Research Article

Statistics in Medicine

Received 17 August 2016, Accepted 23 May 2017 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.7373

An EM algorithm for nonparametric estimation of the cumulative incidence function from repeated imperfect test results

Birgit I. Witte,[a*†] **Johannes Berkhof**[a] **and Marianne A. Jonker**[a,b]

In screening and surveillance studies, event times are interval censored. Besides, screening tests are imperfect so that the interval at which an event takes place may be uncertain. We describe an expectation–maximization algorithm to find the nonparametric maximum likelihood estimator of the cumulative incidence function of an event based on screening test data. Our algorithm has a closed-form solution for the combined expectation and maximization step and is computationally undemanding. A simulation study indicated that the bias of the estimator tends to zero for large sample size, and its mean squared error is in general lower than the mean squared error of the estimator that assumes the screening test is perfect. We apply the algorithm to follow-up data from women treated for cervical precancer. Copyright © 2017 John Wiley & Sons, Ltd.

Keywords: cervical post-treatment surveillance; EM algorithm; interval censoring; nonparametric maximum likelihood estimation; screening test

1. Introduction

In many longitudinal studies, the primary objective is to estimate the cumulative incidence of an event. In case of a manifest event, such as death or clinical disease, time to event is unambiguous. However, the occurrence of an asymptomatic event can only be detected by a diagnostic test. Such a test is imperfect, which means that false positive or false negative results may occur. If the diagnostic test is positive, an invasive confirmatory test is performed, which is considered to be a gold standard test with sensitivity and specificity of 100%. Screening and surveillance studies have as end point a subclinical (asymptomatic) disease stage, in which the diagnostic test is performed repeatedly. This leads to mixed case interval censored observations, where the time to event is only known up to a time interval and patients are allowed to differ with regard to the number of follow-up times.

A motivating example is a Dutch surveillance study in which 435 women treated for cervical precancer (CIN2/3) were monitored outside the hospital for recurrent disease by a cytological diagnostic test [1]. Women with abnormal cytology were referred to the gynecologist for a confirmatory diagnosis. The aim of the study was to evaluate the safety of the monitoring schedule by estimating the cumulative incidence of recurrent CIN2/3.

Several authors have considered estimation of a cumulative incidence function based on mixed case interval censored data with possibly missed events. Richardson and Hughes [2] study the nonparametric maximum likelihood estimator (MLE) under the assumption of discrete censoring times. More specifically, they assume that subjects are tested at every scheduled time point until they are either right censored or have a positive test. In practice, subjects have incomplete follow-up, and actual time points differ from scheduled time points. For example, in the surveillance study of Kocken *et al.* [1], women were

[a] *Department of Epidemiology and Biostatistics, Amsterdam Public Health Research Institute, VU University Medical Center, Amsterdam, The Netherlands*

[b] *Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, The Netherlands*

[*] *Correspondence to: Birgit I. Witte, Department of Epidemiology and Biostatistics, Amsterdam Public Health Research Institute, VU University Medical Center, Amsterdam, The Netherlands.*

[†] *E-mail: B.Witte@vumc.nl*

scheduled to be screened at three time points, but 735 unique follow-up times were observed. Moreover, some women had fewer than three follow-up screens (5.6%), and many women had at least four screens (40.8%) instead of three. A model that assumes continuous screening times seems an adequate solution to allow for incomplete and unscheduled follow-up. Balasubramanian and Lagakos [3] assume continuous screening times and propose sequential quadratic programming (SQP) methods to maximize the nonparametric log-likelihood. SQP methods have been successful in small and medium but generally not in high-dimensional problems [4]. The maximization problem can become high dimensional if each subject has his or her own distinct follow-up times. Then, SQP methods tend to be slow and may even fail to converge because they involve (high-dimensional) matrix inversion. More recently, Titman [5] also considered the estimation of the cumulative incidence from imperfect test results and studied the iterative convex minorant algorithm [6] and an adapted version of the constrained Newton algorithm of Wang [7]. In the two scenarios that Titman [5] studies, follow-up is censored either directly after the first positive screening test or only at the end of study, whereas in our setting, follow-up is censored after the first positive confirmatory test.

In this article, we assume, like in Balasubramanian and Lagakos [3] and Titman [5], that the screening times are continuous, which is usually most appropriate for screening and surveillance studies. However, we propose an expectation–maximization (EM) algorithm for maximization of the log-likelihood that takes into account that screening tests are imperfect. We show that the combined expectation (E) and maximization (M) steps of this algorithm has a closed-form solution and does not involve (high-dimensional) matrix inversion so that the algorithm is fast even when the number of distinct follow-up times is large. The R-code of our algorithm is available in the Supporting Information and can be applied to data from screening or surveillance studies.

The outline of this article is as follows. We introduce the model and define the MLE in Section 2 and describe the EM algorithm in Section 3. We perform a simulation study in Section 4 and apply our method to the post-treatment surveillance data of Kocken *et al.* [1] in Section 5. We end with a discussion in Section 6. We provide detailed derivations in the Appendix.

2. Definitions, assumptions, and the likelihood

In this section, we introduce the mixed case interval censoring model that takes into account that the screening test is imperfect and define the nonparametric MLE for the cumulative incidence function of the event of interest. We first consider the screening process and impose several assumptions. Then, because this underlying process is not observed completely, we describe the observed data and derive the log-likelihood.

