

Himmelfarb Health Sciences Library, The George Washington University Health Sciences Research Commons

Epidemiology and Biostatistics Faculty Publications

Epidemiology and Biostatistics

6-28-2017

Mixture models for undiagnosed prevalent disease and interval-censored incident disease: applications to a cohort assembled from electronic health records.

Li C Cheung

Qing Pan

George Washington University


Noorie Hyun

Mark Schiffman

Barbara Fetterman

See next page for additional authors

Follow this and additional works at: https://hsrc.himmelfarb.gwu.edu/sphhs_epibiostats_facpubs

 Part of the [Biostatistics Commons](#), [Epidemiology Commons](#), and the [Statistical Models Commons](#)

APA Citation

Cheung, L., Pan, Q., Hyun, N., Schiffman, M., Fetterman, B., Castle, P., Lorey, T., & Katki, H. (2017). Mixture models for undiagnosed prevalent disease and interval-censored incident disease: applications to a cohort assembled from electronic health records.. *Statistics in Medicine*, (). <http://dx.doi.org/10.1002/sim.7380>

This Journal Article is brought to you for free and open access by the Epidemiology and Biostatistics at Health Sciences Research Commons. It has been accepted for inclusion in Epidemiology and Biostatistics Faculty Publications by an authorized administrator of Health Sciences Research Commons. For more information, please contact hsrc@gwu.edu.

Authors

Li C Cheung, Qing Pan, Noorie Hyun, Mark Schiffman, Barbara Fetterman, Philip E Castle, Thomas Lorey,
and Hormuzd A Katki

Mixture models for undiagnosed prevalent disease and interval-censored incident disease: applications to a cohort assembled from electronic health records

Li C. Cheung,^{a,b,*†} Qing Pan,^a Noorie Hyun,^b Mark Schiffman,^b Barbara Fetterman,^c Philip E. Castle,^d Thomas Lorey^c and Hormuzd A. Katki^b

For cost-effectiveness and efficiency, many large-scale general-purpose cohort studies are being assembled within large health-care providers who use electronic health records. Two key features of such data are that incident disease is interval-censored between irregular visits and there can be pre-existing (prevalent) disease. Because prevalent disease is not always immediately diagnosed, some disease diagnosed at later visits are actually undiagnosed prevalent disease. We consider prevalent disease as a point mass at time zero for clinical applications where there is no interest in time of prevalent disease onset. We demonstrate that the naive Kaplan–Meier cumulative risk estimator underestimates risks at early time points and overestimates later risks. We propose a general family of mixture models for undiagnosed prevalent disease and interval-censored incident disease that we call prevalence–incidence models. Parameters for parametric prevalence–incidence models, such as the logistic regression and Weibull survival (logistic–Weibull) model, are estimated by direct likelihood maximization or by EM algorithm. Non-parametric methods are proposed to calculate cumulative risks for cases without covariates. We compare naive Kaplan–Meier, logistic–Weibull, and non-parametric estimates of cumulative risk in the cervical cancer screening program at Kaiser Permanente Northern California. Kaplan–Meier provided poor estimates while the logistic–Weibull model was a close fit to the non-parametric. Our findings support our use of logistic–Weibull models to develop the risk estimates that underlie current US risk-based cervical cancer screening guidelines. Published 2017. This article has been contributed to by US Government employees and their work is in the public domain in the USA.

Keywords: cervical cancer; cumulative risk estimation; HPV; Kaplan–Meier; prevalence–incidence models

1. Introduction

Screening is used to identify individuals with asymptomatic disease or disease precursors for effective early intervention. A goal of ‘precision medicine’ is to develop screening guidelines that are based on the risk of disease, given an individual’s risk factors and screening test results [1]. To inform screening guidelines, many large-scale general-purpose epidemiologic cohort studies are being organized within large healthcare providers who use electronic health records [2]. These providers have the benefit of having large patients populations to recruit from, pre-existing infrastructure to support longitudinal visits, and electronic health records to facilitate data collection. Although potentially cost-effective and efficient, such cohorts present many challenges for analysis [3]. We consider two key features of such data that make it inappropriate to calculate risk using standard methods, such as Kaplan–Meier methods [4] or Cox models [5].

^aDepartment of Statistics, The George Washington University, Washington, DC, U.S.A.

^bDivision of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Rockville, MD, U.S.A.

^cRegional Laboratory, Kaiser Permanente Northern California, Berkeley, CA, U.S.A.

^dAlbert Einstein College of Medicine, Bronx, NY, U.S.A.

*Correspondence to: Li C. Cheung, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, 9609 Medical Center Drive, Room 7E612, Rockville, MD 20850, U.S.A.

†E-mail: li.cheung@nih.gov.

First, the exact time of disease onset for an individual is unobserved, and only the partial information of the occurrence falling between the last time observed to be disease-free and the time of diagnosis, that is, mixed-cases interval censoring [6], is available. For approximately regular intervals, which can be enforced in controlled trials, discrete time or Kaplan–Meier methods could be applied as approximations. However, researchers working with data from health providers typically cannot influence the timing of visits and patients return at intervals that are quite irregular, which render discrete-time or Kaplan–Meier methods inappropriate to use [7].

Second, people can enter the screening program with prevalent disease at baseline that is not always immediately diagnosed. In particular, people with missing or negative screening test results generally do not undergo definitive disease ascertainment, such as biopsies. Consequently, among those for whom prevalent disease was not ascertained but disease is diagnosed at future screens, it is not possible to determine whether the disease is undiagnosed prevalent disease that occurred prior to enrollment into the screening program or incident disease that occurred over the course of screening. Accurately estimating risk of prevalent disease is important because clinicians are primarily concerned with the risk that disease is present. Guidelines for referring women for immediate biopsies should be based on the prevalent disease risk [8].

In our application, it suffices to consider prevalent disease as a left-censored point mass at time zero, because a clinician takes little interest on the time in the past when disease detected today might have arisen. Standard approaches to mixed cases interval censoring [9–11] do not handle a point mass at time zero for prevalent disease. One could either set aside those with ascertained prevalent disease or approximate ascertained prevalent disease by offsetting to a small time just after zero and then apply traditional interval-censoring approaches. However, both of these methods would treat unascertained prevalent disease that are diagnosed at future screens as incident disease and would result in underestimating prevalent disease risks while inflating incident disease risks. For covariates, a Cox model cannot properly separate covariate effects on prevalent disease from those for incident disease. Furthermore, applying Kaplan–Meier and Cox model methods by ignoring the interval censoring and using time of diagnosis to approximate time of disease onset results in biased estimates [12, 13]. In Section 2, we demonstrate that there is an interesting pattern to this bias: The Kaplan–Meier cumulative risk estimator generally underestimates early risks and overestimates later risks.

In this article, we propose a general family of mixture models, which we call prevalence–incidence survival models, for estimating the cumulative risk and assessing covariate effects using data from screening cohorts assembled from electronic health records (Section 3). Prevalence–incidence models are akin to cure mixture models [14–16], but the point mass is at zero rather than infinity. Prevalent disease is observed for some individuals, mitigating identifiability issues that can affect cure mixture models [17]. Identifiability of prevalence–incidence models is considered in the Appendix S1. Moreover, in prevalence–incidence survival models, accurately identifying those at high risk of prevalent disease and in need of immediate intervention is of primary importance, whereas, in cure models, it is not often crucial for clinicians to differentiate long-term survivors from those who are cured.

