A prevalence-incidence mixture model for interval-censored screening and post-treatment surveillance data in a population with a temporarily increased disease risk

Running title: Prevalence-incidence mixture model

Authors and affiliations:

Kelsi R. Kroon*[1,2], Johannes A. Bogaards[1,2], Johannes Berkhof[1,2, 3]

[1] Department of Epidemiology and Data Science, Amsterdam UMC, location Vrije Universiteit Amsterdam, De Boelelaan 1117, Amsterdam, the Netherlands

[2] Amsterdam Public Health Research Institute, Methodology, Amsterdam, the Netherlands

3 Cancer Center Amsterdam, Imaging and Biomarkers, Amsterdam, the Netherlands

*** Corresponding Author:**

Kelsi R Kroon (k.kroon@amsterdamumc.nl, ORCiD: 0009-0005-7546-4410)

Address: Department of Epidemiology and Data Science, Amsterdam UMC – location Vrije Universiteit, De Boelelaan 1117, 1081 HV, Amsterdam, The Netherlands

---

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/PI3M).

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, 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, and JB declare no conflicts of interests.

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 participants provided written informed consent.

Authors' contributions: KRK: Conceptualization, Formal analysis, Methodology, Visualisation, Writing - Original Draft Preparation; **JAB**: Conceptualization, Supervision, Writing - Reviewing and Editing; **JB**: 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.

Acknowledgements: The authors thank C JLM Meijer for providing the POBASCAM data.

Abstract

To optimize screening and surveillance programs, it is essential to have information on the duration between the onset of a high-risk condition and treatable disease. Modelling of this duration is complicated when the high-risk condition only applies to some of the subjects and/or is temporary (e.g., oncogenic infections). Additionally, interval-censored data between visits makes exact disease onset time unknown and some of the disease cases may already be present at baseline. We propose a prevalence-incidence mixture model for interval-censored data to estimate cumulative disease risk based on individual risk factors and to estimate the mean duration of the high-risk condition. Subjects without the high-risk condition may also develop disease which is modelled by adding background risk. Parameters are estimated with an Expectation-Maximisation (EM) algorithm with weakly informative Cauchy priors. The duration between the baseline high-risk condition and treatable disease is directly obtained from the model. A score test is derived to check for poor fit. The algorithm is validated through simulation studies and applied to Dutch cervical cancer screening and post-treatment surveillance data. The screening example is longitudinal data of women who tested positive for human papillomavirus (HPV) at baseline. Some of these had undiagnosed, treatable disease (CIN2+) at baseline, some developed CIN2+ during follow-up, and some were assumed to clear the infection. The second example is surveillance data of women treated for CIN2+. They either had residual CIN2+ at baseline, were HPV-infected (without CIN2+) despite the treatment or were HPV-uninfected. The HPV and disease status at baseline were not observed. Our EM algorithm showed very good convergence for simulated data and the parameter estimates were close to the true values for realistic sample sizes. In the examples, our model had lower Akaike's Information Criterion values than models with Weibull, B-spline and lognormal hazards for incident events (difference 0.7 to 26.7). To conclude, our prevalence-incidence mixture model can provide an accurate fit to the data and can be used for designing risk-based screening and surveillance programs.

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

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. In screening and post-treatment surveillance, there are two potential scenarios for disease detection. Individuals who are screen-positive or treated for disease may have underlying *prevalent* disease at baseline, or disease might develop in-between two visits (*interval-censored incident disease*). Furthermore, incident disease can develop either from the baseline condition or from a newly acquired condition during follow-up. The baseline condition may only apply to a proportion of subjects and may be temporary so that the cumulative disease risk is likely to level off.

A motivating example is human papillomavirus (HPV) infections, which cause almost all cervical cancers worldwide[1] and are also associated with several other cancers and warts[2]. After an initial HPV infection, individuals are at increased risk of developing a high-grade cervical lesion (CIN2+) until the infection clears. To optimize a cervical cancer screening or post-treatment surveillance program, it is essential to estimate the duration between the high-risk condition and CIN2+.