2.1. Screening process and assumptions

Let T be the time to event of interest with cumulative incidence function F and let C be the screening time at which occurrence of the event is assessed. Define by D the true status of the event at time C, that is,

$$D = \begin{cases} 0, & \text{if the event did not yet occur } (T > C), \\ 1, & \text{if the event occurred } (T \leqslant C), \end{cases}$$

and by X the outcome of the screening test, that is,

$$X = \begin{cases} 0, & \text{if the screening test is negative,} \\ 1, & \text{if the screening test is positive.} \end{cases}$$

Denote the sensitivity of this test by $\alpha_1 = \mathrm{P}(X=1 \mid D=1)$ and the specificity by $\alpha_2 = \mathrm{P}(X=0 \mid D=0)$. In screening and surveillance programs, a positive screening test is followed by a confirmatory test whose outcome at time C is denoted by

$$Y = \begin{cases} 0, & \text{if the confirmatory test is negative,} \\ 1, & \text{if the confirmatory test is positive.} \end{cases}$$

In general, the status of the event is not checked once but repeatedly at random, continuous, follow-up times $0 < C_1 < C_2 < \ldots$ until the end of the study. Define K as the (random) number of follow-up times up to this moment. Denote the status of the event at a time C_j by D_j and the outcomes of the screening and confirmatory tests by X_j and Y_j, respectively. Throughout, we assume that

(A.1) the event time T is independent of the number of follow-up visits K and the follow-up times $C_1, C_2, \dots, C_K$,

(A.2) the outcomes of the screening tests are independent of each other conditional on $K = k$ and $D_1, D_2, \dots, D_k$:

$$\mathrm{P}\left(X_1 = x_1, \dots, X_k = x_k \mid K = k, D_1, \dots, D_k\right) = \prod_{j=1}^{k} \mathrm{P}\left(X_j = x_j \mid D_j\right),$$

(A.3) the events $\{X_j = 1\}$ and $\{Y_j = y\}$ are independent conditional on D_j:

$$\mathrm{P}\left(X_j = 1, Y_j = y \mid D_j\right) = \mathrm{P}\left(X_j = 1 \mid D_j\right) \mathrm{P}\left(Y_j = y \mid D_j\right), \text{and}$$

(A.4) the confirmatory test has perfect sensitivity and specificity; that is, both $\mathrm{P}(Y_j = 1 \mid D_j = 1)$ and $\mathrm{P}\left(Y_j = 0 \mid D_j = 0\right)$ are equal to 1.

2.2. *Observed data and the likelihood*

We assume that screening is stopped after a confirmed event. This means that the screening times $C_1, C_2, \dots, C_K$ are observed until the time at which a positive screening test is observed and confirmed by a positive confirmatory test. Moreover, Y_j is only observed when $X_j = 1$. Therefore, define by Z_j the outcome of the combined screening and confirmatory test at a time C_j, that is,

$$Z_j = \begin{cases} 0, & \text{if } X_j = 0, \\ Y_j, & \text{if } X_j = 1, \end{cases}$$

and by E the variable indicating whether or not an event was observed during the follow-up; that is, $E = 1$ if the event was observed and $E = 0$ if not. Denote by R the number of observed screening tests, that is, $R = \min\{j : Z_j = 1\}$ if $E = 1$ and $R = K$ if $E = 0$ and vectors of length R by boldface letters, that is, $\mathbf{C} = \left(C_1, \dots, C_R\right)$. Then, instead of the complete process $\left(K, C_1, \dots, C_K, X_1, \dots, X_K, Y_1, \dots, Y_K\right)$, the process $W = (R, \mathbf{C}, \mathbf{X}, \mathbf{Z}, E)$ is observed.

Based on n i.i.d. observations of W, we estimate the cumulative incidence function F as well as the sensitivity α_1 and specificity α_2 by maximizing the log-likelihood

$$\ell\left(F, \alpha_1, \alpha_2\right) = \sum_{i=1}^{n} \log h_{F,\alpha_1,\alpha_2}\left(R_i, \mathbf{C}_i, \mathbf{X}_i, \mathbf{Z}_i, E_i\right) \tag{1}$$

over the class of all distribution functions $\mathcal{F}$, and $\alpha_1, \alpha_2 \in [0, 1]$. Note that the log-likelihood $\ell\left(F, \alpha_1, \alpha_2\right)$ is conditional on the observed data $W_1, \dots, W_n$, but this is omitted in the notation. In the Appendix, we show that h_{F,α_1,α_2}, the density of one observation $w = (r, \mathbf{c}, \mathbf{x}, \mathbf{z}, e)$ of W, can be decomposed into a product of two terms. The first term is a function of the parameters of interest F, α_1, and α_2. The second term is a function of the density of the underlying screening process $(K, C_1, \dots, C_K)$ and can be ignored when maximizing the log-likelihood (1) because it does not depend on F, α_1, and α_2. In the Appendix, we show, under assumptions (A.1–A.4), that the first term of the log-likelihood is equal to

$$\begin{aligned} &\left(1 - \alpha_2\right)^{l^+} {\alpha_2}^{l-l^+} \times \\ &\left\{ \sum_{j=l+1}^{r} {\alpha_2}^{j-l-1} \left(1 - \alpha_1\right)^{r-j} {\alpha_1}^{e} \left(1 - \alpha_1\right)^{1-e} \left(F\left(c_j\right) - F\left(c_{j-1}\right)\right) + (1 - e){\alpha_2}^{r-l} \left(1 - F\left(c_r\right)\right) \right\}. \end{aligned} \tag{2}$$

Here, l is the observed value of L that represents the index of the last confirmed false positive screening test (and $L = 0$ if there are no screening tests with $X_j = 1$ and $Y_j = 0$). Furthermore, l^+ is the observed value of $L^+ = \sum_{j=1}^{L} X_j$ that represents the number of confirmed false positive screening tests up to and including time C_L (and $L^+ = 0$ if $L = 0$).

3. The expectation–maximization algorithm

For given α_1 and α_2, the MLE $\hat{F}_{\alpha_1,\alpha_2}$ for F can be obtained by the EM algorithm [8]. In this section, we derive a closed-form solution for the combined E and M step of the EM algorithm. Subsequently, the MLE for (F, α_1, α_2) can be obtained via a grid search over all $(\alpha_1, \alpha_2) \in [0,1]^2$. At every point (α_1, α_2) on the grid, $\hat{F}_{\alpha_1,\alpha_2}$ and the corresponding value of the log-likelihood $\ell(\hat{F}_{\alpha_1,\alpha_2}, \alpha_1, \alpha_2)$ are computed and the MLE for (F, α_1, α_2) is defined as the maximizer of $\ell(\hat{F}_{\alpha_1,\alpha_2}, \alpha_1, \alpha_2)$ on the grid.