We employed the expectation-maximization (EM) algorithm [18] for inference in parametric prevalence–incidence survival models. All details are presented for the useful special case where prevalent disease is modelled with logistic regression and incident disease is modelled with a Weibull survival model, which is supported by the multi-stage theory of carcinogenesis [19]. We also present a non-parametric cumulative risk estimator for the marginal model that has no covariates; this can be used to check the distribution assumptions of the parametric prevalence–incidence survival models. Our R package, PIMixture (<https://dceg.cancer.gov/tools/analysis/PIMixture>), provides the non-parametric estimator, the logistic–Weibull model, and various other parametric prevalence–incidence models. We use simulations to examine the robustness of our cumulative risk estimators when the logistic–Weibull model is misspecified (Section 4).

In section 5, we use the Kaiser Permanente Northern California (KPNC) cervical cancer screening cohort to estimate the cumulative risk of cervical precancer and cancer following HPV-positive/Pap-negative baseline screening results. The KPNC cohort consists of 1.4 million women screened from 2003–2013 and was assembled by linking various electronic records of patient information, test results, and disease outcomes [20]. Not all women with HPV-positive/Pap-negative results underwent definitive disease ascertainment at baseline, so while only 0.18% of these women had precancer/cancer diagnosed at baseline, precancer/cancer could have been present at baseline for another 2.05% of these women. Compared with the non-parametric estimates, the Kaplan–Meier estimates show the expected under/over estimation. Estimates from the logistic–Weibull model agree with the non-parametric estimates and

demonstrates that the logistic–Weibull model can estimate prevalent disease risk even when disease is rarely ascertained at baseline. We illustrate how we used the risk estimates from the logistic–Weibull model to inform the current U.S. cervical cancer screening guidelines [21, 22]. In Section 6, some useful extensions of prevalence–incidence models are discussed.

2. The behavior of the Kaplan–Meier estimator in screening data

While originally proposed for analysis of mortality data, Kaplan–Meier methods are often used to estimate disease-free survival by using the time of diagnosis to approximate time of disease onset [13]. This approximation is reasonable for symptomatic disease, as the diagnosis time is close to the time of disease onset. However, for asymptomatic disease, the difference between time of diagnosis and time of onset depends on the density of screening visits and the sensitivity of screening tests. In cervical screening, women who test negative usually have long-term (3-year) intervals [21]. Consequently, diagnosis of cervical precancer/cancer may lag onset by years, and Kaplan–Meier estimates based on diagnosis time may be biased for the cumulative risk of disease onset.

Consider this simple example. For n subjects disease-free at baseline, denote the cumulative distribution of time of disease onset as F . Suppose these subjects randomly return for a single follow-up visit with definitive disease ascertainment at one of two time points: θn subjects are seen at t_1 , and the remaining $(1 - \theta)n$ subjects are seen at time t_2 , where $\theta \in (0, 1)$ and $0 < t_1 < t_2$. From Table I, we see that Kaplan–Meier methods using time of diagnosis will underestimate the cumulative risk at t_1 by the fraction of people, θ , who return at the early time point. Intuitively, disease with early onset is not detected until later, and thus, Kaplan–Meier methods underestimates early disease risk. This underestimation can be substantial; if for example, people are asked to return at t_1 and 70% do return, then the cumulative risk at t_1 is underestimated by 30%.

Interestingly, at the later time point, t_2 , Kaplan–Meier methods treating diagnosis time as occurrence time overestimates the cumulative risk when $\theta F(t_1) > 0$. From Table I, it is easily seen that the amount of overestimation is given as $\theta F(t_1)\{1 + F(t_2)\}$. Intuitively, we not only have the extra disease from those whose disease onset was earlier but also the risk set contains fewer people than at the earlier time, leading to overestimation. The amount of overestimation at t_2 increases as follows

- 1 θ increases to one. When θ is close to zero, bias is small at time t_2 as almost all subjects return for the first time at time t_2 . As θ increases to one, the bias in the cumulative risk at t_1 shrinks while the bias at t_2 increases.
- 2 $F(t_1)$ increases to one. When $F(t_1)$ is zero, no events have occurred among the n subjects by t_1 . Thus, for the $(1 - \theta)n$ who return at time t_2 , time t_1 can be seen as renewal of the survival process with all subjects returning at time t_2 .

Note that t_1 can be arbitrarily close to t_2 , but the Kaplan–Meier estimate at t_1 and t_2 can differ considerably with biases in opposite directions. This is because the Kaplan–Meier estimator sees $(1 - \theta)nF(t_2)$ events occurring in the very short time period between t_1 and t_2 , with a smaller population at risk, $(1 - \theta)n$ instead of n , and thus estimates very large hazards for t_2 . A numerical example is presented in Appendix S2.

The bias in the Kaplan–Meier estimator does not depend on n but only on the proportions returning at each of the two time points, so increasing the sample size to infinity does not remove the bias. This pattern of the Kaplan–Meier cumulative risk estimator underestimating at earlier times and overestimating at later times persists for screening data with more than two return time points.

Table I. Performance of the Kaplan–Meier (KM) cumulative risk estimator using time of diagnosis to approximate time of disease onset.

Time	At risk	Diagnosed	KM cumulative risk estimate	True cumulative risk
t_1	n	$n\theta F(t_1)$	$\theta F(t_1)$	$F(t_1)$
t_2	$(1 - \theta)n$	$(1 - \theta)nF(t_2)$	$1 - (1 - \theta F(t_1))(1 - F(t_2))$	$F(t_2)$

Note: N subjects whose time of disease onset have cumulative distribution function, F , are randomly seen at one of two times: θn are seen at time t_1 and $(1 - \theta)n$ are seen at time t_2 .

3. Prevalence–incidence survival models

We propose a general family of mixture models, which we call prevalence–incidence survival models. For subjects $i = 1, 2, \dots, n$, let T_i be the time of disease onset, which is never directly observed. Let $c_i = I(T_i \leq 0)$, which is the subject-specific indicator variable for prevalent disease present at the baseline visit (time zero). Let L_i be the last time the i th subject is observed to be disease-free, which is defined only when c_i is known to be zero. Let R_i be the first time that the i th subject is diagnosed with the disease. For many subjects, c_i is unobserved, and we know only that $T_i \leq R_i$. We allow $R_i = \infty$ to include cases of right-censoring or when subjects with unobserved c_i provide no information regarding T_i . Let \mathbf{x}_i and \mathbf{z}_i be subject-specific vectors of covariates for prevalent and incident disease respectively.

