$$+ \sum_{z_i \text{ missing}} \log\left(\pi(\boldsymbol{x}_1; \boldsymbol{\theta}_1) + (1 - \pi(\boldsymbol{x}_1; \boldsymbol{\theta}_1))F(R_i; \boldsymbol{x}_2, \boldsymbol{\theta}_2, \lambda_2, \beta)\right) \tag{9}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \lambda_2, \beta\}$.

3.2 Expectation step

Given the current parameter estimates $\boldsymbol{\theta}^{(l)}$ and the observed data, the expected value of the latent variable z_i, $\mathbb{E}\left(z_i|\boldsymbol{\theta}^{(l)}\right)$, is estimated. When z_i is known then the expected value of z_i is simply its value. When z_i is unknown, the left bound is always zero, so $F(L_i;\boldsymbol{x}_2,\boldsymbol{\theta}_2,\lambda_2,\beta)=0$, and the expected value of z_i is given by

$$\mathbb{E}\left(z_i|\boldsymbol{\theta}^{(l)}\right)=\frac{\pi(\boldsymbol{x}_1;\boldsymbol{\theta}_1^{(l)})}{\pi(\boldsymbol{x}_1;\boldsymbol{\theta}_1^{(l)})+\left(1-\pi(\boldsymbol{x}_1;\boldsymbol{\theta}_1^{(l)})\right)F\left(R_i;\boldsymbol{x}_2,\boldsymbol{\theta}_2^{(l)},\lambda_2^{(l)},\beta^{(l)}\right)}. \tag{10}$$

3.3 Maximization step

Given the observed data and the latent variables, we estimate the parameters by maximizing the expected complete data log-likelihood, $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(l)}\right)$

$$Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(l)}\right)=\sum_{i=1}^{n}\mathbb{E}\left(z_i|\boldsymbol{\theta}^{(l)}\right)\log\left(\pi(\boldsymbol{x}_1;\boldsymbol{\theta}_1)\right)+\left(1-\mathbb{E}\left(z_i|\boldsymbol{\theta}^{(l)}\right)\right)\left[\log(1-\pi(\boldsymbol{x}_1;\boldsymbol{\theta}_1))+\log\left(F(R_i;\boldsymbol{x}_2,\boldsymbol{\theta}_2,\lambda_2,\beta)-F(L_i;\boldsymbol{x}_2,\boldsymbol{\theta}_2,\lambda_2,\beta)\right)\right]. \tag{11}$$

To update the parameters, we use a single Newton step where the next parameter value is given by

$$\boldsymbol{\theta}^{(l+1)}=\boldsymbol{\theta}^{(l)}-\left[\mathrm{H}_Q\left(\boldsymbol{\theta}^{(l)}\right)\right]^{-1}\mathrm{U}_Q\left(\boldsymbol{\theta}^{(l)}\right). \tag{12}$$

In the equation above, $\mathrm{U}_Q(\boldsymbol{\theta})$ is the gradient of $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(l)}\right)$ consisting of first derivatives, and $\mathrm{H}_Q(\boldsymbol{\theta})$ is the Hessian matrix of second partial derivatives of $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(l)}\right)$.

The expectation and maximisation step are repeated until the absolute change in the log-likelihood between iterations is less than 10^{-8}. After the model has converged, the Hessian of the observed log-likelihood is calculated using numerical differentiation, then variances of the parameter estimates are calculated by the inverse of the negative of the Hessian of the observed data log-likelihood at $\widehat{\boldsymbol{\theta}}$.

3.4 Initial values

Direct implementation of (11) and (12) will lead to convergence problems. Instead, a two-step process is implemented in order to estimate the background risk. To find initial values for the other model parameters, we use a version of the Short EM algorithm[16]. This method involves using multiple sets of random starting values and performing short runs of the EM algorithm with a relaxed convergence criterion (e.g., 10^{-2} instead of 10^{-8}) and low number of maximum iterations (e.g., 10). The parameter

estimates that yield the highest log-likelihood are then used as initial values in the EM gradient algorithm.

In the second step, we take a very small value of background risk (e.g., fixed at $\log(\beta) = -12$) with the initial values found in the previous step for the remaining parameters, then we iteratively increase the fixed value of background risk and at each step use as initial values for the remaining parameters the estimates found in the previous iteration until the log-likelihood stops increasing. Lastly, the estimates from the most recent iteration are used as initial values in a final run of the EM algorithm where the background risk is no longer kept at a fixed value to find maximum likelihood estimates with standard errors for all parameters. This procedure yields a high-speed algorithm with a high convergence rate which makes it suitable for general consumption, speed and convergence results are described in the simulation study.

3.5 Separability issues

A situation might arise where there are no or a very low number of subjects with a specific combination of a covariate and the outcome. This causes identifiability issues when estimating the maximum likelihood estimate for this covariate, and although the algorithm may converge, it is likely that the confidence intervals for the estimate will be very large. To address these issues in the logistic prevalence mixture, we choose a Cauchy distribution with a centre of 0 and scale of 2.5, as recommended by Gelman et al. (2008)[17], as a weakly informative prior. The first and second derivatives of the log of the Cauchy$(0, 2.5)$ distribution with respect to each parameter can be directly evaluated in the maximisation step and added to calculations of the gradient and Hessian, respectively. Credible intervals for the parameters in the model with a Cauchy prior can be obtained through Taylor expansion[18].