Let C_{ij} denote the jth follow-up time of subject i. Note that only the values of F at the follow-up times $C_{iL_i}, \dots, C_{iR_i}$ $(i = 1, 2, \dots, n)$ matter when maximizing the log-likelihood, because the log-likelihood only depends on F via these points. Therefore, maximization of the relevant part of (1) can be restricted to a class of purely discrete distribution functions F whose corresponding density f assigns mass to only these time points (and possibly to an additional point to give the estimate mass 1). Denote by $\tau_1, \tau_2, \dots, \tau_U$ the ordered U distinct values of all follow-up times $C_{1L_1}, \dots, C_{1R_1}, \dots, C_{nL_n}, \dots, C_{nR_n}$. Let $\tau_{U+1} > \tau_U$ be an additional point for the remaining mass and define $p_u = f(\tau_u)$.

Define by $p_u^{(m)}$ the probability that T is equal to τ_u in the mth iteration, that is, $p_u^{(m)} = \mathrm{P}^{(m)}(T = \tau_u)$, and let $\mathbf{p}^{(m)} = \left(p_1^{(m)}, \dots, p_{U+1}^{(m)}\right)$ be the vector of these probabilities such that $\sum_{u=1}^{U+1} p_u^{(m)} = 1$. In the E step of the mth iteration of the EM algorithm, the conditional expectation of the log-likelihood for the full likelihood conditional on the observed data is taken under $\mathbf{p}^{(m)}$ and equals

$$\mathrm{E}^{(m)}\left\{ \sum_{i=1}^{n} \log f(T_i) \,\middle|\, W_1, \dots, W_n \right\} = \sum_{i=1}^{n} \mathrm{E}^{(m)}\left(\log f(T_i) \mid W_1, \dots, W_n \right) =$$
$$= \sum_{i=1}^{n} \left\{ \sum_{u=1}^{U+1} \log f(\tau_u) \cdot \mathrm{P}^{(m)}\left(T_i = \tau_u \mid W_i\right) \right\} = \sum_{u=1}^{U+1} \log p_u \sum_{i=1}^{n} \mathrm{P}^{(m)}\left(T_i = \tau_u \mid W_i\right).$$

In the M step, the conditional expectation is maximized with respect to $\mathbf{p} = (p_1, \dots, p_{U+1})$ over the set $\mathcal{P} = \left\{ \mathbf{p} \in [0,1]^{U+1} : \sum_{u=1}^{U+1} p_u = 1 \right\}$. By the Lagrange multiplier method, it follows that the maximizer in the M step satisfies

$$p_u^{(m+1)} = n^{-1} \sum_{i=1}^{n} \mathrm{P}^{(m)}\left(T_i = \tau_u \mid W_i\right). \tag{3}$$

The probability $\mathrm{P}^{(m)}(T_i = \tau_u \mid W_i)$ in (3) can be computed explicitly. Define

$$F^{(m)}(t) = \sum_{u:\tau_u \leqslant t} \mathrm{P}^{(m)}\left(T = \tau_u\right) = \sum_{u:\tau_u \leqslant t} p_u^{(m)};$$

then, in case of a confirmed event (i.e., $E_i = 1$), it holds that

$$\mathrm{P}^{(m)}\left(T_i = \tau_u \mid W_i\right) = p_u^{(m)} \frac{\sum_{j=L_i+1}^{R_i} \alpha_2^{j-L_i-1}\left(1-\alpha_1\right)^{R_i-j} \alpha_1 1_{(C_{ij-1}, C_{ij}]}\left(\tau_u\right)}{\sum_{j=L_i+1}^{R_i} \alpha_2^{j-L_i-1}\left(1-\alpha_1\right)^{R_i-j} \alpha_1 \left(F^{(m)}\left(C_{ij}\right) - F^{(m)}\left(C_{ij-1}\right)\right)}, \tag{4}$$

where $1_A(x)$ is the indicator function, that is, $1_A(x)$ is equal to one if $x \in A$ and zero otherwise. A derivation of (4) can be found in the Appendix. In an analogous way, it can be shown that if $E_i = 0$,

$$\mathrm{P}^{(m)}\left(T_i = \tau_u \mid W_i\right) = p_u^{(m)} \times$$
$$\frac{\sum_{j=L_i+1}^{R_i} \alpha_2^{j-L_i-1}\left(1-\alpha_1\right)^{R_i-j+1} 1_{(C_{ij-1}, C_{ij}]}\left(\tau_u\right) + \alpha_2^{R_i-L_i} 1_{\left(C_{iR_i}, \infty\right)}\left(\tau_u\right)}{\sum_{j=L_i+1}^{R_i} \alpha_2^{j-L_i-1}\left(1-\alpha_1\right)^{R_i-j+1} \left(F^{(m)}\left(C_{ij}\right) - F^{(m)}\left(C_{ij-1}\right)\right) + \alpha_2^{R_i-L_i}\left(1 - F^{(m)}\left(C_{ij}\right)\right)}. \tag{5}$$

Combining the E and M step of the algorithm by substituting the probabilities $\mathrm{P}^{(m)}\left(T_i = \tau_u \mid W_i\right)$ of (4) and (5) into (3) yields the estimated probability after $m+1$ iterations

$$p_u^{(m+1)} = n^{-1} \sum_{i=1}^{n} p_u^{(m)} \times \frac{\sum_{j=L_i+1}^{R_i+1-E_i} \alpha_2{}^{j-L_i-1} \left(1-\alpha_1\right)^{R_i-j} \alpha_1{}^{E_i} \left(1-\alpha_1\right)^{1-E_i} 1_{(C_{ij-1},C_{ij}]}(\tau_u)}{\sum_{j=L_i+1}^{R_i+1-E_i} \alpha_2{}^{j-L_i-1} \left(1-\alpha_1\right)^{R_i-j} \alpha_1{}^{E_i} \left(1-\alpha_1\right)^{1-E_i} \left(F^{(m)}\left(C_{ij}\right) - F^{(m)}\left(C_{ij-1}\right)\right)}, \tag{6}$$

where by definition $C_{iR_i+1} = \tau_{U+1}$ and $F^{(m)}(\tau_{U+1}) = 1$. The updating formula in (6) has a closed form that makes the computation of $\mathbf{p}^{(m+1)}$ undemanding.