Non-parametric and parametric models have widely been used for estimating the cumulative risk but they have 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 does not have covariates which are crucial when designing risk-based screening or surveillance programs. Parametric models allow for more flexibility with respect to the modelling of the effects of covariates. Prevalence-incidence mixture models have been proposed[7,8] which assume a logistic regression model for prevalent disease, some of which are undiagnosed at baseline, and a parametric model (e.g. Weibull) for incident disease. These models may provide an accurate fit of the incidence function, but do not provide credible estimates of the mean time to incident disease. This is because they assume that the cumulative incidence approaches 1 for large time T whereas the observed incidence is typically low in screening and post-treatment surveillance settings. A reason for the low incidence is that only a proportion of the subjects without prevalent disease at

60 baseline have a high-risk baseline condition[9]. For example, in post-treatment surveillance
61 studies, subjects are at high risk of a recurrent lesion when not all causative cells were
62 removed by the treatment. Furthermore, the high-risk condition may be temporary, for
63 example, because a competing event prevents progression from the high-risk condition to
64 disease. If the competing event is observed during follow-up, then a prevalence-incidence
65 competing risk model can be specified[10]. However, there are several situations where such a
66 model cannot be applied. For example, in screening studies the competing state is not
67 measured during follow-up when follow-up is done in accordance with national guidelines
68 and the new screening instrument has not yet been implemented.

69 We propose to capture the heterogeneity in the data through a prevalence-incidence
70 mixture model where a proportion of prevalent disease cases is undiagnosed at baseline
71 and subjects without prevalent disease have a temporarily increased disease risk that
72 converges with the background risk over time. Incident cases developed from the baseline
73 condition are assumed to follow a mixture of an exponential distribution and non-
74 progressive condition, whereas other incident disease cases are described by adding
75 background risk. The model for incident disease developed from the baseline condition is
76 mathematically equivalent to a model with an exponential time distribution for two
77 competing events, of which one is not observable[11]. To examine whether the model is
78 violated by the data, we present a score test comparing the exponential distribution for
79 incident disease from the baseline condition against a Gamma alternative.

80 The article is structured as follows. In Section 2, we describe the model and an expectation-
81 maximisation (EM) algorithm for estimating the model parameters with interval-censored
82 screening data, which has been implemented in an R package. In Section 3, we present
83 simulation studies to verify that the algorithm can accurately recover true parameter values
84 and a score test to check the exponential distribution for incident disease from the baseline
85 condition. In Section 4, we apply the model to estimate the cumulative risk of CIN2+ in two
86 different data sets. The first dataset consists of cytological and histological follow-up of a
87 screening population of women who tested positive for HPV at baseline. Some of these
88 women had undiagnosed CIN2+ at baseline, some developed CIN2+ during follow-up and
89 some were assumed to clear the infection. The second dataset consists of cytological and
90 histological follow-up of a cohort of patients who were treated for CIN2+ at baseline. Some

of these women had (undiagnosed) residual CIN2+ at baseline, some were assumed to be HPV-infected despite the treatment and were at high risk of developing recurrent CIN2+ during follow-up, and some were assumed to be HPV-uninfected at baseline. The HPV status at baseline was not measured. Finally, in Section 5 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

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

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 or from another condition (e.g. a new infection acquired after baseline). Events from the baseline condition follow a competing events framework, where the baseline condition

either progresses to the disease or transitions to an unobservable, competing state, both of which are assumed to follow an exponential distribution.

Let us assume a competing risk framework with progression parameter $\lambda_1(\boldsymbol{x_2}, \boldsymbol{\theta_2})$ and clearance parameter λ_2. Then the cumulative risk of developing disease from the baseline condition at time t is given by