Simulation study

4.1 Validation of model fit

Simulation study: setup

We assessed the performance of our model in a simulation study where data was simulated according to the following settings. For the disease process, values for the parameters $\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \lambda_2$ and β were set, and for each subject i, a Bernoulli random variable for prevalent disease, P_i, was drawn with probability $\pi(\boldsymbol{x}_1, \boldsymbol{\theta}_1)$. In subjects

without prevalent disease (i.e. $P_i = 0$), a second Bernoulli variable for "early" incident disease, A_i, was drawn with probability $\frac{\lambda_1(\boldsymbol{x_2},\boldsymbol{\theta_2})}{\lambda_1(\boldsymbol{x_2},\boldsymbol{\theta_2})+\lambda_2}$. Additionally, two event times were generated for "early" incident disease (T_1) and "late" incident disease (T_2). For incident disease caused by the baseline condition, T_1 was drawn from an exponential distribution with rate parameter $\lambda_1(\boldsymbol{x_2}, \boldsymbol{\theta_2}) + \lambda_2$ if $A_i = 1$, and set to infinity otherwise. For incident disease due to background risk, T_2 was drawn from an exponential distribution with rate parameter β. We then took the minimum of T_1 and T_2 as the onset time of incident disease, i.e., $T_{\mathrm{D}} = \min(T_1, T_2)$.

For the observation process, screening (or surveillance) times were drawn from a Beta distribution with mean around a specified interval (e.g., 3-year intervals) until there was at least 25 years of follow-up and for each individual a Bernoulli variable with a pre-defined attendance probability was drawn to indicate whether they showed up for a screening visit at a specific time point. From this, left and right intervals were created using the last time an individual was observed before the time of disease onset and the first time observed after disease onset.

The following parameter values for prevalence, progression, and clearance were used for the simulation study without covariates: $\pi = \log(0.20), \log(\lambda_1) = -1$, and $\log(\lambda_2) = -1.5$. These parameter values were loosely based on values observed in cervical cancer screening studies. We investigated the effect of simulation settings on model performance by varying the (1) sample size (N=1000 or 10,000), (2) attendance probability per screening round (75% or 100%), (3) value of background risk ($\beta = 0, 0.005, 0.01,$ or 0.02) and (4) weakly informative prior (none or Cauchy(0,0.25)). This resulted in 32 separate simulation conditions.

The simulation study was repeated with baseline covariates, again loosely based on values in cervical cancer screening studies. A continuous variable Age was generated from a Uniform(30, 70) distribution to reflect age at baseline, and a binary variable HPV16 (e.g., positive for HPV genotype 16 at baseline) was generated with probability of being equal to 1 set at a 0.3. This parameter was assumed to positively affect both the progression and prevalence probabilities in the model. The following model was used:

$$\text{logit}\big(\pi(\boldsymbol{x_1}, \boldsymbol{\theta_1})\big) = \theta_{1,0} + \theta_{1,1} \times \text{Age} + \theta_{1,2} \times \text{HPV16} \qquad (14)$$

$$\log\left(\lambda_1(\boldsymbol{x_2}, \boldsymbol{\theta_2})\right) = \theta_{2,0} + \theta_{2,1} \times \text{HPV16}$$

with parameters set at: $\boldsymbol{\theta_1} = \{\theta_{1,0} = -3, \theta_{1,1} = 1, \theta_{1,2} = 1\}, \boldsymbol{\theta_2} = \{\theta_{2,0} = -1, \theta_{2,1} = 2\}$, and $\lambda_3 = -1.5$. All simulated data sets were fit with our EM algorithm using 10 different random starting values per data set. For each simulation setting 200 data sets were generated and the median estimate, the percent bias, and 95% confidence/credible interval coverage probability (CP) were calculated (shown in Figure 2). Additionally, the time that it takes for the algorithm to converge and the convergence rate were calculated for models with and without covariates in the setting with low background risk $(h = 0.005)$, imperfect attendance probability (75%) and priors included.

Simulation study: results

The model performed well in percentage bias and coverage across all parameters in simulations, regardless of covariates, attendance probability, or priors, for low background risk (Figure 2). Without covariates, percentage bias ranged from -1.64% to 1.93% (N=1000) and -0.42% to 0.45% (N=10,000), with coverage probability between 91.0% to 98.5% (N=1000) and 91.0% to 98.0% (N=10,000). Percentage bias decreased with higher attendance probability ranging between -1.45% and 1.93% (75% attendance) and between -1.64% and 0.91% (perfect attendance). Percentage bias ranged from -1.64% to 0.95% with weakly informative priors and from -1.30% to 1.93% without priors. Attendance probability and inclusion of priors had minimal impact on coverage probability. With covariates, similar trends were observed. For N=1000, percentage bias ranged from -6.05% to 5.63%, with coverage probability between 91.0% to 99.0%. For N=10,000, absolute percentage bias remained low (<1.45%) across settings, while coverage probability ranging from 91.5% to 99.0%.

The average CPU time (in seconds) to fit each of the 200 simulated data sets increased with covariates. Without covariates, average time was 2.42s (N=1000) and 19.77s (N=10,000), but with covariates the average time rose to 14.96s (N=1000) and 143.85s (N=10,000). There were no cases of non-convergence in any of the settings.

4.2 Checking exponential progression rate

Score test

The exponential progression rate for early events may not hold even after adjusting for latent clearance and the occurrence of late events. This limits the generalizability of the model when used to support decision making for a screening or surveillance protocol with multiple test moments. An example is the outcome cancer developing through multiple stages which can be described by assuming an increasing hazard since onset of the baseline condition. We evaluated the fit of an exponential time distribution for early events by devising a score test with a Gamma distribution as alternative. The Gamma distribution is a mathematically attractive alternative because if the progression to disease follows a Gamma distribution with shape parameter k and rate parameter λ_1, then the cumulative incidence function for the early events model becomes

$$G(t;\lambda_1,\lambda_2,k) = \frac{\lambda_1^k}{(\lambda_1+\lambda_2)^k}\frac{1}{\Gamma(k)}\int_0^t (\lambda_1+\lambda_2)^k\, u^{k-1} e^{-(\lambda_1+\lambda_2)u} du, \qquad (15)$$

which in turn is the distribution function of a Gamma variable with shape parameter k and rate parameter $\lambda_1+\lambda_2$ multiplied by the constant $\frac{\lambda_1^k}{(\lambda_1+\lambda_2)^k}$.