A possible starting vector $\mathbf{p}^{(0)}$ for the EM algorithm is the naive MLE that ignores the imperfection of the screening test, defined as

$$\tilde{F} = \underset{F \in \mathcal{F}}{\operatorname{argmax}} \sum_{i=1}^{n} \left\{ E_i \log\left(F\left(C_{iR_i}\right) - F\left(C_{iR_i-1}\right)\right) + (1-E_i) \log\left(1 - F\left(C_{iR_i}\right)\right) \right\}.$$

The EM algorithm is terminated when the change in the value of the log-likelihood becomes small. Pointwise confidence bounds for F can be obtained via nonparametric bootstrap [9], as the central limit theory is presumably not applicable [10, 11].

4. Simulation study

In this section, we compare the MLE $\hat{F}$ with the naive estimator $\tilde{F}$ and assess the performance of the EM algorithm in a simulation study where we simulate data according to the following settings. The event time T has a Weibull distribution with shape parameter θ (set at $2/3$, 1, and $3/2$) and scale parameter λ (set at $2/\theta$, $10/\theta$, and $100/\theta$). The values of the shape parameter correspond to a monotone decreasing, constant or monotone increasing hazard function, and the values of the scale parameter correspond to a short (1.2–4.0), a moderate (6.0–20), and a long (60–200) mean time to event. For each subject, the number of visits K is randomly selected from the set $\{1, 2, \dots, 16\}$. Although throughout this paper we assume the screening times $C_1, C_2, \dots, C_K$ are continuous, for ease of computations (and without loss of generality), in this simulation study we simulate them from a discrete distribution. More specifically, they are randomly selected from the set $\{0.5, 1, 1.5, 2, 3, 4, 5, 6, 8, 10, 12, 14, 16, 18, 20, 25\}$, where each value has equal probability of being selected. Under these follow-up times, the cumulative incidence at the end of follow-up (i.e., $t = 25$) is above 75% for $\lambda = 2/\theta$ or $\lambda = 10/\theta$ and below 25% for $\lambda = 100/\theta$. The result of the screening test X_j at time C_j ($j = 1, \dots, K$) has a Bernoulli distribution with parameter $\left(1-\alpha_2\right)$ for $C_j < T$ and α_1 for $C_j \geq T$, where $\left(\alpha_1, \alpha_2\right)$ are set at $(0.5, 0.5)$, $(0.5, 0.9)$, and $(0.9, 0.9)$. The sample size n is set at 100 and 1000. For each simulation setting, we draw $B = 1000$ data sets. We terminate the EM algorithm when the change in the log-likelihood is less than 10^{-9}.

We follow two different approaches with respect to the estimation of the sensitivity and specificity: (i) only α_2 is estimated simultaneously with F, while α_1 is treated as fixed, and (ii) both α_1 and α_2 are estimated simultaneously with the distribution function F. In both approaches, we apply a grid search algorithm: first on a coarse grid of size 0.1, subsequently on a refined grid of size 0.01 around the maximizer on the 0.1 grid, and finally on a grid of size 0.001 around the maximizer over the grid 0.01.

4.1. *Comparison of the maximum likelihood estimator and the naive estimator*

The bias (i.e., the difference between the estimator and true value) and mean squared error (MSE) of $\hat{F}$ and $\tilde{F}$ are estimated at several fixed time points t_0 (set at 1, 2, 5, 10, and 20), based on the simulated data sets under each simulation setting of n, θ, and λ. Denote by $\hat{F}_b$ the MLE (with estimated α_2 and either known or estimated α_1) and by $\tilde{F}_b$ the naive estimator for F based on the bth simulated data set. Then, the bias and MSE of $\hat{F}$ and $\tilde{F}$ at time t_0 are estimated by

$$\mathrm{bias}\left(\hat{F}\left(t_0\right)\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{F}_b\left(t_0\right) - F\left(t_0\right)\right),\ \mathrm{MSE}\left(\hat{F}\left(t_0\right)\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{F}_b\left(t_0\right) - F\left(t_0\right)\right)^2,\ \text{and}$$

$$\mathrm{bias}\left(\tilde{F}\left(t_0\right)\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\tilde{F}_b\left(t_0\right) - F\left(t_0\right)\right),\ \mathrm{MSE}\left(\tilde{F}\left(t_0\right)\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\tilde{F}_b\left(t_0\right) - F\left(t_0\right)\right)^2.$$

The estimated bias of $\hat{F}$ is smaller than the estimated bias of $\tilde{F}$ for setting (i); that is, the specificity is estimated with $\hat{F}$ and the sensitivity is fixed at the simulation design value (Figure S.1A–C in Supporting Information). The absolute value of the estimated bias of $\hat{F}$ is largest when the sensitivity is 0.5, the specificity is 0.9, and the sample size is 100. The bias of $\hat{F}$ tends to zero as n increases for all settings of $\{t_0, \alpha_1, \alpha_2, \theta, \lambda\}$, whereas the bias of $\tilde{F}$ does not converge to zero and is large for $\alpha_1 = 0.5$. Similar results are observed in setting (ii), when both the sensitivity and the specificity are estimated with $\hat{F}$ (Figure S.1D–F). In line with the results on the estimated bias, the MSE of $\hat{F}$ is in general much smaller than the MSE of $\tilde{F}$ (Figure S.2). This is most pronounced for a short time to event ($\lambda = 2/\theta$, black lines). Only for small samples ($n = 100$) and long time to event ($\lambda = 100/\theta$, blue lines), the MSE of $\hat{F}$ was larger than MSE of $\tilde{F}$.