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

In this specification, $\delta(\boldsymbol{x_2}, \boldsymbol{\theta_2}) = \lambda_1\,(\boldsymbol{x_2}, \boldsymbol{\theta_2})/(\lambda_1\,(\boldsymbol{x_2}, \boldsymbol{\theta_2}) + \lambda_2\,)$ is the probability of disease during follow-up developed from the baseline high-risk condition. The expected time from the high-risk condition and the event is given by $1/(\lambda_1\,(\boldsymbol{x_2}, \boldsymbol{\theta_2}\,) + \lambda_2\,)$. So far, we assumed that all subjects have a high-risk condition at baseline, but the model can also be applied when it is not known who is at high-risk of developing disease. In this situation, only a proportion of the subjects have a high-risk condition and either develop disease or transition to a competing, unobservable event. The expected duration from the high-risk condition to disease is still given by $1/(\lambda_1\,(\boldsymbol{x_2}, \boldsymbol{\theta_2}\,) + \lambda_2\,)$, but $\delta(\boldsymbol{x_2}, \boldsymbol{\theta}_2)$ is now the proportion of subjects who have a high-risk condition and develop disease from this high-risk condition. The progression parameter $\lambda_1(\boldsymbol{x_2}, \boldsymbol{\theta_2})$ can in both situations be interpreted as the derivative of the cumulative incidence $G\,(t; \boldsymbol{x}_2, \boldsymbol{\theta}_2, \lambda_2)$ at $t = 0$, i.e. $G'\,(0; \boldsymbol{x}_2, \boldsymbol{\theta}_2, \lambda_2)$.

For events that do not develop from the baseline condition, 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:

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

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:

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

where m_2 is the number of covariates used for the rate parameter λ_1.

In principle, the background risk β can also be linked to covariates, but this is not considered here as its regression coefficients tend to be less well identified as those of π and λ_1. The main role of β is to describe how the baseline hazard is influenced by the risk of an event that did not develop from the baseline condition.

In our model, the regression coefficient $\theta_{2,i}$ $(i \geq 1)$ can also be interpreted as the change in the log-odds of progression probability $\delta(\boldsymbol{x_2}, \boldsymbol{\theta}_2)$ associated with a one-unit increase in the corresponding covariate. To see this, let $\log\big(\lambda_1(\boldsymbol{x_2}, \boldsymbol{\theta}_2)\big) = \theta_{2,0} + \sum_{i=1}^{m_2} \theta_{2,i}x_{2,i} = \gamma(\boldsymbol{x_2}, \boldsymbol{\theta}_2)$, and $\log(\lambda_2) = \phi$. Then:

$$\delta(\boldsymbol{x_2}, \boldsymbol{\theta}_2) = \frac{\lambda_1(\boldsymbol{x_2}, \boldsymbol{\theta}_2)}{\lambda_1(\boldsymbol{x_2}, \boldsymbol{\theta}_2) + \lambda_2} = \frac{e^{-\phi+\gamma(\boldsymbol{x_2}, \boldsymbol{\theta}_2)}}{e^{-\phi+\gamma(\boldsymbol{x_2}, \boldsymbol{\theta}_2)} + 1}, \tag{7}$$

so the logit is linear in the covariates: $\text{logit}\, \delta(\boldsymbol{x_2}, \boldsymbol{\theta}_2) = -\phi + \gamma(\boldsymbol{x_2}, \boldsymbol{\theta}_2)$.

A variant of the traditional expectation-maximisation (EM) algorithm called the EM gradient algorithm[12] 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[13]. Details about the estimation can be found in the Supplementary Appendix.

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 parameter for this covariate, and although the algorithm may converge, it is likely that the confidence intervals for the parameter 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)[14], 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[15].

3. Simulation study

3.1 Validation of model fit

Simulation study I: 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 incident disease from the baseline condition, A_i, was drawn with probability $\delta(\boldsymbol{x_2}, \boldsymbol{\theta}_2)$. Additionally, two event times were generated for incident disease developed from the baseline condition (T_1) and from the new condition acquired after baseline (T_2). 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 and 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., $T_{\mathrm{D}} = \min(T_1, T_2)$.

Screening (or post-treatment surveillance) times were drawn from a Beta(20, 20) distribution scaled to have a mean around 3-yearly intervals until there was at least 25 years of follow-up. Attendance of each individual at each screening visit was drawn from a Bernoulli distribution. This resulted in left and right intervals representing the last time an individual was observed before disease onset and the first time 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 (0.75 or 1), (3) value of background risk ($\beta = 0, 0.005, 0.01$, or 0.02) and (4) weakly

informative prior (none or Cauchy(0,2.5)). 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 prevalence and progression probabilities in the model. The following model was used:

$$\begin{aligned} \text{logit}\big(\pi(\boldsymbol{x_1}, \boldsymbol{\theta_1})\big) &= \theta_{1,0} + \theta_{1,1} \times \text{Age} + \theta_{1,2} \times \text{HPV16} \\ \log\big(\lambda_1(\boldsymbol{x_2}, \boldsymbol{\theta_2})\big) &= \theta_{2,0} + \theta_{2,1} \times \text{HPV16} \end{aligned} \tag{8}$$

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 1). 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 ($\beta = 0.005$), imperfect attendance probability (0.75) and priors included.