Let θ_1 and θ_2 be disjoint vectors of parameters. Given covariates \mathbf{x} and \mathbf{z} , let $P(T \leq 0; \mathbf{x}, \mathbf{z}) = P(c = 1; \mathbf{x}) = \pi(\mathbf{x}; \theta_1)$ be the probability of prevalent baseline disease (prevalence model) and $P(T > t | c = 0, \mathbf{x}, \mathbf{z}) = P(T > t | c = 0, \mathbf{z}) = S(t; \theta_2 | c = 0, \mathbf{z})$ be the conditional survival function of time to disease onset given that the disease is not present at baseline (incidence model). The cumulative risk can be modelled as

$$P(T \leq t; \mathbf{x}, \mathbf{z}, \theta) = \pi(\mathbf{x}; \theta_1) + \{1 - \pi(\mathbf{x}; \theta_1)\} \{1 - S(t; \theta_2 | c = 0, \mathbf{z})\}, \quad t \geq 0, \quad (1)$$

where $\theta = \theta_1 \cup \theta_2$. Thus, the cumulative risk, $P(T \leq t; \mathbf{x}, \mathbf{z}, \theta)$, can be seen as a mixture distribution having mixing proportions $\pi(\mathbf{x}; \theta_1)$ and $1 - \pi(\mathbf{x}; \theta_1)$ with component distributions $P(T \leq t | c = 1) = 1$, a degenerate distribution among those with prevalent disease at baseline, and $P(T \leq t; \theta_2 | c = 0, \mathbf{z}) = 1 - S(t; \theta_2 | c = 0, \mathbf{z})$, the conditional failure time distribution among those without prevalent disease at baseline.

3.1. Non-parametric estimator of the cumulative risk

In this subsection, we show how to obtain the non-parametric maximum likelihood estimate (NPMLE) of the cumulative risk. In addition to its intrinsic value, the NPMLE can be used to assess the distribution assumptions of parametric prevalence–incidence models with no covariates.

The NPMLE of the cumulative risk can be obtained by applying the EM-iterative convex minorant (ICM) algorithm [23] after mapping (c_i, L_i, R_i) to new intervals (L'_i, R'_i) in three steps as followed:

- 1 if $c_i = 1$, set (L'_i, R'_i) to the interval $(0, \epsilon]$;
- 2 if $c_i = 0$, set (L'_i, R'_i) to intervals $(\max(\epsilon, L_i), R_i]$;
- 3 if c_i is unknown, set (L'_i, R'_i) to intervals $(0, R_i]$,

where ϵ is half of the value of the smallest positive L_i in the data set.

Because the NPMLE assumes that the survival function is one at time zero, we approximate prevalent disease by assuming known prevalent disease occurs in the small time interval $(0, \epsilon]$ immediately after time zero (step 1). The estimated risk of disease prevalent at time zero is then the NPMLE cumulative risk estimate at ϵ . Known incident disease is mapped to time intervals after ϵ (step 2). We set ϵ to half of the value of the smallest positive L_i in the data set; however, ϵ can take any value between zero and the smallest positive L_i . The intuition is to use ϵ to assign separate time intervals to known prevalent disease and known incident disease, as bias will ensue if the intervals overlap. The third step assigns time intervals that span the prevalent and incident disease intervals when it is unknown whether the disease is prevalent or incident.

We cannot directly extend methods for the NPMLE to include covariates because such an approach assumes that the size and relationship (e.g., proportional hazards or additive risks) of the covariate effects for incident disease also applies to prevalent disease. Instead, we suggest using prevalence–incidence survival models when incorporating covariates.

3.2. Likelihood for prevalence–incidence survival models

Let \mathbf{K}_1 and \mathbf{K}_0 be partitions of subjects into observed and missing c_i , respectively. We assume that c_i is missing at random (MAR).

When $c_i = 1$, the subject contributes $\log\{\pi(\mathbf{x}_i; \theta_1)\}$ to the log-likelihood. When $c_i = 0$, the subject contributes $\log\{1 - \pi(\mathbf{x}_i; \theta_1)\} + \log\{S(L_i; \theta_2 | c_i = 0, \mathbf{z}_i) - S(R_i; \theta_2 | c_i = 0, \mathbf{z}_i)\}$ to the log-likelihood. Finally, when c_i is unobserved, the subject contributes $\log[\pi(\mathbf{x}_i; \theta_1) + \{1 - \pi(\mathbf{x}_i; \theta_1)\} \{1 - S(R_i; \theta_2 | c_i = 0, \mathbf{z}_i)\}]$. Let \mathbf{y}^{obs} and \mathbf{y}^{mis} denotes the observed and missing data, respectively, contained in $\mathbf{y}_i = \{\mathbf{x}_i, \mathbf{z}_i, L_i, R_i, c_i\}$, for $i = 1, 2, \dots, n$. Then the observed log-likelihood is given as:

$$\begin{aligned}
 l(\mathbf{y}^{obs}; \boldsymbol{\theta}) = & \sum_{i \in \mathbf{K}_1} [c_i \log\{\pi(\mathbf{x}_i; \boldsymbol{\theta}_1)\} + (1 - c_i)\{\log(1 - \pi(\mathbf{x}_i; \boldsymbol{\theta}_1)) \\
 & + \log(S(L_i; \boldsymbol{\theta}_2|c_i = 0, \mathbf{z}_i) - S(R_i; \boldsymbol{\theta}_2|c_i = 0, \mathbf{z}_i))\}] \\
 & + \sum_{i \in \mathbf{K}_0} \log[\pi(\mathbf{x}_i; \boldsymbol{\theta}_1) + \{1 - \pi(\mathbf{x}_i; \boldsymbol{\theta}_1)\}\{1 - S(R_i; \boldsymbol{\theta}_2|c_i = 0, \mathbf{z}_i)\}].
 \end{aligned} \tag{2}$$

A full derivation of the observed log-likelihood from the complete data joint likelihood is given in Appendix S3.

3.3. Inference for parametric prevalence–incidence survival models

When parametric forms are chosen for both $\pi(\mathbf{x}; \boldsymbol{\theta}_1)$ and $S(t; \boldsymbol{\theta}_2|c = 0, \mathbf{z})$ in (1), the likelihood (2) can be directly maximized by Newton–Raphson method [24]. However, Newton–Raphson method does not converge for KPNC data analyses that have a large proportion of subjects (greater than 99%) with right-censored event times, which is a situation that is not uncommon for population-based cancer screening. For those analyses, we employ the EM algorithm [18] as followed:

Initialization Set initial values for the parameters $\boldsymbol{\theta}^{(0)}$. Two useful sets of initial values are the parameter values that maximize (2) after assuming that disease diagnosed in \mathbf{K}_0 are either all prevalent disease, ($c_i = 1$) or all incident disease ($c_i = 0$). When the EM algorithm based on these two extreme starting values converge to the same set of parameter values, that provides additional confidence that the EM algorithm is converging to a global maximum.

Alternate between the expectation (E) step and the maximization (M) step until convergence:

E-step Use $\boldsymbol{\theta}^{(l)}$ to compute the expected log-likelihood given by

$$\begin{aligned}
 Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(l)}) = & \sum_{i=1}^n [E(c_i|\mathbf{y}_i^{obs}; \boldsymbol{\theta}^{(l)}) \log\{\pi(\mathbf{x}_i; \boldsymbol{\theta}_1)\} \\
 & + \{1 - E(c_i|\mathbf{y}_i^{obs}; \boldsymbol{\theta}^{(l)})\}\{\log(1 - \pi(\mathbf{x}_i; \boldsymbol{\theta}_1))\} \\
 & + \log(S(L_i; \boldsymbol{\theta}_2|c_i = 0, \mathbf{z}_i) - S(R_i; \boldsymbol{\theta}_2|c_i = 0, \mathbf{z}_i))],
 \end{aligned} \tag{3}$$

where $S(L_i; \boldsymbol{\theta}_2|c_i = 0, \mathbf{z}_i) = 1$ when $i \in \mathbf{K}_0$ and

$$E(c_i|\mathbf{y}_i^{obs}; \boldsymbol{\theta}^{(l)}) = \begin{cases} c_i & : i \in \mathbf{K}_1 \\ \frac{\pi(\mathbf{x}_i; \boldsymbol{\theta}_1^{(l)})}{\pi(\mathbf{x}_i; \boldsymbol{\theta}_1^{(l)}) + \{1 - \pi(\mathbf{x}_i; \boldsymbol{\theta}_1^{(l)})\}\{1 - S(R_i; \boldsymbol{\theta}_2^{(l)}|c_i=0, \mathbf{z}_i)\}} & : i \in \mathbf{K}_0. \end{cases} \tag{4}$$

For subjects in \mathbf{K}_0 , as R_i approaches zero, $E(c_i|\mathbf{y}_i^{obs}; \boldsymbol{\theta}^{(l)})$ approaches one, reflecting the fact that the detected disease is more likely to have been prevalent at baseline. When $R_i = \infty$, then there is no information regarding the subject’s time of disease onset and the conditional expectation of c_i equals the average probability of prevalence.