4.2. *Estimation of the sensitivity and specificity*

For setting (i), the estimates for the specificity seem to be consistent (Figure S.3A–F), even in case of small sample size ($n=100$). The bias is largest when $\alpha_2=0.5$, $\alpha_1=0.5$, $n=100$, and cumulative incidence is high ($\lambda=2/\theta$). For these settings, the interquartile range (IQR) of the estimated specificity is widest when $\theta=3/2$ (0.457 − 0.546). For $\alpha_2=0.9$ and $n=100$, the IQR (0.884–0.932) is widest when $\alpha_1=0.5$, $\lambda=2/\theta$, and $\theta=2/3$. For $n=1000$, the bias and IQRs are markedly smaller. Over all simulation settings, the median bias ranges from −0.0025 to 0.01. Similar results are found for setting (ii) (Figure S.3G–L).

The estimator for the sensitivity performs worse than the estimator for the specificity (Figure S.4). For $n=100$, the sensitivity tends to be overestimated and the bias is largest when only a few events are observed ($\lambda=100/\theta$). This may be explained by the fact that the data contain little information about the sensitivity, and consequently, the log-likelihood is a flat function of α_1. When the sample size increases ($n=1000$), the bias vanishes, but the IQR is still large when the cumulative incidence is low (that is, when $\lambda=100/\theta$). The IQR is smaller for low specificity than for high specificity. A likely explanation for this is that a false positive test, which is common when the specificity is low, is followed by a verification test that provides conclusive information about the event status.

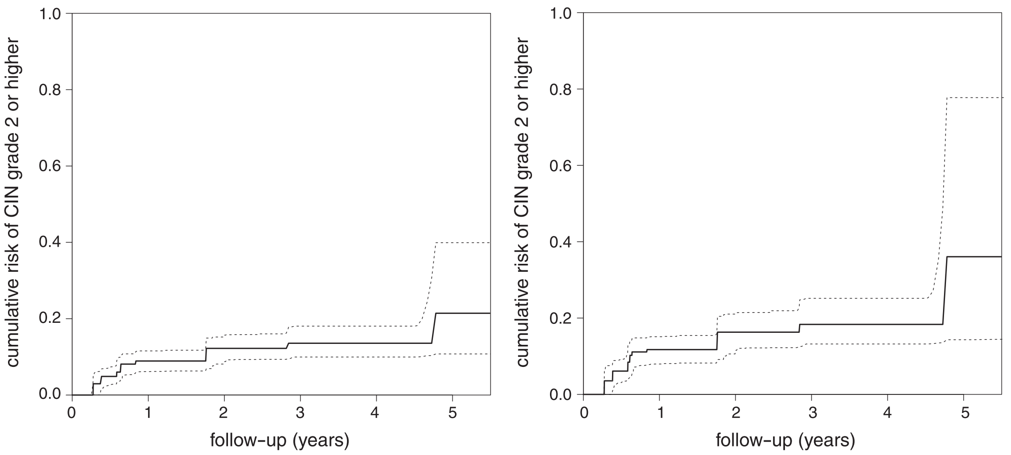
4.3. *Performance of the EM algorithm*

To examine the convergence of the EM algorithm, we simulate one data set for each setting of θ, λ, n, α_1, and α_2. We then simulate 1000 starting vectors $\mathbf{p}^{(0)}$ from the uniform[0, 1] distribution yielding estimates $\hat{F}_b$ ($b = 1, \ldots, 1000$) for each data set separately. Sensitivity and specificity are fixed instead of estimated. For each setting of the simulation design parameters, the maximal difference between the estimates

$$\max_{1\leqslant u\leqslant U+1}\left\{\max_b \hat{F}_b(\tau_u) - \min_b \hat{F}_b(\tau_u)\right\}$$

is computed. The convergence results are encouraging: for 52 out of 54 simulation parameter settings, the maximal difference is smaller than $5\cdot 10^{-4}$. For two settings, ($\theta=3/2$, $\lambda=100/\theta$, $n=100$, $\alpha_1=0.9$, and α_2 either 0.5 or 0.9), the maximal differences are 0.016 ($\alpha_2=0.5$) and 0.015 ($\alpha_2=0.9$).

We investigate the speed of the algorithm by increasing the number of possible support points $\tau_1, \tau_2, \ldots, \tau_U$ of $\hat{F}$. First, the censoring times are randomly selected from the set $\{1, 2, \ldots, 24, 25\}$. Then, the grid resolution is increased to 0.1 and finally to 0.01. The maximum numbers of follow-up times to which positive mass can be assigned, denoted by $U_{\max}$, are then equal to 25, 250, and 2500, respectively. We set the parameters θ, λ, n, α_1, and α_2 as before and treated α_1 and α_2 as fixed. For each combination of simulation parameters, we draw 1000 data sets and record the CPU time (in seconds) needed to compute $\hat{F}$ on a Dell precision workstation (Windows XP, 48 GB RAM, 12 core 2.93 GHz, intel X5670 xeon processor). The naive estimator is again used as starting estimator of the EM algorithm. For $U_{\max}$ equal to



A. $\tilde{F}$

B. $\hat{F}$ with $\alpha_1 = 0.53, \alpha_2 = 0.829$

Figure 1. The estimators $\tilde{F}$ (A) and $\hat{F}$ (B) with 95% bootstrap confidence interval (dashed lines) for the cumulative risk of recurrent CIN2+ after treatment.

25, the median CPU time ranges from 0.10 to 0.88 s; for $U_{\max}$ equal to 250, the median CPU time ranges from 7.6 to 24.1 s; and for $U_{\max}$ equal to 2500, the median CPU time ranges from 110.0 to 487.7 s.