Simulation study I: 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 1). 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 ranged between -1.45% and 1.93% at attendance probability of 0.75 and between -1.64% and 0.91% at perfect attendance. Percentage bias ranged from -1.30% to 1.93% without priors and from -1.64% to 0.95% with weakly informative 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, the absolute percentage bias was lower than for N=1000 (<1.45% across settings), and the coverage probability ranged 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), and 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.

3.2 Checking exponential progression rate

Score test

The exponential progression rate for events developed from the baseline condition may be violated by the data even after adjusting for competing, unobserved events and events that did not develop from the baseline condition. This limits the applicability of the model when used to support decision making for a screening or post-treatment 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 events developed from the baseline condition by devising a score test with a Gamma distribution as alternative. If the progression to disease follows a Gamma distribution with shape parameter k and rate parameter λ_1, then the cumulative incidence function for events developed from the baseline condition becomes

$$G(t;\lambda_1,\lambda_2,k) = \frac{\lambda_1^k}{(\lambda_1+\lambda_2)^k}\frac{1}{\Gamma(k)}\int_0^t (\lambda_1+\lambda_2)^k\, u^{k-1} e^{-(\lambda_1+\lambda_2)u} du, \tag{9}$$

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 $\lambda_1^k/(\lambda_1+\lambda_2)^k$.

The null hypothesis is that the Gamma shape parameter, k, in equation (9) 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 equation (9) 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 $(S(\widetilde{\boldsymbol{\theta}}) = U(\widetilde{\boldsymbol{\theta}})^2/I(\widetilde{\boldsymbol{\theta}}))$ which is asymptotically Chi-squared distributed with 1 degree of freedom under the null.

Simulation study II: 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 incident disease developed from the baseline condition, A_i, was drawn with probability $\lambda_1^k/(\lambda_1+\lambda_2)^k$ and T_1 was generated from a gamma distribution with shape parameter k and rate parameter $\lambda_1+\lambda_2$, according to equation (9). 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 II: results

[Place Table 1 near here] The results are in Table 1. For N=1000, the Type I error ranges from 0.03 to 0.10 and for N=10,000, it does not exceed 0.05. For N=1000, the power under *k* > 1 ranges from 0.43 to 0.99 when *k* = 2 and is negligible or low when *k* < 2. For N=10,000, the power ranges from 0.71 to 1.00.

4. Examples

4.1 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, and the maximum follow-up was 19 years. The study has been described before[16,17]. In this application, we used a subset of 2,269 women who were HPV-positive at their first screening date

278 (baseline). The date of the first histology test reporting CIN2+ was set as the right bound. If
279 an index smear was reported within 3 months before the first CIN2+ histology diagnosis, it
280 was used as the right bound instead. For the left bound, the most recent normal cytological
281 finding and/or histological finding reporting <CIN2 occurring before the right bound was
282 used.

283 The information loss, measured by the AIC, was low for our model without covariates (AIC =
284 4740.12) compared to the logistic-Bspline model[7] with five knots (AIC = 4740.83), the
285 logistic-Weibull (AIC = 4767.84) and the logistic-lognormal model (AIC = 4747.46).

286 [Place Table 2 near here] Table 2 gives parameter estimates with 95% confidence intervals,
287 log-likelihood values and score test results for our model with abnormal baseline cytology
288 defined as a binary variable in the prevalent disease regression equation, and HPV16
289 genotype defined as a binary variable in the prevalent and incident disease regression
290 equations. The selection of the variables was based on HPV literature, which reported
291 cytology as prognostic for prevalent disease and HPV16 genotype as prognostic for
292 prevalent and future disease[18,19]. The score test was not significant for the models with and
293 without background risk. Adding background risk improved the fit of the model (LR 5.56, p-
294 value 0.018), and adding the Cauchy prior reduced the width of the credible intervals in the
295 model with background risk.