M-step The updated MLE, $\boldsymbol{\theta}^{(l+1)}$, are the values of $\boldsymbol{\theta}$ that maximizes the expected log-likelihood (3), which can be separated into terms containing only $\pi(\mathbf{x}_i; \boldsymbol{\theta}_1)$ and only $S(t; \boldsymbol{\theta}_2|c_i = 0, \mathbf{z}_i)$ as followed:

$$\begin{aligned}
 Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(l)}) = & \sum_{i=1}^n [E(c_i|\mathbf{y}_i^{obs}; \boldsymbol{\theta}^{(l)}) \log\{\pi(\mathbf{x}_i; \boldsymbol{\theta}_1)\} + \{1 - E(c_i|\mathbf{y}_i^{obs}; \boldsymbol{\theta}^{(l)})\}\{\log(1 - \pi(\mathbf{x}_i; \boldsymbol{\theta}_1))\}] \\
 & + \sum_{i=1}^n \{1 - E(c_i|\mathbf{y}_i^{obs}; \boldsymbol{\theta}^{(l)})\}\{\log(S(L_i; \boldsymbol{\theta}_2|c_i = 0, \mathbf{z}_i) - S(R_i; \boldsymbol{\theta}_2|c_i = 0, \mathbf{z}_i))\}.
 \end{aligned}$$

This allows $\boldsymbol{\theta}_1^{(l+1)}$ and $\boldsymbol{\theta}_2^{(l+1)}$ to be estimated separately.

The variance of $\hat{\boldsymbol{\theta}}$ is the inverse of the observed Fisher information. The variance of the cumulative risk can be derived using the delta method:

$$\hat{Var}\{P(T \leq t; \mathbf{x}, \mathbf{z}, \hat{\boldsymbol{\theta}})\} = \nabla P(T \leq t; \mathbf{x}, \mathbf{z}, \hat{\boldsymbol{\theta}})^T \cdot Var(\hat{\boldsymbol{\theta}}) \cdot \nabla P(T \leq t; \mathbf{x}, \mathbf{z}, \hat{\boldsymbol{\theta}}),$$

where $\nabla P(T \leq t; \mathbf{x}, z, \hat{\theta})$ is the gradient of $P(T \leq t; \mathbf{x}, z, \hat{\theta})$.

The complementary log–log of the cumulative risk can be shown to be asymptotically normal via the multivariate delta method (Appendix S4). Confidence limits for the cumulative risk can be constructed on the complementary log–log scale and then converted to the cumulative risk scale.

3.4. logistic–Weibull model

A parametric prevalence–incidence survival model that is relevant to cancer screening is a mixture of logistic regression prevalence model and Weibull regression incidence model. The prevalence and incidence models are parameterized as

$$\pi(\mathbf{x}; \theta_1) = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})}$$

and

$$S(t; \theta_2 | c = 0, z_i) = \exp \left[- \left\{ \frac{t}{\exp(z_i^T \boldsymbol{\gamma})} \right\}^{\frac{1}{\tau}} \right]$$

respectively, where $\boldsymbol{\beta} = (\beta_0, \beta_2, \dots, \beta_{m_1-1})$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_2, \dots, \gamma_{m_2-1})$ are vectors of regression coefficients and $\tau > 0$ governs the shape of the Weibull distribution. Note that $\theta_1 = \boldsymbol{\beta}$ and $\theta_2 = (\boldsymbol{\gamma}, \tau)$. Inference for parameters, cumulative risk, odds ratios, and hazard ratios for the logistic–Weibull models are presented in Appendix S5. The M-step is solved using just one iteration of Newton’s method. This EM gradient algorithm [25] is locally equivalent to the EM algorithm but converges more quickly.

The Weibull survival model is supported by the multi-stage theory of carcinogenesis [19], which suggests that cancer incidence has a Weibull distribution. The time of follow-up for most cancer studies is relatively short, in the sense that time to cancer incidence is right-censored for most subjects. Given that we expect the hazard of cancer to be monotonic over this short time scale, the Weibull incidence model can be an adequate working model for cancer screening.

In the special case of no undiagnosed baseline disease (i.e., everyone at baseline undergoes definitive disease ascertainment), then the likelihood (2) can be factored into prevalence and incidence model parts, and thus, parameters in each model can be estimated separately. For the logistic–Weibull model, this means that the logistic regression and Weibull regression can be conducted separately to obtain their own parameter estimates.

4. Performance of the logistic–Weibull model

We evaluated the performance of the logistic–Weibull prevalence–incidence survival model for cases when the model was correctly specified and for cases when it was misspecified. Within each set-up, 1,000 simulation data sets, each containing 10,000 patients were used. Because most applications will use the same covariates for both the prevalence and incidence models, we set $x_i = z_i = (x_{1i}, x_{2i})$, where x_{1i} and x_{2i} were drawn from independent Bernoulli(0.5) and Uniform(0,1) distributions, respectively. Using different sets of covariates did not change the overall conclusions of the simulations.

For simulations under a correctly specified logistic–Weibull model, prevalent disease followed a logistic regression model and time to onset of incident disease followed a Weibull regression model:

$$\pi(\mathbf{x}; \theta_1) = \frac{\exp(-3.5 + .5x_{1i} + x_{2i})}{1 + \exp(-3.5 + .5x_{1i} + x_{2i})}$$

and

$$S(t; \theta_2 | c = 0, z) = \exp \left\{ - \left(\frac{t}{\exp(3.5 - .5x_{1i} - .5x_{2i})} \right)^2 \right\}.$$

For simulations under a misspecified logistic–Weibull model, prevalent disease followed a probit regression model and time to onset of incident disease followed a log–logistic survival model:

$$\pi(\mathbf{x}; \boldsymbol{\theta}_1) = \Phi(-2 + .35x_{1i} + .5x_{2i})$$

and

$$S(t; \boldsymbol{\theta}_2 | c = 0, z_i) = \frac{1}{1 + t^2 \exp\{-(4.3 + 1.5x_{1i} + 2x_{2i})\}},$$

where Φ is the cumulative distribution function of the standard normal distribution.