The null hypothesis is that the Gamma shape parameter, k, in equation (15) equals one, thereby reducing the Gamma distribution to an exponential distribution. To calculate the gradient and Hessian of the log-likelihood function evaluated at $k=1$, we used numerical methods to obtain the first and second derivatives of $G(t)$ in (15) with respect to the shape parameter. Let $\widetilde{\boldsymbol{\theta}} = \{k=1, \widehat{\boldsymbol{\theta}}\}$ be the constrained maximum likelihood estimator (MLE) where $\widehat{\boldsymbol{\theta}}$ is the MLE for the remaining parameters, $U(\widetilde{\boldsymbol{\theta}})$ is the score vector, and $I(\widetilde{\boldsymbol{\theta}})$ is the Fisher information matrix. Then the score test statistic is given by

$$S(\widetilde{\boldsymbol{\theta}}) = \frac{U(\widetilde{\boldsymbol{\theta}})^2}{I(\widetilde{\boldsymbol{\theta}})} \qquad (16)$$

which is asymptotically Chi-squared distributed with 1 degree of freedom under the null.

Simulation study: setup

Simulation studies were performed to evaluate the power and Type 1 error of the model under varying combinations of the (1) shape parameter, (2 background risk,

and (3) sample size). Data was generated according to the procedure described previously, however in this case the Bernoulli variable for "early" incident disease, A_i, was drawn with probability $\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k$ and T_1 was generated from a gamma distribution with shape parameter k and rate parameter $\lambda_1 + \lambda_2$, according to equation (15). The shape parameter was set at either $k = 1, 1.5$ or 2, where $k = 1.5$ and $k = 2$ are time distributions with increasing hazards. Background risk was set at either $\beta = 0$, $0.005, 0.01,$ or 0.02, and was assumed to be either known and fixed at its correct value or estimated with the EM algorithm. Sample size was either $N = 1000$ or $10{,}000$. This makes $2\ (N) \times 3\ (k) \times 3\ (\beta)\ = 36$ combinations in case the background risk is positive, plus $2\ (N) \times 3\ (k) = 6$ combinations in case the background risk is zero resulting in 42 combinations. Data sets were generated for each combination until 100 data sets had a successful model fit.

Simulation study: results

We can read from Table 1 that the Type I error ranges from 0.03 to 0.10 when N=1000 and from 0.03 to 0.05 when N=10,000. The power under k > 1 is moderate to high when k = 2 and N=1000 and negligible or low when k < 2 and N=1000. The power ranges from 0.71 to 1.00 when N=10,000. Estimating the background risk (instead of assuming a fixed background risk) substantially decreased the power when N=1000 but not when N=10,000. A likely explanation is that a deviation from an exponential distribution competes with the "late" event distribution and it becomes hard to identify lack of model fit after having estimated background risk β in moderately sized data sets. This is supported by Table 1 as the power under k > 1 decreases with background risk β.

5. Examples

5.1 Cervical cancer screening data

The Population-based screening study Amsterdam (POBASCAM) is a randomized controlled screening trial (Netherlands trial registration ID: NTR218) in the Netherlands consisting of women from the Dutch population recruited between 1999 and 2002. The study has been described before[12,19]. In this study, we used a subset of 2,269 women who were HPV-positive at their first screening date (baseline). The maximum follow-up was 19 years.

To create interval censored data from this screening data set, the date of the first histology test reporting CIN2+ was set as the right bound (or infinity if CIN2+ was not reported). However, if an index smear was reported within 3 months before the first CIN2+ histology diagnosis, it was used as the right bound instead. For the left bound, the most recent normal cytological finding and/or histological finding reporting <CIN2 occurring before the right bound was used. In the absence of prior results, the baseline date was used as the left bound.

For illustration, three baseline covariates were used in the application. Age was included as a continuous variable and was standardised to have a mean of 0 and standard deviation of 0.5, as suggested by [17]. HPV genotype was defined as a binary variable indicating the presence of either HPV genotype 16 (HPV16 = 1) or another high-risk type (HPV16 = 0). Cytology was defined as a binary variable with score 1 if abnormal and 0 if normal. The link function was

$$\text{logit}(\pi(\boldsymbol{x_1}, \boldsymbol{\theta_1})) = \theta_{1,0} + \theta_{1,1} \times \text{Age} + \theta_{1,2} \times \text{HPV16} + \theta_{1,3} \times \text{Cytology}. \quad (17)$$

Progression rate parameter λ_1 (progression) was assumed to depend only on HPV16 genotype, while λ_2 (“clearance”) was kept constant:

$$\log\big(\lambda_1(\boldsymbol{x_2}, \boldsymbol{\theta_2})\big) = \theta_{2,0} + \theta_{2,1} \times \text{HPV16} \quad (18)$$

All analysis was performed in R 4.4.0 (The R Foundation) using our R package *[PICmodel],* which is available through GitHub.

We started by comparing the fit of the PIC model without covariates with those of three existing models: a non-parametric model for interval-censored data with prevalent disease, a logistic-Weibull and a logistic-lognormal prevalence-incidence model[7]. Our model performed well in capturing the shape of the non-parametric estimate (Figure 3A) whereas the logistic-Weibull and the logistic-lognormal models slightly underestimated the cumulative incidence for years 3 to 12 after baseline. This is also reflected in the log-likelihood values at the maximum likelihood estimates which was -2366.06 for the PIC model, -2380.92 for the logistic-Weibull model and -2370.73 for the logistic-lognormal model.