296 For HPV16-positive women, the mean time to CIN2+ was 3.32 years (95% CI: 2.23, 4.43) and
297 the proportion of women who developed CIN2+ from the baseline infection was 0.33 (95%
298 CI: 0.29, 0.38). For HPV16-negative women, the mean time to CIN2+ was 4.34 years (95% CI:
299 2.83, 5.86) and the proportion of women who developed CIN2+ from the baseline condition
300 was 0.13 (95% CI: 0.10, 0.16).

301 **4.2 Cervical post-treatment follow-up data**

302 After treatment of CIN2+, there is a chance of residual disease due to incomplete excision
303 (prevalent disease) or HPV infected cells (the baseline condition) regrow to CIN2+. Women
304 also remain at risk of developing a new CIN2+ (background risk)[20]. We used data from a
305 long-term multi-cohort study following women after treatment for CIN2+ (Netherlands trial
306 registration ID: NTR1468)[21]. This cohort study included 435 women who were monitored
307 with HPV testing and cytology at 6, 12, and 24 months after treatment. Thereafter, women

308 were followed through the routine screening programme with 5-year intervals. The
309 maximum follow-up was 21 years.

310 We compared our model without covariates to other prevalence-incidence models using
311 Akaike's Information Criterion (AIC). Our model (AIC = 610.13) fitted slightly better than a
312 logistic-Bspline model with five knots (AIC = 617.39), logistic-Weibull (AIC = 614.64) and
313 logistic-lognormal model (AIC = 613.48).

314 [Place Table 3 near here] Table 3 gives parameter estimates with 95% confidence intervals,
315 log-likelihood values and score test results for three models with severity of disease at
316 treatment as covariate (CIN3: 1 if CIN3+, 0 if CIN2) in the prevalent and incident disease
317 regression equations. The score test was significant for a model without background risk (p-
318 value = 0.0047), but was not significant when background risk was included (p-value = 0.51)
319 indicating that an exponential progression model for events developed from the baseline
320 condition and background risk for other events was not violated by the data. Adding
321 background risk also significantly improved model fit (LR 8.52, p-value: 0.0035), but led to
322 wide confidence intervals for the intercept and effect of CIN3. Smaller credible intervals
323 could be obtained by adding a Cauchy prior.

324 For women with CIN3 at treatment, the expected time to recurrence was 1.05 years (95%
325 CI: 0.45, 1.64) and the proportion that developed disease was 0.13 (95% CI: 0.04, 0.21). For
326 women with CIN2 at treatment, the expected time to recurrence was 1.11 years (95% CI:
327 0.50, 1.72) and the proportion that developed disease was 0.07 (95% CI: 0.00, 0.14).

328 **5. Discussion**

329 We present a model for a population with a heterogenous disease risk that adds to the
330 previously introduced concept of prevalence-incidence mixture models for the cumulative
331 risk of disease[7,8]. By combining concepts from this model with additional parameters, we
332 were able to decompose a complex process into simpler key components that together
333 provide accurate estimates of the cumulative disease risk compared to non-parametric
334 estimates.

335 To model progression, a competing risks framework with exponential distribution for
336 disease progression was used which produced cumulative incidence curves with an AIC fit
337 comparable to or lower than those of existing methods (i.e., Weibull, lognormal, B-splines).

338 A specific feature of our model is that it separates events that are caused by the baseline
339 condition and events that are not caused by the baseline condition. Models that do not
340 separate these events are less suitable when the aim is to estimate the duration from the
341 baseline condition and treatable disease.

342 Our estimated model parameters can be used in Markov simulation models that are widely
343 used to estimate the health gains and harms (i.e., QALYs) of new screening algorithms over
344 a patient's screening lifetime. Our score test can be used to detect lack of fit and hence be
345 used to evaluate whether a Markov simulation model is violated by the data. If the model is
346 consistent with the data, our model can be used to predict the impact of multiple rounds of
347 screening and inform changes in the screening interval when moving towards a more
348 personalised approach based on individual risk factors. Furthermore, our model can be used
349 in post-treatment surveillance settings to assess how long the disease risk exceeds the
350 background risk, warranting surveillance.

351 A limitation of our model is that it does not account for the sensitivity of the screening test
352 results collected after baseline. This is important when determining the left and right bound
353 of the interval-censored data. The bias in the cumulative incidence is expected to be small
354 when the sensitivity is above 0.50, but can be substantial if the screening test sensitivity is
355 lower[3]. Notably, the sensitivity of repeat cytology for detecting CIN2+ is about 0.85[22].