For each subject, an indicator variable for whether disease was ascertained at baseline was drawn from an independent Bernoulli distribution. We examined the performance of the prevalence–incidence model while varying the levels of disease ascertainment at baseline: 90% – high ascertainment, 50% – moderate ascertainment, and 10% – low ascertainment. Following baseline, time to each subsequent disease ascertainment followed a Gamma(3,1) distribution, with no disease ascertainment after time 20. The amount of right-censoring in the data varied from 63% to 69% for each of the simulation set-ups. Confidence intervals for the cumulative risks were obtained by constructing confidence limits on the asymptotically normal complementary log–log cumulative risk scale and converting those confidence limits to the cumulative risk scale.

When correctly specified, the logistic–Weibull prevalence–incidence model performed well regardless of the level of disease ascertainment at baseline, with unbiased parameter and cumulative risk estimates and good coverage. Table II gives the results of the logistic–Weibull model under low disease ascertainment at baseline.

Table III shows the result of fitting a misspecified logistic–Weibull prevalence–incidence survival model to a population where prevalent baseline disease followed a probit regression model and time to

Table II. Performance of the logistic–Weibull prevalence–incidence model when the model is not misspecified and with low disease ascertainment at baseline.

Parameter	True	Percentage bias	ASE	ESE	CP
β_0	−3.5	−0.25	0.13	0.13	0.952
β_1	0.5	0.86	0.10	0.10	0.943
β_2	1.0	0.26	0.18	0.18	0.952
γ_0	3.5	0.06	0.028	0.028	0.951
γ_1	−0.5	0.08	0.021	0.022	0.942
γ_2	−0.5	−0.11	0.034	0.034	0.960
τ	0.5	0.38	0.0082	0.010	0.956
$P(T = 0)$	0.076	−0.12	0.0049	0.0048	0.954
$P(T = 1)$	0.080	0.02	0.0048	0.0047	0.952
$P(T = 3)$	0.109	0.19	0.0045	0.0046	0.949
$P(T = 5)$	0.166	0.17	0.0050	0.0051	0.946
$P(T = 10)$	0.386	−0.08	0.0069	0.0069	0.949

Note: $\beta_0, \beta_1,$ and β_2 are coefficients for the logistic regression portion of the model; $\gamma_0, \gamma_1,$ and γ_2 are regression coefficients for the Weibull scale and τ is the inverse of the Weibull shape. $P(T = t)$ is the cumulative risk at time t for those with a specific set of covariates. Percentage bias is negative to denote underestimation and positive to denote overestimation. ASE, asymptotic standard error; ESE, empirical standard error; CP, coverage probability.

Table III. Performance of the misspecified logistic–Weibull prevalence–incidence model under high and low disease ascertainment at baseline.

Parameter	True	High ascertainment				Low ascertainment			
		Percentage bias	ASE	ESE	CP	Percentage bias	ASE	ESE	CP
$P(T = 0)$	0.081	−0.9	0.0039	0.0038	0.952	−8.3	0.0045	0.0044	0.726
$P(T = 1)$	0.082	0.7	0.0039	0.0038	0.949	−5.9	0.0045	0.0043	0.830
$P(T = 3)$	0.090	4.8	0.0039	0.0039	0.815	0.5	0.0043	0.0042	0.963
$P(T = 5)$	0.106	6.7	0.0041	0.0040	0.563	4.5	0.0044	0.0042	0.812
$P(T = 10)$	0.173	3.8	0.0051	0.0049	0.748	3.9	0.0053	0.0051	0.748

Note: $P(T = t)$ is the cumulative risk at time t for those with a specific set of covariates. Percentage bias is negative to denote underestimation and positive to denote overestimation. ASE, asymptotic standard error; ESE, empirical standard error; CP, coverage probability.

Table IV. Percent bias and coverage probability (CP) for the non-parametric cumulative risk estimate under different levels of disease ascertainment at baseline: high (H), moderate (M), and low (L).

Parameter	True	Percentage bias (H)	Percentage bias (M)	Percentage bias (L)
$P(T = 0)$	0.076	0.08	-0.03	-0.55
$P(T = 1)$	0.078	0.05	-0.11	-0.10
$P(T = 3)$	0.096	0.11	0.03	0.01
$P(T = 5)$	0.130	0.03	0.15	0.12
$P(T = 10)$	0.280	-0.06	-0.05	0.05

Note: Percentage bias is negative to denote underestimation and positive to denote overestimation.

onset of incident disease followed a log-logistic survival model. While coverage can be poor, the bias in the cumulative risk was below 9% at all time points. Under high baseline disease ascertainment, there was virtually no bias at early time points as logit and probit link functions often yield very similar outputs. The misspecified survival model resulted in biased cumulative risks at later times, with the greatest amount of overestimation at time 5.6 (6.76%). When there was less data on the baseline disease status, as in the case of low baseline disease ascertainment, the misspecified survival model had a greater effect on bias at early time points, with the baseline disease risk underestimated by 8.3%. At later times, the misspecified survival model resulted in overestimating the cumulative risk, with the greatest amount of overestimation at time 6.9 (5.4%).

We also applied the methods described in Section 3.1 to the simulation data to show that the bias for the NPMLC cumulative risk estimator was less than 1%, regardless of the level of disease ascertainment. These results are presented in Table IV.

5. Application to women who test HPV-positive/Pap-negative

The most common abnormal screening result at KPNC is HPV-positive but Pap-negative [22]. It is the most challenging abnormality to manage because the Pap test might have missed precancer/cancer, but more likely, the HPV infection might naturally clear without intervention. Accurately estimating the cumulative risk of cervical intraepithelial neoplasia grade 3 and cancer (CIN3+), that is, precancer/cancer, is an important step in developing appropriate management guidelines. We compared the cumulative risk estimates of CIN3+ obtained using different methods. We also compared the overall and age-stratified risk curve of HPV-positive/Pap-negative women to the risk curves of other abnormal results for which management strategies are well established. We assumed that prevalent disease status was MAR given the baseline cotest result, because whether women have definitive disease ascertainment at baseline depended primarily on the KPNC protocols for their baseline cotest result.

From 2003–2013, 34,261 women age 30–65, with no prior history of screening abnormalities, precancers, or treatments, tested HPV-positive/Pap-negative at their first ‘cotesting’ (testing with both HPV and Pap) screen. According to KPNC protocols, these women would have been told to forgo immediate colposcopy and undergo a repeat cotesting screen in 1 year, where if they tested either HPV-positive or Pap-positive, they would be referred to colposcopy for definitive disease ascertainment. However, only 12,058 (35.2%) of women had repeat cotesting 1 year later; 1,249 (3.6%) opted for immediate colposcopy, while others either returned late for repeat cotesting (the largest return time was over 10 years later) or did not return at all.