Table 2 gives the parameter estimates with 95% confidence/credible intervals, log-likelihood values and score test results for three PIC models fit with covariates: (1) without background risk, (2) with background risk, and (3) with background risk and

Cauchy(0, 0.25) prior. Background risk improves model fit, as the log-likelihood (LL) significantly increased when background risk was added to the model (LL without background risk: -2036.40, LL with background risk: -2033.97, p-value for likelihood ratio test (LRT) = 0.027). Score tests for models with and without background risk were not significant (p-value = 0.68 and p-value = 0.95, respectively) suggesting that the data do not provide evidence for a Gamma, rather than exponential distribution for progression to CIN2+. We also investigated the inclusion of a weakly informative Cauchy(0, 0.25) prior. The results illustrate that the weakly informative Cauchy prior can substantially affect the width of the confidence intervals, as the coefficients for the intercept and effect of cytology in the prevalence model have much wider (frequentist) confidence intervals compared to the credible intervals in the model with the Cauchy prior. Further inspection indicated that this is likely due to a separability issue as the disease status of women with normal cytology is not ascertained at baseline, but in the model inferred from women with a missing disease status at baseline.

5.2 Cervical post-treatment follow-up data

Our model is also suitable for follow-up data after treatment for CIN2+, because there is a chance of residual disease due to incomplete excision (prevalent disease), the baseline condition (HPV infection) may have persisted and regrow to a lesion ("early" event), and women remain at increased risk of developing CIN2+ from a new HPV infection compared to the general population ("late" event)[20]. We used data from a long-term multi-cohort study following women after treatment for CIN2+ (Netherlands trial registration ID: NTR1468)[13]. This cohort study included 435 women who had participated in one of three earlier studies that monitored them using HPV testing and cytology at 6-, 12-, and 24-month intervals after large loop excision of the transformation zone (LLETZ) or cold-knife conization treatment for CIN2/3+. After this period of monitoring, women were followed through the routine screening programme with 5-year intervals. The maximum follow-up was 21 years.

For our example, the prevalence model had standardised continuous age (Age) and disease status at treatment (CIN3; 1 if CIN3+, 0 if <CIN3) as covariates:

$$\text{logit}(\pi(\boldsymbol{x_1}, \boldsymbol{\theta_1})) = \theta_{1,0} + \theta_{1,1} \times \text{Age} + \theta_{1,2} \times \text{CIN3} \tag{19}$$

and the progression parameters λ_1 and λ_2 were constant.

We again started by fitting the PIC, logistic-Weibull and logistic-lognormal model without covariates. Figure 3B clearly illustrates that the PIC model is closer to the nonparametric model than the logistic-Weibull and logistic-lognormal model which is also reflected in the loglikelihood values at the maximum likelihood estimate: -301.04 for the PIC model, -304.32 for the logistic-Weibull model, and -303.74 for the logistic-lognormal model.

Table 3 gives the parameter estimates with 95% confidence intervals, log-likelihood values and score test results for three PIC models with covariates. The results show that background risk significantly improves model fit for these data (LL without background risk: -301.84, LL with background risk: -297.83, LRT p-value: 0.004). Additionally, the score test was significant for a model without background risk (p-value = 0.008), but was not significant when background risk had been included (p-value = 0.39) indicating that an exponential progression model for the early events and background risk for the late events is not violated by the data. Using the Cauchy prior led to substantially smaller 95% (credible) intervals for the intercept and effect of CIN3 in the prevalent disease model with background risk.

6. Discussion

We present a model for a population with a temporarily elevated baseline risk that adds to the concept of prevalence-incidence models for the cumulative risk of disease introduced by Cheung et al.[7]. By combining concepts from this model with additional parameters, we were able to decompose a multifaceted process into simpler key components that together provide accurate estimates of the cumulative disease risk compared to non-parametric estimates.

To model progression, a competing risks framework with exponential distributions for disease progression and latent clearance was used, which better modelled cumulative incidence curves than Weibull and lognormal models . A specific feature of our model is that it distinguishes between early and late events. This distinction is particularly important in cervical cancer screening, as the time between HPV infection and CIN2+ is determined by the early events. Models that do not separate these events may overestimate the lead time gain from HPV testing, which could influence decisions about the screening interval. An advantage of modelling early events with an

exponential distribution is that the hazard for individuals still at risk remains constant over time, ensuring comparability across studies regardless of when data collection begins. This is important when analysing screening cohort data from different countries with different age ranges. To examine whether the exponential distribution is not violated by the data, we developed a score test and results from two applications showed that we would not reject the null hypothesis that progression is exponential when the background risk is taken into account. A limitation of the examples is that the sample size and hence the power of the score test was limited, but with the emergence of HPV screening registry data, the statistical power will be much higher in future studies.

Another advantage of using exponential competing risks is that the estimated model parameters can be used in Markov simulation models that are used to estimate the health gains and harms (i.e., QALYs) of new screening algorithms over a patient's screening lifetime. More specifically, our model can be used to predict the impact of multiple rounds of screening and inform changes in the screening interval when moving towards a more personalised approach based on individual risk factors.

We identified the following limitations of our model. First, our model does not account for the sensitivity of test results, e.g., cytology test results in a cervical cancer setting. This is especially important since we occasionally use the date of a woman's last normal cytology result in place of CIN0/1 histology results (i.e., confirmed <CIN2+) for the left bound to achieve a smaller interval around the time of CIN2+ development. In cohort studies, the sensitivity of repeat cytology for detecting CIN2+ was reported to be 85%[21]. At this level of sensitivity, the bias introduced by the assumption of perfect cytology is expected to be small[3]. However, in settings where the sensitivity of the screening test is low, an extension of our work is to extend the model to include parameters for the sensitivity of screening tests to allow for more accurate cumulative risk estimation. Second, we did not use HPV test results in follow-up. A main reason was that clinical HPV assays are not intended to detect viral clearance and a considerable percentage of HPV-negative results in the follow-up are false negative [22,23]. Additionally, follow-up HPV tests need to provide full genotyping information in order to rule out a genotype switch, but such tests are not (widely) used in screening programs.