356 Additionally, our model, when viewed as a competing risk model, does not use follow-up
357 measurements from the baseline condition in order to assess clearance. Our model was
358 motivated by cohort studies where the baseline condition is not reassessed at follow-up
359 measurements. A Weibull prevalence-incidence mixture model with a clearance event was
360 developed by Hyun (2020)[10]. This model still assumes that the high-risk condition can be
361 measured without error and a possible path for future research is to account for the limited
362 sensitivity of the screening test for assessing the high-risk condition during follow-up. A
363 Bayesian prevalence-incidence mixture model was recently proposed that takes
364 misclassification into account due to limited sensitivity of the screening test[23]. Such an
365 approach, adapted to a competing risk setting, could be beneficial when analysing large
366 data with detailed information recorded at every visit.

367 In conclusion, our model provides a flexible and interpretable approach for estimating
368 disease 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 of follow-up screening test results into the model.

References

1. Walboomers JM, Jacobs M V, Manos MM, Bosch FX, Kummer JA, Shah K V, et al. Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. J Pathol. 1999/08/19. 1999;189(1):12–9. doi:10.1002/(SICI)1096-9896(199909)189:1<12::AID-PATH431>3.0.CO;2-F PubMed PMID: 10451482.
2. Roden RBS, Stern PL. Opportunities and challenges for human papillomavirus vaccination in cancer. Nat Rev Cancer. 2018 Apr;18(4):240–54. doi:10.1038/nrc.2018.13 PubMed PMID: 29497146.
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 Sep 20;36(21):3412–21. doi:10.1002/sim.7373
4. Turnbull BW. Nonparametric Estimation of a Survivorship Function with Doubly Censored Data. J Am Stat Assoc. 1974 Mar;69(345):169–73. 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):290–5. 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):457–81. doi:Doi 10.2307/2281868 PubMed PMID: WOS:A1958WX09300012.
7. Cheung LC, Pan Q, Hyun N, Schiffman M, Fetterman B, Castle PE, et al. Mixture models for undiagnosed prevalent disease and interval-censored incident disease: applications to a cohort assembled from electronic health records. Stat Med. 2017/07/01. 2017;36(22):3583–95. doi:10.1002/sim.7380 PubMed PMID: 28660629.
8. Hyun N, Cheung LC, Pan Q, Schiffman M, Katki HA. Flexible risk prediction models for left or interval-censored data from electronic health records. Ann Appl Stat. 2017 Jun 1;11(2). doi:10.1214/17-AOAS1036
9. Srivastava S, Koay EJ, Borowsky AD, De Marzo AM, Ghosh S, Wagner PD, et al. Cancer overdiagnosis: a biological challenge and clinical dilemma. Nat Rev Cancer. 2019 Jun 25;19(6):349–58. doi:10.1038/s41568-019-0142-8
10. Hyun N, Katki HA, Graubard BI. Sample-weighted semiparametric estimation of cause-specific cumulative risk and incidence using left- or interval-censored data from electronic health records. Stat Med. 2020 Aug 15;39(18):2387–402. doi:10.1002/sim.8544
11. Titman AC, Sharples LD. Semi-Markov models with phase-type sojourn distributions. Biometrics. 2009/11/17. 2010;66(3):742–52. doi:10.1111/j.1541-0420.2009.01339.x PubMed PMID: 19912172.
12. Lange K. A Gradient Algorithm Locally Equivalent to the Em Algorithm. Journal of the Royal Statistical Society Series B-Methodological. 1995;57(2):425–37. PubMed PMID: WOS:A1995QL97000006.
13. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data Via the *EM* Algorithm. J R Stat Soc Series B Stat Methodol. 1977 Sep 1;39(1):1–22. doi:10.1111/j.2517-6161.1977.tb01600.x