Among the population of HPV-positive/Pap-negative women, 1,056 CIN3+ were diagnosed, of which 61 (8.8%) were classified as cancers. Because early cervical cancers and precancers usually have no symptoms, we treated censoring as uninformative. We applied an algorithm to patients’ longitudinal history of cotesting and colposcopy results so that the intervals (L_i, R_i) reflect medical opinion of the smallest reasonable interval in which disease onset can occur. For example, we accepted negative cotesting results as a surrogate for ascertaining that a woman did not have precancer/cancer (<CIN3), because HPV-positivity is a necessary precursor to cervical cancer and negative cotesting results have a very low false positive rate [26]. Among the 34,261 women testing HPV-positive/Pap-negative at their first cotesting screen, 61 (0.18%) were diagnosed with CIN3+ at an immediate colposcopy visit. The 995 (2.90%) were diagnosed with CIN3+ in follow-up, of which we could rule out prevalent CIN3+ for only 292 (0.85%). The 24,282 (70.87%) were right-censored, while 8,923 (26.0%) did not have sufficient history to classify

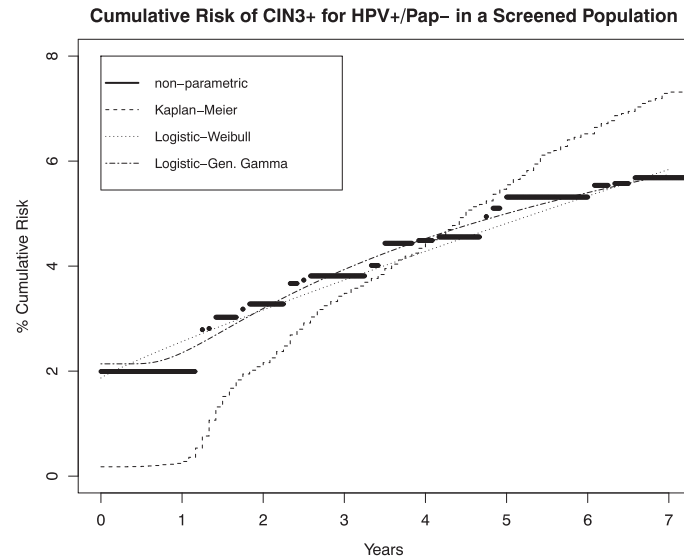


Figure 1. Cumulative risk of CIN3+ for HPV+/Pap- in a screened population. Percentage cumulative risk of cervical intraepithelial neoplasia grade 3 and cancer (CIN3+) following a HPV positive/ Pap-negative result at the initial screen.

as either CIN3+ or <CIN3 at any time point. The percentage risk of a prevalent CIN3+ in the sample is then between $100 \times 61 / (24,282 + 61 + 995)\% = 0.24\%$ and $100 \times (61 + 995 - 292) / (24,282 + 61 + 995)\% = 3.02\%$.

5.1. Comparison of cumulative risk estimates

For women testing HPV-positive/Pap-negative, we estimated the cumulative risk of CIN3+ using our non-parametric risk estimator, the Kaplan–Meier estimator treating time of diagnosis as occurrence time, and various parametric prevalence incidence models fitted without covariates. For parametric prevalence incidence models, we modelled prevalent disease using an intercept-only logistic regression and considered the Weibull, exponential, loglogistic, lognormal, generalized gamma, and gamma distributions for the incidence model. Confidence intervals were obtained by estimating confidence limits on the asymptotically normal complementary log–log scale and converting those confidence limits to the cumulative risk scale. Figure 1 shows the cumulative risk curve of CIN3+ from the non-parametric risk estimator, the Kaplan–Meier estimator, the logistic–Weibull model, and the logistic-generalized-gamma model, which is the best fitting parametric prevalence–incidence model according to the Bayesian information criterion (BIC).

The non-parametric cumulative risk curve was a step function that is constant for long intervals followed by large jumps, reflecting the clustering in the times that women return for further screening. For example, both the non-parametric prevalent and 1-year cumulative risk of CIN3+ were 1.99%, but the non-parametric cumulative risk jumped to 2.79% at 14 months. The non-parametric risk curve was flat in the first year because few women returned early for screening, and the jump at 14 months reflects the large clustering of diagnoses from those who returned at 1 year, tested HPV-positive or Pap-positive at that visit, and were referred to colposcopy visits for definitive disease ascertainment. Because decision frameworks for action are based on risks at specific time points [8], the non-parametric estimator cannot be directly used for determining guidelines. However, the non-parametric cumulative risk can be used to assess fit of other estimators. The non-parametric estimator estimated 1.99% prevalent CIN3+ risk and 5.68% 7-year cumulative risk of CIN3+. In contrast, Kaplan–Meier methods estimated merely 0.18% prevalent CIN3+ risk, reflecting the proportion of HPV-positive/Pap-negative women who were diagnosed with CIN3+ at immediate colposcopy visits. Kaplan–Meier methods presumed that all CIN3+ diagnosed at future times were truly incident disease and overestimated the hazard of incident CIN3+, leading to a 7.31% 7-year cumulative risk estimate.

The logistic–Weibull model estimate of 1.87% prevalent CIN3+ risk was close to the non-parametric estimate of 1.99%, showing that prevalence–incidence models can estimate prevalent disease risk even when baseline disease is rarely ascertained. Confidence intervals for the logistic–Weibull model were

Table V. Percentage cumulative risk and 95% confidence intervals of acquiring cervical intraepithelial neoplasia grade 3 or cancer (CIN3+) following an HPV-positive/Pap-negative result at the initial screen as estimated by the non-parametric estimator and by logistic–Weibull prevalence incidence models.

Year	Non-parametric			Logistic–Weibull			% IR
	Est.	LCL	UCL	Est.	LCL	UCL	
0	1.99	0.00	2.30	1.87	1.59	2.19	74
1	1.99	1.70	2.30	2.56	2.36	2.78	30
2	3.28	2.90	3.55	3.17	2.95	3.39	32
3	3.81	3.47	4.11	3.73	3.51	3.98	27
4	4.49	4.12	4.80	4.28	4.03	4.55	24
5	5.31	4.60	5.67	4.81	4.53	5.11	46
6	5.31	5.03	5.83	5.33	5.01	5.67	60
7	5.68	5.25	6.12	5.84	5.47	6.24	11

Note: The % interval reduction (IR) is the percent reduction in confidence interval length from the parametric assumptions.

tighter than that of the bootstrap confidence intervals for the non-parametric estimator (Table V). The parametric assumption resulted in a 74% reduction in confidence interval for the baseline cumulative risk, and a smaller reduction at later time points. The reduction in confidence interval lengths can be considered a positive or a negative feature, depending on how much one believes the parametric assumptions.

The logistic–generalized gamma model (BIC = 10, 106) was a better fit to the data than the logistic–Weibull model (BIC = 10, 117). However, the logistic–generalized gamma model estimated zero hazards for the first 7 months following the initial HPV-positive/Pap-negative cotesting result; even though there was little data within this time period, zero hazards for HPV-infected women is unrealistic [27] and more likely a result of overfitting. Regardless of which distribution (Weibull, exponential, loglogistic, lognormal, generalized gamma, or gamma) was used for the incidence model, the cumulative risk curves were similar to one another in respect to the resulting recommendations for guidelines and were close fits to the non-parametric risk curve. We favor using the logistic–Weibull model over other parametric prevalence–incidence models, as it allows covariate effects to be described in terms of odds ratios for prevalent disease and hazard ratios for incident disease.

5.2. Implications for screening guidelines

Management of abnormal Pap screening results in the absence of an HPV test is well established. Women with atypical cells of undetermined significance (ASC-US) Pap results are asked to return for repeat screening in 1 year, while women with low-grade squamous intraepithelial lesion (LSIL) or worse Pap results are referred to immediate colposcopy [22]. We fitted the logistic–Weibull model to KPNC data and compared the cumulative risk curves for these groups of women with women with HPV-positive/Pap-negative cotesting results (Figure 2). For HPV-positive/Pap-negative women, the estimated baseline CIN3+ risk of 1.87% (95% CI: 1.59–2.19%) was lower than the 1.99% (95% CI: 1.24–3.17%) baseline CIN3+ risk estimated for women with ASC-US Pap results and much lower than the 3.87% (95% CI: 3.49–4.29%) baseline CIN3+ risk estimated for women with LSIL Pap results. Applying an equal management of equal risk principle [8], the baseline CIN3+ risk estimates suggest that women with HPV-positive/Pap-negative cotesting results should not be referred to immediate colposcopy.