5. Application

In this section, we apply the EM algorithm described in Section 3 to the data of Kocken *et al.* [1]. In this long-term multi-cohort study, 435 women treated for CIN2/3 were monitored for development of recurrent cervical precancer or cancer (CIN2+). Cytology tests were used for surveillance. In case of an abnormal cytology test, specimens were obtained by colposcopy and a histological confirmatory diagnosis was carried out. If histology was negative, that is, no CIN2+, women remained under surveillance. Ten women were excluded from our analysis because follow-up information was not available or study protocol violation. Of the remaining 425 women, 52 women (12.2%) developed recurrent CIN2+ after a median follow-up of 299 days (range 101–1746 days). In total, 735 unique follow-up visits are observed, ranging from 43 to 2328 days.

We estimate both the cumulative incidence of recurrent CIN2+ and the specificity simultaneously and keep the sensitivity fixed, because only few women developed recurrent CIN2+ and our simulation study shows that the estimator of the sensitivity behaves poorly in such a setting. The sensitivity of cytology is set to 53.0%, the pooled estimate in an international meta-analysis of screening cohorts [12]. Figure 1 shows the estimates $\tilde{F}$ (left panel) and $\hat{F}$ (right panel) together with 95% pointwise confidence bounds obtained via nonparametric bootstrap. The number of bootstrap samples is set at 2000. The CPU time needed to compute $\hat{F}$ was 6.1 s. Of note, both $\hat{F}$ and $\tilde{F}$ are not unique between two successive observed time points, and therefore, we chose to linearly interpolate the values of $\hat{F}$ and $\tilde{F}$ between time points.

The estimated specificity is 82.9% (95% confidence interval: 80.1–85.4%). This is lower than the pooled specificity of 96.3% reported by Cuzick *et al.* [12], where all studies except one (84.2%) showed a specificity above 92.9%. Note however that our population is a high-risk population of women treated for CIN2/3, and it is well known that the specificity in a clinical cohort can differ from the specificity in a screening population.

The corrected risk estimate $\hat{F}$ is always higher than the uncorrected risk estimate $\tilde{F}$ (Figure 1). After 4.5 years of follow-up, the corrected estimate shows a stronger increase than the uncorrected estimate, and also the pointwise confidence interval is markedly wider. This difference might be explained by the fact that in only a few women CIN2+ recurred by the end of follow-up. As the cytology test is imperfect, the event might have been undetected for some of these women. Therefore, the MLE that accounts for the imperfection of the screening test puts more mass at the last observation point than the uncorrected MLE does.

Because we did not directly estimate the sensitivity of cytology but imputed an estimate from a meta-analysis, we performed a sensitivity analysis. Figure 2 displays the 1-, 2-, and 4-year risks of recurrent CIN2+ for different values of the sensitivity of cytology, varying from 30% to 100% because a sensitivity

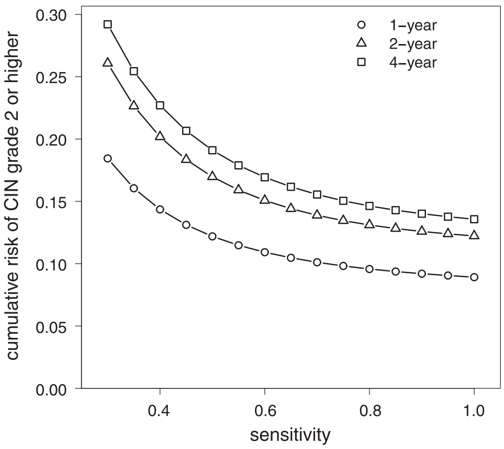


Figure 2. The 1-, 2-, and 4-year risk of recurrence of CIN2+ for different values of the sensitivity of the cytology test. The specificity was estimated by 82.9% based on the data.

below 30% is deemed very unlikely, while the specificity is fixed at our maximum likelihood estimate of 82.9%. The estimated 1-, 2-, and 4-year cumulative incidence of recurrent CIN2+ decreases when the sensitivity increases, but a strong impact of sensitivity is observed only if the sensitivity is below 50%.

6. Discussion

In this article, we propose an EM algorithm to compute the MLE that accounts for the imperfection of the screening test in the mixed case interval censoring model with continuous screening times. The combined E and M step of the algorithm has a closed form (6) and does not involve high-dimensional matrix inversion like the SQP methods described in Balasubramanian and Lagakos [3]. Convergence properties are good, and our algorithm is fast even when the maximum number of distinct time points is 2500 and the number of subjects is 1000. Our estimator shows minimal bias in the simulation study, whereas the naive estimator may be severely biased. Furthermore, our simulation study shows that the specificity of the screening test can be consistently estimated on the basis of the data, even if the sample size is small. Estimating the sensitivity works well in large ($n = 1000$) samples but appears to be inaccurate in small samples in particular in combination with a low cumulative incidence. The main reason for this is that a confirmatory test is only performed in subjects with a positive screening test so that the data provide ample information on the specificity but only limited information on the sensitivity. In general, however, good estimates of the sensitivity may be available from external sources. When this is not the case, one may consider estimating the sensitivity together with the distribution function, but this strategy can only be successful when both the study size and the number of observed events are high.

Our method assumes that the confirmatory test is perfect. In cancer screening and surveillance, histopathological results are used for disease confirmation and are assumed to be very accurate. If the follow-up confirmatory test had been imperfect, estimation of its sensitivity would have been challenging because treatment is recommended after a positive result. In some clinical settings, a confirmatory test is not available. Then multiple positive screening tests may be observed before treatment, which improves the accuracy of the estimate of the screening test sensitivity considerably. This setting has been studied by Titman [5].

We think that our method is particularly useful when the sensitivity of the screening test is low. In our example on recurrent CIN2+, the corrected and uncorrected cumulative risks for the first 4.5 years after treatment differ, but a sensitivity of cytology below 30% is deemed very unlikely. For other screening tests, for example, fecal occult blood testing and prostate-specific antigen in colorectal and prostate cancer screening, respectively, sensitivities of 20% have been reported [13, 14]. With our EM algorithm, the cumulative incidence of the precancerous disease can now be estimated more accurately in these screening settings.