In conclusion, our model provides a flexible and interpretable approach for estimating CIN2+ risk that can be used to inform changes in the screening program and the incorporation of individual risk factors as we move towards a more personalized approach. Future work could explore incorporating test sensitivity as well as follow-up HPV test results into the model.

Additional information

Authors' contributions: Kelsi R Kroon: Conceptualization, Formal analysis, Methodology, Visualisation, Writing - Original Draft Preparation; **Johannes A Bogaards**: Conceptualization, Supervision, Writing - Reviewing and Editing; **Chris JLM Meijer**: Data curation, Writing - Reviewing and Editing; **Johannes Berkhof**: Conceptualization, Supervision, Writing - Reviewing and Editing. All authors gave final approval of the submitted version of the manuscript. The work reported in the paper has been performed by the authors, unless clearly specified in the text.

Ethics approval and consent to participate: The POBASCAM trial (Trial registration ID: NTR218) was approved by the Medical Ethics Committee of the VU University Medical Centre (Amsterdam, The Netherlands; no. 96/103) and the Ministry of Public Health (The Hague, The Netherlands; VWS no. 328650). All participants provided written informed consent. The post-treatment cohort study (Trial registration ID: NTR1468) was approved by the ethics board at all hospitals. All women provided written informed consent.

Funding: This work was supported by the Horizon 2020 research and innovation program of the European Commission [RISCC project, grant agreement No 847845] (to KRK, JAB, CJLMM, and JB). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of interest: KRK, JAB, CJLMM, and JB declare no conflicts of interests.

Data availability: The data that support the findings of this study are available from the corresponding author upon reasonable request. The code for the accompanying R package is publicly available on GitHub (https://github.com/kelsikroon/PICmodel).

References

1. Walboomers JM, Jacobs M V, Manos MM, et al. Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *J Pathol*. 1999;189(1):12-19. doi:10.1002/(SICI)1096-9896(199909)189:1<12::AID-PATH431>3.0.CO;2-F
2. Roden RBS, Stern PL. Opportunities and challenges for human papillomavirus vaccination in cancer. *Nat Rev Cancer*. 2018;18(4):240-254. doi:10.1038/nrc.2018.13
3. Witte BI, Berkhof J, Jonker MA. An EM algorithm for nonparametric estimation of the cumulative incidence function from repeated imperfect test results. *Stat Med*. 2017;36(21):3412-3421. doi:10.1002/sim.7373
4. Turnbull BW. Nonparametric Estimation of a Survivorship Function with Doubly Censored Data. *J Am Stat Assoc*. 1974;69(345):169-173. doi:10.1080/01621459.1974.10480146
5. Turnbull BW. The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1976;38(3):290-295. doi:https://doi.org/10.1111/j.2517-6161.1976.tb01597.x
6. Kaplan EL, Meier P. Nonparametric-Estimation from Incomplete Observations. *J Am Stat Assoc*. 1958;53(282):457-481. doi:Doi 10.2307/2281868
7. Cheung LC, Pan Q, Hyun N, et al. Mixture models for undiagnosed prevalent disease and interval-censored incident disease: applications to a cohort assembled from electronic health records. *Stat Med*. 2017;36(22):3583-3595. doi:10.1002/sim.7380
8. Boag JW. Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy. *J R Stat Soc Series B Stat Methodol*. 1949;11(1):15-44. doi:10.1111/j.2517-6161.1949.tb00020.x
9. Berkson J, Gage RP. Survival Curve for Cancer Patients Following Treatment. *J Am Stat Assoc*. 1952;47(259):501-515. doi:10.1080/01621459.1952.10501187
10. Jakobsen LH, Bøgsted M, Clements M. Generalized parametric cure models for relative survival. *Biometrical Journal*. 2020;62(4):989-1011. doi:10.1002/bimj.201900056
11. De Angelis R, Capocaccia R, Hakulinen T, Soderman B, Verdecchia A. Mixture models for cancer survival analysis: application to population-based data with covariates. *Stat Med*. 1999;18(4):441-454. doi:10.1002/(SICI)1097-0258(19990228)18:4<441::AID-SIM23>3.0.CO;2-M
12. Rijkaart DC, Berkhof J, Rozendaal L, et al. Human papillomavirus testing for the detection of high-grade cervical intraepithelial neoplasia and cancer: final results of the POBASCAM randomised controlled trial. *Lancet Oncol*. 2012;13(1):78-88. doi:10.1016/S1470-2045(11)70296-0
13. Kocken M, Helmerhorst TJ, Berkhof J, et al. Risk of recurrent high-grade cervical intraepithelial neoplasia after successful treatment: a long-term multi-cohort study. *Lancet Oncol*. 2011;12(5):441-450. doi:10.1016/S1470-2045(11)70078-X
14. Lange K. A Gradient Algorithm Locally Equivalent to the Em Algorithm. *Journal of the Royal Statistical Society Series B-Methodological*. 1995;57(2):425-437. <Go to ISI>://WOS:A1995QL97000006
15. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data Via the *EM* Algorithm. *J R Stat Soc Series B Stat Methodol*. 1977;39(1):1-22. doi:10.1111/j.2517-6161.1977.tb01600.x
16. Biernacki C, Celeux G, Govaert G. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Comput Stat Data Anal*. 2003;41(3-4):561-575. doi:Pii S0167-9473(02)00163-9 Doi 10.1016/S0167-9473(02)00163-9