The 1-year CIN3+ risk among women with HPV-positive/Pap-negative results (2.56%, 95% CI: 2.36–2.78%) was similar to the 1-year CIN3+ risk among women with ASC-US Pap results (2.50%, 95% CI: 1.97–3.17%). However, the hazards of incident CIN3+ were greater among women with HPV-positive/Pap-negative cotesting results than women with ASC-US Pap results, so that at 5 years, the cumulative risks of CIN3+ were 4.81% (95% CI: 4.53–5.11%) and 3.37% (95% CI: 2.84–4.00%) for women with HPV-positive/Pap-negative and ASC-US Pap results, respectively. These risk estimates suggest that women with HPV-positive/Pap-negative results, like women with ASC-US Pap results, should return for a repeat cotesting screen in one year. However, for HPV-positive/Pap-negative results, clinicians will need to be especially cautious that women do not skip their return visits.

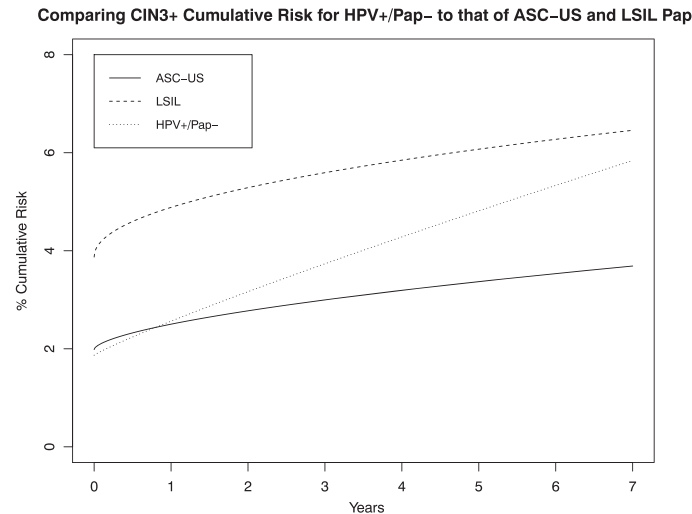


Figure 2. Comparing CIN3+ cumulative risk for HPV+/Pap- to that of ASC-US and LSIL Pap. Curves showing the percentage cumulative risk of cervical intraepithelial neoplasia grade 3 and cancer (CIN3+) following (1) HPV-positive/Pap-negative cotesting results, (2) Pap results of atypical cells of undetermined significance (ASC-US), and (3) Pap results of low-grade squamous intraepithelial lesion (LSIL).

Table VI. Result of fitting a logistic–Weibull model with categorical age to women with an initial HPV-positive/Pap-negative cotesting result at KPNC.

Age category	Percentage sample	OR	HR	Percentage prevalent risk	1 year % CR	2 years % CR	3 years % CR
30–34	37.7	1.00	1.00	1.91	2.88	3.75	4.57
35–39	18.4	1.13	0.63	2.14	2.76	3.30	3.83
40–44	13.7	1.38	0.59	2.61	3.17	3.68	4.16
45–49	10.6	0.88	0.38	1.67	2.05	2.38	2.71
50–54	8.4	0.55	0.37	1.06	1.42	1.75	2.06
55–65	11.2	0.63	0.64	1.21	1.85	2.41	2.94

Note: Percentage sample is the percentage of women with HPV-positive/Pap-negative cotesting result in that age category. OR are odds ratio of having CIN3+ prevalent at the initial cotesting screen for that age group compared with the reference group (age 30–34). HR are hazard ratios of acquiring incident CIN3+ after the initial cotesting screen for that age group compared with the reference group (age 30–34). Percentage prevalent risk refers to the percent risk of having CIN3+ prevalent at the initial cotesting screen. 1 year, 2 years, and 3 years Percentage CR refers to the percent cumulative risk of having CIN3+ 1 year, 2 years, and 3 years after the initial cotesting screen, respectively.

Women with HPV-positive/Pap-negative cotesting result may consist of sub-populations with distinctly different risk profiles [28]; additional information can be used to further personalize management. We considered how risk profiles may differ for women of different ages by fitting age as a categorical covariate in the logistic–Weibull model (Table VI). No age group had risks of prevalent CIN3+ high enough to warrant referral to immediate colposcopy. The 2-year cumulative risks of CIN3+ among women of ages 45–49, 50–54, and 55–65 were 2.38% (95% CI: 1.89–3.00%), 1.75% (95% CI: 1.30–2.36%), and 2.41% (95% CI: 1.94–2.99%), respectively. Compared with the 2.50% implicit cumulative risk threshold for returning women for repeat screening in 1-year that is implied by women with ASC-US Pap results, these risks suggest that a longer return time may be reasonably safe with older women. Further analysis that consider the benefits (e.g., potential reduction in overtreatment) would need to be considered before recommending an extended screening interval for these age groups.

6. Discussion

Cohorts assembled from electronic health records at health providers can have irregularly interval-censored outcomes, and prevalent left-censored disease is undetermined when disease ascertainment is not conducted at the first available visit. We demonstrated that Kaplan–Meier methods are inappropriate and proposed a general family of mixture models, called prevalence–incidence survival models. We

presented an EM algorithm to fit parametric prevalence–incidence models with covariates and obtained a non-parametric estimate (no covariates) by adapting standard NPML methods. The non-parametric estimate can be used to assess the fit of parametric prevalence–incidence models such as the logistic–Weibull. For a cohort undergoing cervical cancer screening at KPNC, Kaplan–Meier yielded poor risk estimates and the non-parametric cumulative risk is a step function with large jumps. In contrast, the logistic–Weibull yielded a smooth risk curve and agreed with the non-parametric estimates. The logistic–Weibull allows for covariate effects to be described in terms of odds ratios for prevalent disease and hazard ratios for incident disease and the Weibull distribution has previously been suggested for modeling cancer [19]. The ability to separate out the risk of disease present at baseline versus disease that occurs during follow-up is important for informing clinical decisions on whether to intervene with a surgical procedure at the initial screening visit (baseline). These findings support the choice of the logistic–Weibull model to estimate risks that underlie current US risk-based cervical cancer screening guidelines [21].

Kaplan–Meier is biased under interval-censoring, even with ad hoc schemes to impute event onset within intervals [29]. We demonstrated that the bias has an interesting pattern: underestimation of risks at early times and overestimation of risks at later times. This bias is exacerbated by the presence of prevalent left-censored outcomes. Most work on left-censoring involves inference on the past time of onset and assumes that it is known a priori whether outcomes are left-censored (prevalent) or interval-censored (incident) [30]. For clinical use, a simple estimate of the total amount of prevalent disease as a point-mass at zero can suffice. Definitive disease ascertainment may not be conducted at the first available visit, so that some disease diagnosed during follow-up are also pre-existing disease.