Appendix

Derivation of the log-likelihood (1)

Define for any $r > 0$ $\mathbf{X}^r = \left(X_1, X_2, \dots, X_r\right)$ the vector containing the first r outcomes of $\mathbf{X}$ and similar for $\mathbf{Z}^r$, $\mathbf{C}^r$, $\mathbf{x}^r$, $\mathbf{z}^r$, and $\mathbf{c}^r$. Then, straightforward calculations imply that

$$\begin{aligned} &h_{F,\alpha_1,\alpha_2}(r, \mathbf{c}, \mathbf{x}, \mathbf{z}, e = 1) = \\ &\quad = \mathrm{P}\left(\mathbf{X}^r = \mathbf{x}^r, \mathbf{Z}^{r-1} = 0, Z_r = 1 \mid K \geq r, \mathbf{C}^r = \mathbf{c}^r\right) g_r(\mathbf{c}^r \mid K \geq r)\mathrm{P}(K \geq r), \end{aligned} \tag{A1}$$

$$\begin{aligned} &h_{F,\alpha_1,\alpha_2}(r, \mathbf{c}, \mathbf{x}, \mathbf{z}, e = 0) = \\ &\quad = \mathrm{P}(\mathbf{X}^r = \mathbf{x}^r, \mathbf{Z}^r = 0 \mid K = r, \mathbf{C}^r = \mathbf{c}^r) g_r(\mathbf{c}^r \mid K = r)\mathrm{P}(K = r), \end{aligned} \tag{A2}$$

where g_r is the conditional density of $\mathbf{C}^r$ (which, for simplicity of notation, does not distinguish between both conditional events in (A1) and (A2)). By assumption (A.1), T is also independent of $\mathbf{C}^r$ and K; hence, $g_r(\mathbf{c}^r \mid K \geq r)\mathrm{P}(K \geq r)$ and $g_r(\mathbf{c}^r \mid K = r)\mathrm{P}(K = r)$ do not depend on F. Because they also do not depend on α_1 and α_2, we only have to determine the first probabilities in (A1) and (A2) to derive (2).

Recall that $l = \max\{j : x_j = 1, z_j = 0\}$ is the (observed) index of the last confirmed false positive screening result and $l^+ = \sum_{j=1}^{l} x_j$ is the total number of observed false positive screening tests up to and including time c_l (with $l = l^+ = 0$ in case of no false positive screening tests). Under assumptions (A.2–A.4),

$$\begin{aligned} &\mathrm{P}(\mathbf{X}^r = \mathbf{x}^r, \mathbf{Z}^{r-1} = 0, Z_r = 1 \mid K \geqslant r, \mathbf{C}^r = \mathbf{c}^r) = \\ &= \sum_{j=1}^{r} \mathrm{P}\left(\mathbf{X}^r = \mathbf{x}^r, \mathbf{Z}^{r-1} = 0, Z_r = 1 \mid c_{j-1} < T \leqslant c_j\right) \mathrm{P}\left(c_{j-1} < T \leqslant c_j\right) \\ &= \sum_{j=l+1}^{r} \mathrm{P}\left(\mathbf{X}^r = \mathbf{x}^r, \mathbf{Z}^{r-1} = 0, Z_r = 1 \mid c_{j-1} < T \leqslant c_j\right) \left(F\left(c_j\right) - F\left(c_{j-1}\right)\right) \\ &= \mathrm{P}\left(\mathbf{X}^l = \mathbf{x}^l, \mathbf{Y}^l = 0 \mid T > c_l\right) \times \\ &\quad \times \sum_{j=l+1}^{r} \left\{ \mathrm{P}\left(X_{l+1} = \dots = X_{r-1} = 0, X_r = 1, Y_r = 1 \mid c_{j-1} < T \leq c_j\right) \times \right. \\ &\quad \left. \times \left(F\left(c_j\right) - F\left(c_{j-1}\right)\right) \right\} \\ &= \left(1 - \alpha_2\right)^{l^+} \alpha_2{}^{l-l^+} \sum_{j=l+1}^{r} \alpha_2{}^{j-l-1} \left(1 - \alpha_1\right)^{r-j} \alpha_1 \left(F\left(c_j\right) - F\left(c_{j-1}\right)\right). \end{aligned}$$

The second equality holds because for all $j \leqslant l$, the probability $\mathrm{P}(\mathbf{X} = \mathbf{x}, \mathbf{Z}^{r-1} = 0, Z_r = 1 \mid c_{j-1} < T \leq c_j)$ is equal to zero by assumption (A.4). Similarly,

$$\begin{aligned} &\mathrm{P}\left(\mathbf{X}^r = \mathbf{x}^r, \mathbf{Z}^r = 0 \mid K = r, \mathbf{C}^r = \mathbf{c}^r\right) = \\ &\quad = \left(1 - \alpha_2\right)^{l^+} \alpha_2{}^{l-l^+} \left\{ \sum_{j=l+1}^{r} \alpha_2{}^{j-l-1} (1 - \alpha_1)^{r-j+1} \left(F\left(c_j\right) - F\left(c_{j-1}\right)\right) + \alpha_2{}^{r-l} \left(1 - F\left(c_r\right)\right) \right\}. \end{aligned}$$

Hence, (2) now follows by substituting these probabilities in (A1) and (A2). □

Proof of equation (4)