17. Gelman A, Jakulin A, Pittau MG, Su YS. A weakly informative default prior distribution for logistic and other regression models. *Ann Appl Stat*. 2008;2(4). doi:10.1214/08-AOAS191
18. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. *Bayesian Data Analysis*. Chapman and Hall/CRC; 2013. doi:10.1201/b16018
19. Bulkmans NW, Rozendaal L, Snijders PJ, et al. POBASCAM, a population-based randomized controlled trial for implementation of high-risk HPV testing in cervical screening: design, methods and baseline data of 44,102 women. *Int J Cancer*. 2004;110(1):94-101. doi:10.1002/ijc.20076
20. Bleeker MCG, Meijer CJLM, Berkhof J. Follow-up after treatment for cervical intraepithelial neoplasia. *BMJ*. 2012;345(oct31 4):e7186-e7186. doi:10.1136/bmj.e7186
21. Verhoef L, Bleeker MCG, Polman N, et al. Colposcopy referrals and <scp>CIN3</scp> detection after triage by host cell <scp>DNA</scp> methylation and/or <scp>HPV</scp> genotyping in <scp>HPV</scp> positive women with low-grade cytology from a population-based Dutch primary <scp>HPV</scp> screening trial. *Int J Cancer*. 2025;156(5):1065-1073. doi:10.1002/ijc.35289
22. Polman NJ, Veldhuijzen NJ, Heideman DAM, Snijders PJF, Meijer CJLM, Berkhof J. HPV-positive women with normal cytology remain at increased risk of CIN3 after a negative repeat HPV test. *Br J Cancer*. 2017;117(10):1557-1561. doi:10.1038/bjc.2017.309
23. Polman NJ, Ebisch RMF, Heideman DAM, et al. Performance of human papillomavirus testing on self-collected versus clinician-collected samples for the detection of cervical intraepithelial neoplasia of grade 2 or worse: a randomised, paired screen-positive, non-inferiority trial. *Lancet Oncol*. 2019;20(2):229-238. doi:10.1016/S1470-2045(18)30763-0

Tables

Table 1. Power of the Gamma score test for different combinations of background risk (β), sample size (N) and shape parameter (k). Data sets were generated for each combination until 100 data sets had a successful model fit. The brackets show the number of data sets that were required to have 100 converged model fits for each simulation setting.

	N = 1000			**N = 10000**		
	k=1	**k=1.5**	**k=2**	**k=1**	**k=1.5**	**k=2**
Estimating β						
$\beta = 0.005$	0.09	0.11	0.84	0.05	1.00	1.00
$\beta = 0.01$	0.08	0.19	0.66	0.04	0.98	1.00
$\beta = 0.02$	0.07	0.11	0.48 (106)	0.05	0.96	1.00
Fixed β						
$\beta = 0$	0.04	0.47	0.99	0.05	1.00	1.00
$\beta = 0.005$	0.10	0.28	0.91	0.03	1.00	0.97
$\beta = 0.01$	0.08	0.11	0.61	0.04	0.97	0.85
$\beta = 0.02$	0.03	0.12	0.43	0.04	0.89	0.71

Table 2. Parameter estimates (95% confidence/credible intervals (CI/ CrI)) for the model fit with baseline covariates on the POBASCAM study with and without the Cauchy prior. Note that age is included as a continuous variable standardised to have a

mean of 0 and standard deviation of 0.5, HPV16 is a binary variable indicating the presence of HPV genotype 16 (HPV16 = 1) or other (HPV16 = 0), and cytology is defined as normal (cytology = 0) or abnormal corresponding to Pap 2 or worse (cytology = 1).

			Without priors or background risk	Without priors, with background risk	With priors, with background risk
Parameter			Estimate (95% CI)	Estimate (95% CI)	Estimate (95% CrI)
Prevalence	Intercept	$\hat{\theta}_{1,0}$	-5.62 (-8.07, -3.08)	-7.32 (-21.92, 7.27)	-5.47 (-7.32, -3.62)
	Effect of age	$\hat{\theta}_{1,1}$	-0.25 (-0.64, 0.14)	-0.22 (-0.62, 0.18)	-0.25 (-0.64, 0.14)
	Effect of HPV16	$\hat{\theta}_{1,2}$	1.21 (0.88, 1.53)	1.20 (0.87, 1.53)	1.19 (0.86, 1.51)
	Effect of cytology	$\hat{\theta}_{1,3}$	4.87 (2.42, 7.33)	6.57 (-8.03, 21.17)	4.72 (2.87, 6.58)
Progression	Intercept	$\hat{\theta}_{2,0}$	-3.39 (-3.59, -3.20)	-3.48 (-3.77, -3.19)	-3.51 (-3.79, -3.23)
	Effect of HPV16	$\hat{\theta}_{2,1}$	0.97 (0.75, 1.19)	1.22 (0.88, 1.57)	1.21 (0.86, 1.55)
Clearance		$\widehat{\lambda_2}$	-2.05 (-2.25, -1.85)	-1.59 (-2.01, -1.18)	-1.62 (-2.02, -1.22)
Log of background risk		$\log(\hat{\beta})$	--	-5.25 (-6.02, -4.47)	-5.25 (-6.02, -4.48)
	Log-likelihood		-2036.40	-2033.97	--
	Gamma score test p-value		0.68	0.95	--

Table 3. Parameter estimates for the post-treatment surveillance data application with and without the Cauchy prior. Note that age is included as a continuous variable standardised to have a mean of 0 and standard deviation of 0.5, histology variable is original baseline diagnosis (CIN2 = 0, CIN3 = 1).