14. Gelman A, Jakulin A, Pittau MG, Su YS. A weakly informative default prior distribution for logistic and other regression models. Ann Appl Stat. 2008 Dec 1;2(4). doi:10.1214/08-AOAS191
15. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis. Chapman and Hall/CRC; 2013. doi:10.1201/b16018
16. Bulkmans NW, Rozendaal L, Snijders PJ, Voorhorst FJ, Boeke AJ, Zandwijken GR, 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. 2004/04/01. 2004;110(1):94–101. doi:10.1002/ijc.20076 PubMed PMID: 15054873.
17. Rijkaart DC, Berkhof J, Rozendaal L, van Kemenade FJ, Bulkmans NW, Heideman DA, 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. 20111214th ed. 2012;13(1):78–88. doi:10.1016/S1470-2045(11)70296-0 PubMed PMID: 22177579.
18. Demarco M, Hyun N, Carter-Pokras O, Raine-Bennett TR, Cheung L, Chen X, et al. A study of type-specific HPV natural history and implications for contemporary cervical cancer screening programs. EClinicalMedicine. 2020 May;22:100293. doi:10.1016/j.eclinm.2020.100293
19. Inturrisi F, Rozendaal L, Veldhuijzen NJ, Heideman DAM, Meijer CJLM, Berkhof J. Risk of cervical precancer among HPV–negative women in the Netherlands and its association with previous HPV and cytology results: A follow-up analysis of a randomized screening study. PLoS Med. 2022 Oct 28;19(10):e1004115. doi:10.1371/journal.pmed.1004115
20. Bleeker MCG, Meijer CJLM, Berkhof J. Follow-up after treatment for cervical intraepithelial neoplasia. BMJ. 2012 Oct 31;345(oct31 4):e7186–e7186. doi:10.1136/bmj.e7186
21. Kocken M, Helmerhorst TJ, Berkhof J, Louwers JA, Nobbenhuis MA, Bais AG, et al. Risk of recurrent high-grade cervical intraepithelial neoplasia after successful treatment: a long-term multi-cohort study. Lancet Oncol. 2011 May;12(5):441–50. doi:10.1016/S1470-2045(11)70078-X
22. Verhoef L, Bleeker MCG, Polman N, Kroon KR, Steenbergen RDM, Ebisch RMF, et al. Colposcopy referrals and CIN3 detection after triage by host cell DNA methylation and/or HPV genotyping in HPV positive women with low-grade cytology from a population-based Dutch primary HPV screening trial. Int J Cancer. 2025 Mar 16;156(5):1065–73. doi:10.1002/ijc.35289
23. Klausch T, Lissenberg-Witte BI, Coupé VM. A Bayesian prevalence-incidence mixture model for screening outcomes with misclassification. 2026 Jan 9.

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 POBASCAM study application with and without the Cauchy prior

			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	$\hat{\theta}_{1,0}$	-5.28 (-6.84, -3.73)	-12.83 (-519.75, 494.09)	-5.66 (-7.73, -3.59)
	Effect of HPV16	$\hat{\theta}_{1,1}$	1.22 (0.9, 1.54)	1.22 (0.89, 1.54)	1.21 (0.88, 1.53)
	Effect of abnormal cytology	$\hat{\theta}_{1,3}$	4.55 (2.99, 6.11)	12.09 (-494.83, 519.02)	4.93 (2.86, 7)
Progression	Intercept	$\hat{\theta}_{2,0}$	-3.4 (-3.59, -3.21)	-3.47 (-3.73, -3.21)	-3.5 (-3.78, -3.23)
	Effect of HPV16	$\hat{\theta}_{2,1}$	0.96 (0.74, 1.18)	1.23 (0.89, 1.56)	1.21 (0.87, 1.55)
Clearance		$\widehat{\lambda_2}$	-2.05 (-2.25, -1.85)	-1.58 (-1.97, -1.2)	-1.61 (-2.01, -1.21)
Log of background risk		$\log(\hat{\beta})$		--	-5.23 (-5.98, -4.49)
	Log-likelihood		-2037.36	-2034.58	--
	Gamma score test p-value		0.97	0.96	0.49

HPV16: human papillomavirus genotype 16

Table 3. Parameter estimates (95% confidence/credible intervals (CI/ CrI)) for the post-treatment surveillance data application with and without the Cauchy prior.