First, note that in case $E_i = 1$

$$\begin{aligned} &\mathrm{P}^{(m)}\left(T_i = \tau_u \mid W_i\right) = \\ &\quad = \sum_{j=L_i+1}^{R_i} \mathrm{P}^{(m)}\left(T_i = \tau_u \mid C_{ij-1} < T_i \leqslant C_{ij}, W_i\right) \mathrm{P}^{(m)}\left(C_{ij-1} < T_i \leqslant C_{ij} \mid W_i\right), \end{aligned}$$

because, by assumption (A.4), all probabilities $\mathrm{P}^{(m)}\left(C_{ij-1} < T_i \leqslant C_{ij} \mid W_i\right)$ are equal to zero for all $j \leqslant L_i$. For fixed $j \in \{L_i + 1, \dots, R_i\}$, it then holds that

$$\begin{aligned}\mathrm{P}^{(m)}\left(T_i = \tau_u \mid C_{ij-1} < T_i \leqslant C_{ij}, W_i\right) &= \\ = \mathrm{P}^{(m)}\left(T_i = \tau_u \mid C_{ij-1} < T_i \leqslant C_{ij}\right) &= \frac{p_u^{(m)} 1_{(C_{ij-1}, C_{ij}]}(\tau_u)}{F^{(m)}\left(C_{ij}\right) - F^{(m)}\left(C_{ij-1}\right)}.\end{aligned} \tag{A3}$$

Furthermore,

$$\begin{aligned}&\mathrm{P}^{(m)}\left(C_{ij-1} < T_i \leqslant C_{ij} \mid W_i\right) = \\ &= \frac{\mathrm{P}^{(m)}\left(W_i \mid C_{ij-1} < T_i \leqslant C_{ij}\right) \mathrm{P}^{(m)}\left(C_{ij-1} < T_i \leqslant C_{ij}\right)}{\sum_{j=L_i+1}^{R_i} \mathrm{P}^{(m)}\left(W_i \mid C_{ij-1} < T_i \leqslant C_{ij}\right) \mathrm{P}^{(m)}\left(C_{ij-1} < T_i \leqslant C_{ij}\right)} \\ &= \frac{\left(1-\alpha_2\right)^{L_i^+} \alpha_2^{L_i - L_i^+} \alpha_2^{j-L_i-1} \left(1-\alpha_1\right)^{R_i - j} \alpha_1 \left(F^{(m)}\left(C_{ij}\right) - F^{(m)}\left(C_{ij-1}\right)\right)}{\left(1-\alpha_2\right)^{L_i^+} \alpha_2^{L_i - L_i^+} \sum_{j=L_i+1}^{R_i} \alpha_2^{j-L_i-1} \left(1-\alpha_1\right)^{R_i - j} \alpha_1 \left(F^{(m)}\left(C_{ij}\right) - F^{(m)}\left(C_{ij-1}\right)\right)}.\end{aligned} \tag{A4}$$

The terms $\mathrm{P}^{(m)}\left(W_i \mid T_i \leqslant C_{iL_i}\right) \mathrm{P}^{(m)}\left(T_i \leqslant C_{iL_i}\right)$ and $\mathrm{P}^{(m)}\left(W_i \mid T_i > C_{iR_i}\right) \mathrm{P}^{(m)}\left(T_i > C_{iR_i}\right)$ are omitted in the denominator of the first equality, because they are zero by assumption (A.4). Combining (A3) and (A4) yields (4). □

Acknowledgement

The authors thank Professor C. J. L. M. Meijer for providing the follow-up data used in Section 5.

References

1. Kocken M, Helmerhorst TJM, Berkhof J, Louwers JA, Nobbenhuis MAE, Bais AG, Hogewoning CJA, Zaal A, Verheijen RHM, Snijders PJF, Meijer CJLM. Risk of recurrent high-grade cervical intraepithelial neoplasia after successful treatment: a long-term multi-cohort study. *Lancet Oncology* 2011; **12**:441–450.
2. Richardson BA, Hughes JP. Product limit estimation for infectious disease data when the diagnostic test for the outcome is measured with uncertainty. *Biostatistics* 2000; **1**:341–354.
3. Balasubramanian R, Lagakos SW. Estimation of a failure time distribution based on imperfect diagnostic tests. *Biometrika* 2003; **90**:171–182.
4. Fletcher R. *Practical Methods of Optimization*. John Wiley & Sons: Chisester, 1987. 2nd edition.
5. Titman AC. Non-parametric maximum likelihood estimation of interval-censored failure time data subject to misclassification. *Statistics and Computing* 2016:1–9.
6. Jongbloed G. The iterative convex minorant algorithm for nonparametric estimation. *Journal of Computational and Graphical Statistics* 1998; **7**:310–321.
7. Wang Y. On fast computation of the non-parametric maximum likelihood estimate of a mixing distribution. *Journal of the Royal Statistical Society: Series B* 2007; **69**:185–298.
8. Dempster AP, Laird NM, Rubin DB. Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society (Series B)* 1977; **39**:1–38.
9. Efron B. Bootstrap methods: another look at the jackknife. *Annals of Statistics* 1979; **7**:1–26.
10. Groeneboom P, Wellner JA. *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhäuser Verlag: Basel, 1992.
11. Huang J, Wellner JA. Interval censored survival data: a review of recent progress. In *Proceedings of the First Seattle Symposium in Biostatistics*, Lin DY, Fleming TR (eds), Lecture Notes in Statistics, vol. 123. Springer-Verlag: Berlin, 1997; 123–169.
12. Cuzick J, Clavel C, Petry KU, Meijer CJLM, Hoyer H, Ratnam S, Szarewski A, Birembaut P, Kulasingam S, Sasieni P, Iftner T. Overview of the European and North American studies on HPV testing in primary cervical cancer screening. *International Journal of Cancer* 2006; **119**:1095–1101.
13. Hol L, van Leerdam ME, van Ballegooijen M, van Vuuren AJ, van Dekken H, Reijerink JCIY, van der Togt ACM, Habbema JDF, Kuipers EJ. Screening for colorectal cancer: randomised trial comparing guaiac-based and immunochemical faecal occult blood testing and flexible sigmoidoscopy. *Gut* 2010; **59**:62–68.
14. Thompson IM, Ankerst DP, Chi C, Lucia MS, Goodman PJ, Crowley JJ, Parnes HL, Coltman CA. Operating characteristics of prostate-specific antigen in men with an initial PSA level of 3.0 ng/ml or lower. *Journal of the American Medical Association* 2005; **294**:66–70.

Supporting information

Additional supporting information may be found online in the supporting information tab for this article.