			Without priors or background risk	Without priors, with background risk	With priors, with background risk
Parameter			Estimate (95% CI)	Estimate (95% CI)	Estimate (95% CrI)
Prevalence	Intercept	$\hat{\theta}_{1,0}$	-3.84 (-5.76, -1.91)	-7.21 (-58.59, 44.18)	-3.91 (-5.75, -2.08)
	Effect of age	$\hat{\theta}_{1,1}$	0.36 (-0.02, 0.74)	0.52 (-0.13, 1.17)	0.45 (-0.05, 0.96)
	Effect of CIN3	$\hat{\theta}_{1,2}$	1.35 (-0.58, 3.29)	4.24 (-46.62, 55.1)	1.02 (-0.69, 2.74)
Progression	Intercept	$\hat{\theta}_{2,0}$	-3.15 (-3.68, -2.61)	-2.27 (-3.1, -1.44)	-2.35 (-3.09, -1.6)
Clearance	Intercept	$\widehat{\lambda_2}$	-1.25 (-1.71, -0.79)	-0.21 (-0.79, 0.36)	-0.24 (-0.79, 0.31)
Log of background risk		$\log(\hat{\beta})$	--	-5.76 (-6.62, -4.9)	-5.72 (-6.56, -4.87)
	Log-likelihood		-301.84	-297.83	--
	Gamma score test p-value		0.01	0.39	

Figures

Figure 1. Schematic representation of the model using hypothetical cervical cancer screening data. The circle dots represent visits where CIN2+ was not present and the crosses represent the first visit where CIN2+ was diagnosed, the brackets give the corresponding left and right interval for each patient.

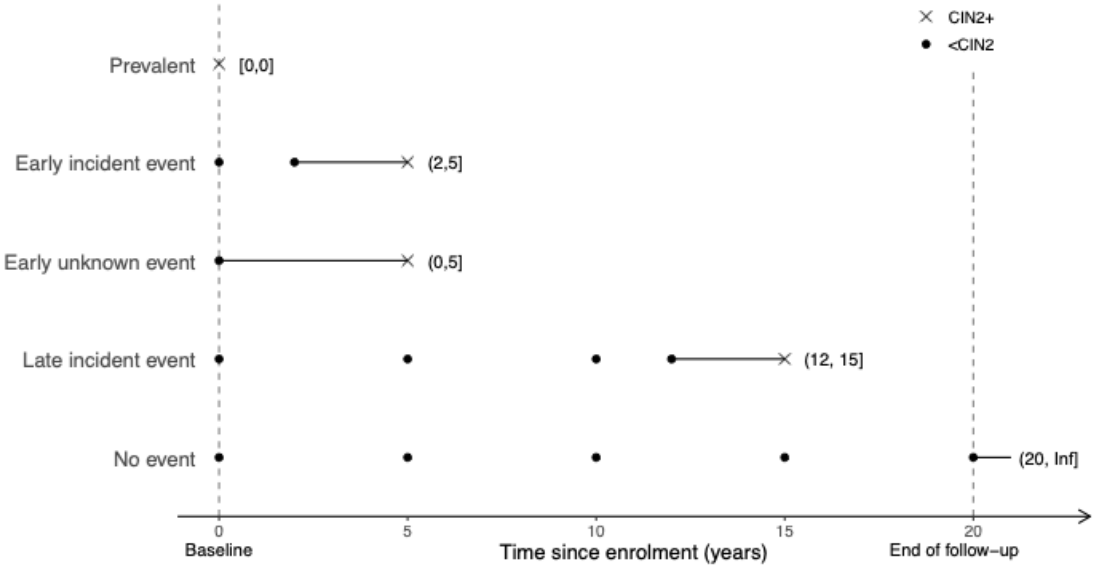

Figure 2. Simulation study results without covariates and with covariates for combinations of prior inclusion, sample size (N), attendance probability and value of background risk. Panels A and B show bias and coverage probability, respectively. Without covariates, parameters are prevalence ($\theta_{1,1}$), progression ($\theta_{2,1}$), clearance (λ_2), and background risk (β). With covariates, parameters are prevalence intercept ($\theta_{1,0}$), effect of age on prevalence ($\theta_{1,1}$), and effect of HPV16 on prevalence ($\theta_{1,2}$), progression intercept ($\theta_{2,0}$), effect of HPV16 on progression ($\theta_{2,1}$), clearance (λ_2), and background risk (β).