The proposed mixture model is particularly applicable to our data, because women who test HPV-positive/Pap-negative at the first available screen consist of two populations: (1) women with long-term HPV infections who, may already have CIN3+, and (2) women with recently acquired HPV infections, most of whom will clear their infections, but can have future CIN3+ if the infection persists [28]. We believe that Kaplan–Meier methods should be avoided. The non-parametric estimate is robust, but constant for long intervals followed by large jumps, and loses efficiency due to the slow $n^{1/3}$ asymptotic convergence rate [31]. The best methodology is to fit a parametric model, such as the logistic–Weibull, and check that it is a good fit to the non-parametric estimate. Our findings support the choice of using logistic–Weibull models for KPNC data to estimate the risks that underlie current US cervical screening guidelines.

Two extensions of prevalence–incidence models would be useful. The first is to develop semi-parametric versions of the prevalence–incidence models that incorporate covariates. Fitting a Cox model to interval-censored data has previously been studied [32–34]. These methods might be adapted to prevalence–incidence survival models. The second is to develop methodology to fit prevalence–incidence models to general epidemiologic study designs, such as case-control studies, while accounting for complex sampling.

Acknowledgements

This research was supported, in part, by the Intramural Research Program of the NIH/NCI. We thank Dr. Joseph Gastwirth and Dr. Huixia Wang for their many suggestions. We thank Dr. Gene Pawlick (Regional Laboratory of the Northern California Kaiser Permanente Medical Care Program) for creating and supporting the data warehouse, and Kaiser Permanente Northern California for allowing use of the data.

References

1. Marcus PM, Freedman AN, Khoury MJ. Targeted cancer screening in average-risk individuals. *American Journal of Preventive Medicine* 2015; **49**:765–771.
2. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: toward better research applications and clinical care. *Nature Reviews Genetics* 2012; **13**:395–405.
3. Sturmer T, Funk MJ, Poole C, Brookhart MA. Nonexperimental comparative effectiveness research using linked healthcare databases. *Epidemiology* 2011; **22**:298–301.
4. Kaplan EL, Meier P. Nonparametric estimation for incomplete observations. *Journal of American Statistical Association* 1958; **53**:457–481.
5. Cox DR. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society – Series B* 1972; **34**:187–220.
6. Schick A, Yu Q. Consistency of the GMLE with mixed case interval-censored data. *Scandinavian Journal of Statistics* 2000; **27**:45–55.

7. Panageas KS, Ben-Porat L, Dickler MN, Chapman PB, Schrag D. When you look matters: the effect of assessment schedule on progression-free survival. *Journal of the National Cancer Institute* 2007; **99**:428–432.
8. Katki HA, Wacholder S, Solomon D, Castle PE, Schiffman M. Risk estimation for the next generation of prevention programmes for cervical cancer. *Lancet Oncology* 2009; **10**:1022–1023.
9. Huang J, Wellner JA. Interval censored survival data: a review of recent progress. *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*. Springer, New York, 1997, 123–169 p.
10. Lindsey JC, Ryan LM. Tutorial in biostatistics: methods for interval-censored data. *Statistics in Medicine* 1998; **17**: 219–238.
11. Zhang Z, Sun J. Interval censoring. *Statistical Methods in Medical Research* 2010; **19**:53–70.
12. Dorey FJ, Little RJA, Schenker N. Multiple imputation for threshold-crossing data with interval censoring. *Statistics in Medicine* 1993; **12**:1589–1603.
13. Rücker G, Messerer D. Remission duration: an example of interval-censored observations. *Statistics in Medicine* 1988; **7**:1139–1145.
14. Boag JW. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society – Series B* 1949; **11**:15–44.
15. Berkson J, Gage RP. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association* 1952; **47**:501–515.
16. Farewell VT. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* 1982; **38**:1041–1046.
17. Li CS, Taylor JMG, Sy JP. Identifiability of cure models. *Statistics and Probability Letters* 2001; **54**:389–395.
18. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society – Series B* 1977; **39**:1–38.
19. Armitage P, Doll R. The age distribution of cancer and a multi-stage theory of carcinogenesis. *British Journal of Cancer* 1954; **8**:1–12.
20. Katki HA, Kinney WK, Fetterman B, Lorey T, Poitras NE, Cheung L, Demuth F, Schiffman M, Wacholder S, Castle PE. Cervical cancer risk for women undergoing concurrent testing for human papillomavirus and cervical cytology: a population-based study in routine clinical practice. *Lancet Oncology* 2011; **12**:663–672.
21. Massad LS, Einstein MH, Huh WK, Katki HA, Kinney WK, Schiffman M, Solomon D, Wentzensen N, Lawson HW. 2012 updated consensus guidelines for the management of abnormal cervical cancer screening tests and cancer precursors. *Journal of Lower Genital Tract Disease* 2013; **17**:S78–S84. <https://www.asccp.org/store.detail2/asccp-mobile-app> [Accessed on 21 June 2017].
22. Katki HA, Schiffman M, Castle PE, Fetterman B, Poitras NE, Lorey T, Cheung LC, Raine-Bennett TR, Gage JC, Kinney WK. Benchmarking CIN 3+ Risk as the basis for incorporating HPV and Pap cotesting into cervical screening and management guidelines. *Journal of Lower Genital Tract Disease* 2013; **17**:S28–S35.
23. Wellner JA, Zhan Y. A hybrid algorithm for computation of the nonparametric maximum likelihood estimator of the distribution function. *Journal of American Statistical Association* 1997; **92**:945–959.
24. Ryaben’kii VS, Tsykov SV. *A Theoretical Introduction to Numerical Analysis*. CRC Press: Boca Raton FL, 2006.
25. Lange K. A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society: Series B* 1995; **57**:425–437.
26. Belinson J, Qiao YL, Pretorius R, Zhang WH, Elson P, Li L, Pan QJ, Fischer AL, Zahniser D, Shanxi Province Cervical Cancer Screening Study: a cross-sectional comparative trial of multiple techniques to detect cervical neoplasia. *Gynecologic Oncology* 2001; **83**:439–444.
27. Schiffman M, Castle PE, Jeronimo J, Rodriguez AC, Wacholder S. Human papillomavirus and cervical cancer. *Lancet* 2007; **370**:809–907.
28. Katki HA, Cheung LC, Fetterman B, Castle PE, Sundaram R. A joint model of persistent human papilloma virus infection and cervical cancer risk: implications for cervical cancer screening. *Journal of the Royal Statistical Society: Series A* 2015; **178**:903–923.
29. Law CG, Brookmeyer R. Effects of mid-point imputation on the analysis of doubly censored data. *Statistics in Medicine* 1992; **11**:1569–1578.
30. Hosmer DW, Lemeshow S, May S. *Applied Survival Analysis: Regression Modeling of Time-to-Event Data* 2nd edition. Wiley: Hoboken, NJ, 2008.
31. Groeneboom P, bounds Wellner JA. *Information Bounds and Nonparametric Maximum Likelihood Estimation*, DMV Seminar, Band 19. NY: Birkhäuser, 1992.
32. Huang J. Efficient estimation for the proportional hazards model with interval censoring. *Annals of Statistics* 1996; **24**: 540–568.
33. Pan W. A multiple imputation approach to Cox regression with interval-censored data. *Biometrics* 2000; **56**:199–203.
34. Goetghebeur E, Ryan L. Semiparametric regression analysis of interval-censored data. *Biometrics* 2000; **56**:1139–1144.

Supporting information

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