			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	$\hat{\theta}_{1,0}$	-3.60 (-5.35, -1.85)	-4.33 (-8.7, 0.04)	-3.6 (-5.21, -1.99)
	Effect of CIN3 [a]	$\hat{\theta}_{1,1}$	1.15 (-0.66, 2.95)	1.39 (-2.96, 5.74)	0.62 (-1.05, 2.29)
Progression	Intercept	$\hat{\theta}_{2,0}$	-3.41 (-4.31, -2.51)	-2.73 (-4.04, -1.42)	-2.73 (-3.8, -1.65)
	Effect of CIN3 [a]	$\hat{\theta}_{2,1}$	0.34 (-0.57, 1.25)	0.58 (-0.67, 1.83)	0.60 (-0.42, 1.62)
Clearance	Intercept	$\widehat{\lambda_2}$	-1.24 (-1.69, -0.78)	-0.20 (-0.77, 0.36)	-0.18 (-0.73, 0.37)
Log of background risk		$\log(\hat{\beta})$	--	-5.76 (-6.61, -4.91)	-5.71 (-6.53, -4.88)
	Log-likelihood		-303.13	-298.88	--
	Gamma score test p-value		0.0047	0.51	0.24

[a] Histology variable is original baseline diagnosis (CIN2 = 0, CIN3 = 1).

CIN2/3: cervical intraepithelial neoplasia grade 2/3

Figures
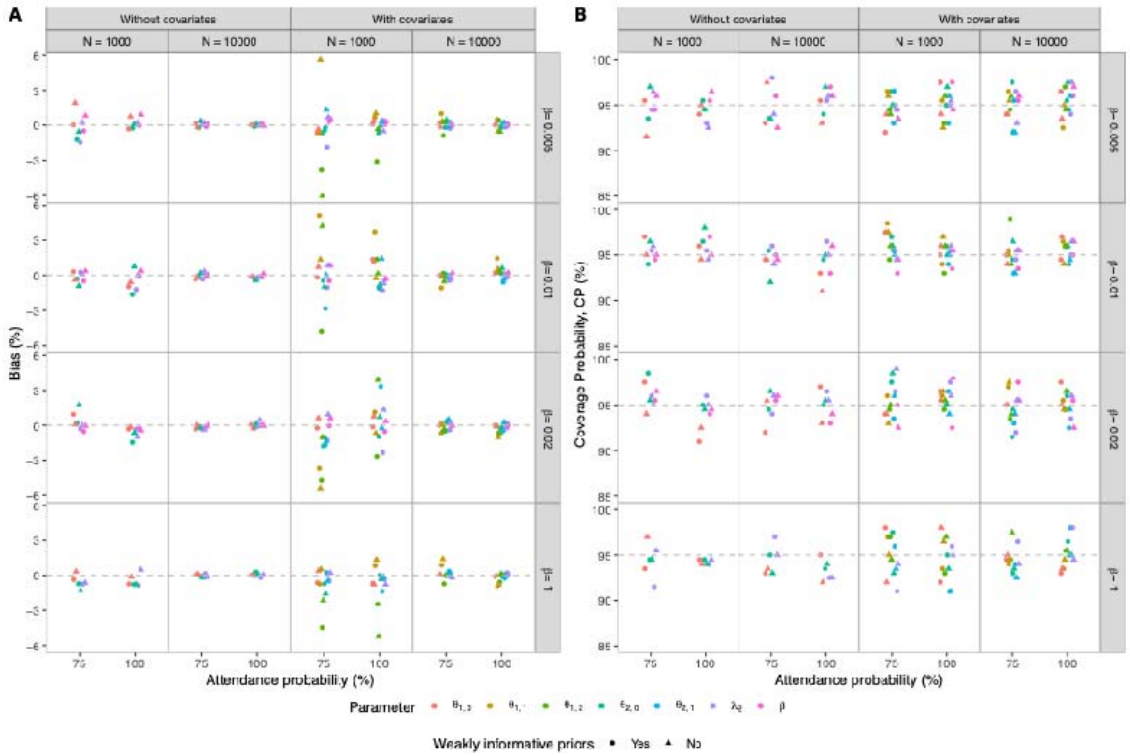
Figure 1. Simulation study results without covariates and with covariates for combinations of prior inclusion, sample size (N), attendance probability and value of background risk. Panels A and B show bias and coverage probability, respectively. Without covariates, parameters are prevalence ($\theta_{1,1}$), 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).