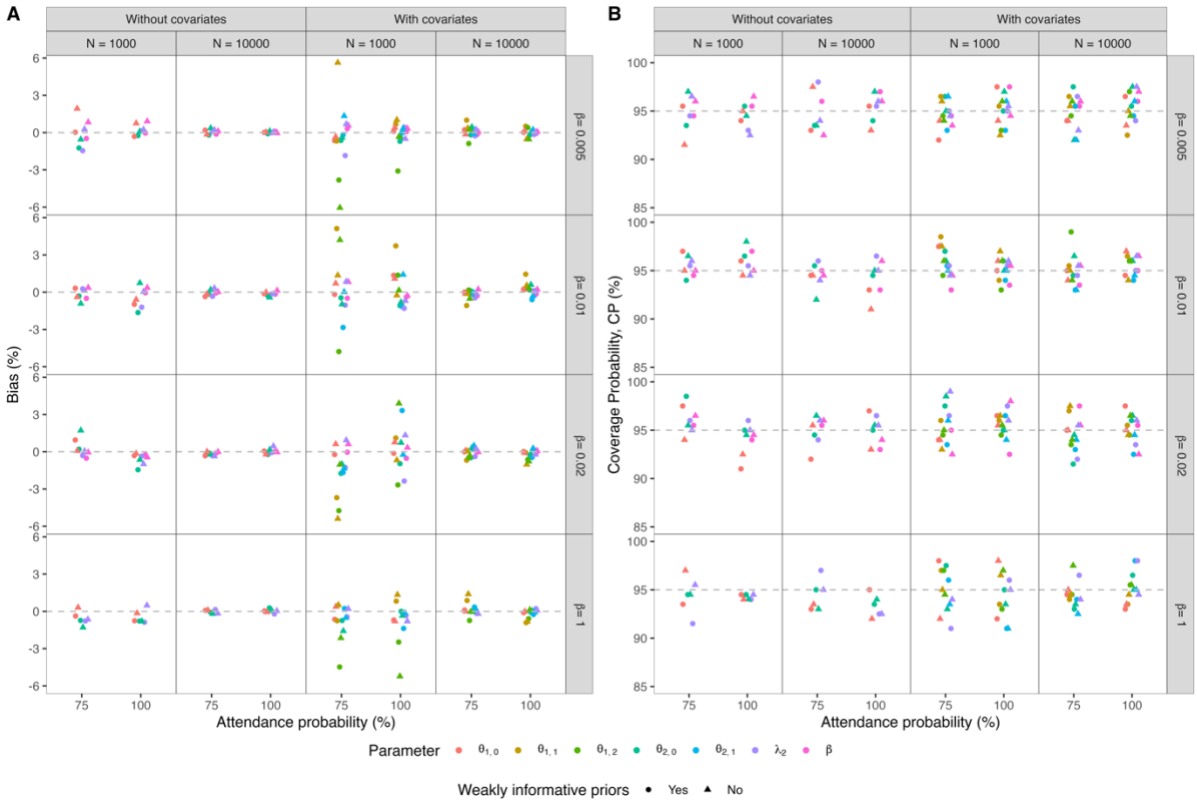
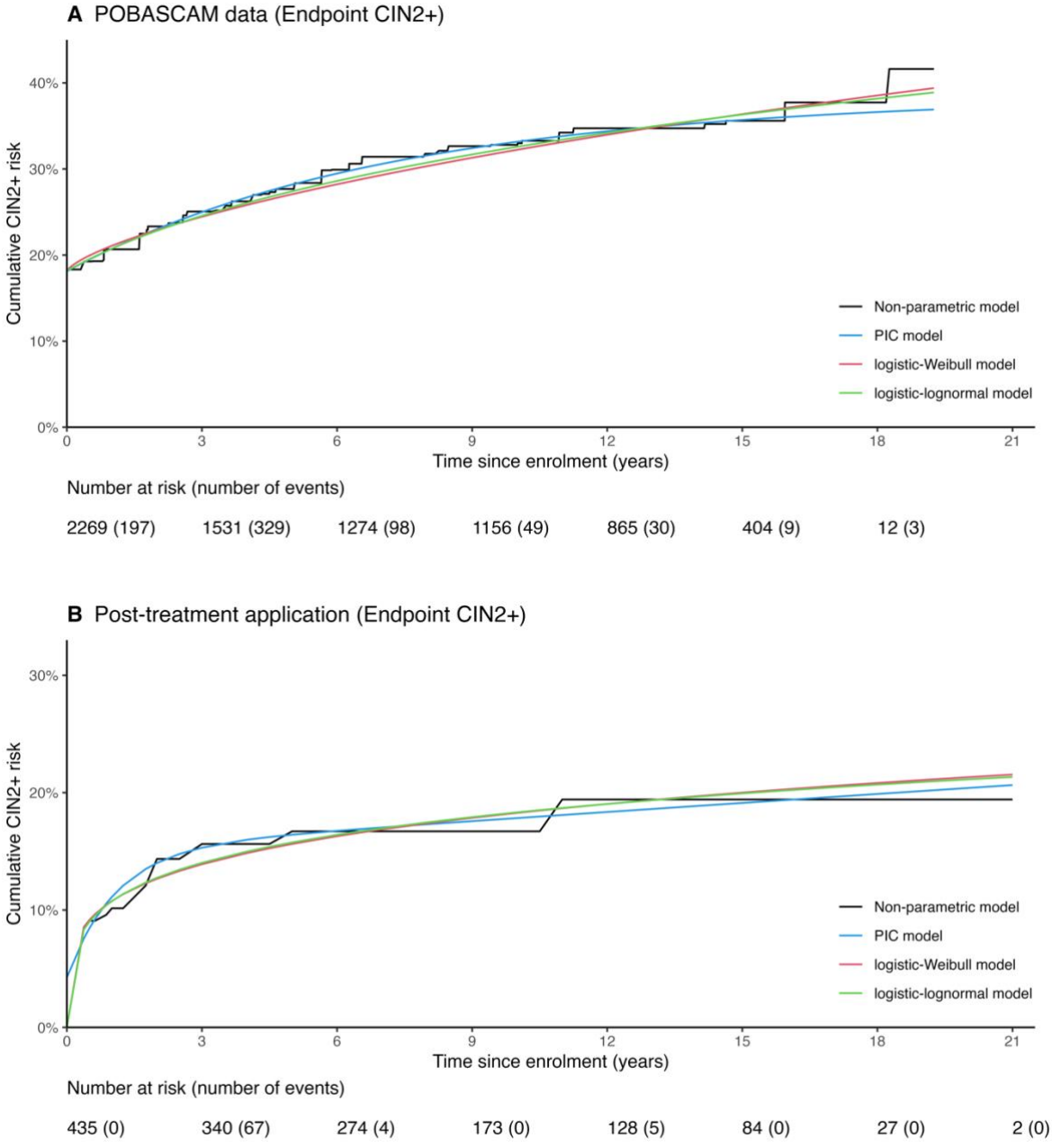
Figure 3. Comparison of model fit (without any baseline covariates) using three different methods for estimating the cumulative risk of CIN2+ in (**A**) HPV-positive women in the POBASCAM data set, and (**B**) women treated for CIN2+ from the post-treatment surveillance data set.