

## FLEXIBLE RISK PREDICTION MODELS FOR LEFT OR INTERVAL-CENSORED DATA FROM ELECTRONIC HEALTH RECORDS

BY NOORIE HYUN\*, LI C. CHEUNG\*, QING PAN<sup>†</sup>, MARK SCHIFFMAN\* AND HORMUZD A. KATKI\*

*National Cancer Institute\* and George Washington University<sup>†</sup>*

Electronic health records are a large and cost-effective data source for developing risk-prediction models. However, for screen-detected diseases, standard risk models (such as Kaplan–Meier or Cox models) do not account for key issues encountered with electronic health record data: left-censoring of pre-existing (prevalent) disease, interval-censoring of incident disease, and ambiguity of whether disease is prevalent or incident when definitive disease ascertainment is not conducted at baseline. Furthermore, researchers might conduct novel screening tests only on a complex two-phase subsample. We propose a family of weighted mixture models that account for left/interval-censoring and complex sampling via inverse-probability weighting in order to estimate current and future absolute risk: we propose a weakly-parametric model for general use and a semiparametric model for checking goodness of fit of the weakly-parametric model. We demonstrate asymptotic properties analytically and by simulation. We used electronic health records to assemble a cohort of 33,295 human papillomavirus (HPV) positive women undergoing cervical cancer screening at Kaiser Permanente Northern California (KPNC) that underlie current screening guidelines. The next guidelines would focus on HPV typing tests, but reporting 14 HPV types is too complex for clinical use. National Cancer Institute along with KPNC conducted a HPV typing test on a complex subsample of 9258 women in the cohort. We used our model to estimate the risk due to each type and grouped the 14 types (the 3-year risk ranges 21.9–1.5) into 4 risk-bands to simplify reporting to clinicians and guidelines. These risk-bands could be adopted by future HPV typing tests and future screening guidelines.

**1. Introduction.** Many large-scale epidemiologic cohort studies are being organized within health-care providers who have large populations of patients to recruit, preexisting infrastructure for longitudinal visits, and electronic health records to facilitate data collection. For example, we collaborated with Kaiser Permanente Northern California (KPNC) to assemble a cohort of women in cervical cancer screening by linking electronic records of patient information, test results and disease outcomes [Castle et al. (2009)]. Nearly all women underwent testing

---

Received July 2016; revised February 2017.

*Key words and phrases.* Mixture model, interval censoring, two-phase sampling, B-splines, weighted likelihood, HIV.

for human papillomavirus (HPV), the cause of nearly all cervical cancer. We previously used this cohort to develop the cancer risk calculations underlying current HPV-based screening guidelines [Katki et al. (2013), Massad et al. (2013)], which are available in the official guidelines App (<http://www.asccp.org/store-detail2/asccp-mobile-app>).

In light of our experience, we have developed new risk modeling methodology for electronic health record data for screen-detected diseases. We address three key issues that make it inappropriate to calculate risk using standard methods, such as Kaplan–Meier [Kaplan and Meier (1958)] or Cox models [Cox (1972)].

First, prevalent disease could exist at enrollment, and separating out risk of prevalent disease is important because clinicians are primarily concerned with the risk that disease is present. Furthermore, doctors have little interest in when a cancer currently detected might have arisen in the past. Thus, it suffices to consider prevalent disease as a left-censored point-mass at time zero, taken as the earliest time at which there exist health records for the outcomes and covariates. The idea of modeling prevalent disease as a point mass at time zero is the obverse of the cure model for two heterogeneous sub-populations, where there is a point mass at time infinity [Li, Taylor and Sy (2001), Ma (2010), Shao et al. (2014)].

However, prevalent disease is not always diagnosed at baseline. People with missing or negative screening test results generally do not undergo definitive disease ascertainment, such as biopsies. Consequently, disease diagnosed at future visits is a mixture of truly incident disease and undiagnosed prevalent disease. A mixture of prevalent and incident disease is a key feature of health record data; it is also commonly found but ignored in epidemiologic cohorts, for example, in case-cohort studies to estimate the incidence rate for an asymptomatic disease, cases diagnosed after baseline are considered to occur after baseline by assuming diagnosis dates are equal to disease onset dates.

The second key issue is that incident disease events are often interval-censored between irregular visits. Researchers working with data from health providers typically cannot influence the timing of visits, and patients return at intervals that are quite irregular. Ignoring interval-censoring leads to invalid inferences [Dorey, Little and Schenker (1993), Odell, Anderson and D’Agostino (1992), Rucker and Messerer (1988)], especially when intervals are irregular. Furthermore, standard interval-censoring methods [cf. Huang and Rossini (1997), Huang and Wellner (1997), Ma (2010), Tian and Cai (2006), Wang et al. (2016), Zhang, Hua and Huang (2010)] do not account for diagnosed or undiagnosed prevalent disease.

The final issue we address is estimating absolute risk from two-phase stratified samples nested within the cohort. Electronic health record information is available on everyone (phase 1), and the new screening tests are available only on a sample of the cohort (phase 2). Conducting biomarker measurements only on a judicious sample can be cost-efficient in using cohort resources [Woodward (1999)]. Estimating absolute risk for the full cohort requires accounting for the sample design,

for example, the sampling fractions. We focus on the Horvitz–Thompson (design-based) estimation [Horvitz and Thompson (1952)] in this manuscript.

We propose a family of mixture models, called “prevalence-incidence” models, for estimating cumulative risk and assessing covariate effects. All details are presented for the useful special case where prevalent disease is modeled with logistic regression and incident disease is modeled with a Cox model (“logistic-Cox”). The semiparametric model is computationally intensive, and estimates asymptotically converge at slow rates though, recent high-level computational resources along with big data can solve the problems. However, low event rates despite a large sample and a set of multiple data analyses can be barriers to using the semiparametric model with a bootstrap-based inference procedure. We propose a weakly-parametric model using a monotone spline for the baseline cumulative hazard. As a practical solution for diagnostic purposes, we propose using the semiparametric estimates to graphically assess the fit of the weakly-parametric model, and an iterative algorithm is used to estimate parameters in a semiparametric framework. We extend our models to account for two-phase stratified sampling via inverse-probability weighting (IPW) by sample inclusion probabilities [cf. Breslow and Wellner (2007), Cai and Zheng (2013), Kovalchik and Pfeiffer (2014), Saegusa (2015)].

We used electronic health records to link data at KPNC for 33,295 HPV-positive women from 2007–2011 to assemble the HPV Persistence and Progression (PaP) Cohort (see Figure 1 for details). We plan to use this data and our prevalence-incidence models to inform the next screening guidelines that will incorporate new screening tests, in particular, HPV typing [Castle et al. (2011)]. Each of the 14 carcinogenic HPV types has different precancer/cancer risk [Schiffman et al. (2011)], but providing information on each of 14 types is too complex for clinicians or developing guidelines. We conducted HPV typing tests in PaP using a residual exfoliated cervical specimen that was stored for study use [Schiffman et al. (2015)]. However, typing tests are too expensive to be used to test all specimens. Instead, we conducted typing tests on a stratified random sample of 9258 women that oversamples women diagnosed with precancer/cancer or are otherwise at high risk (the design will be elaborated in Section 4). Using the IPW logistic-Cox model to calculate risk in PaP, we grouped the 14 types by risk into 4 risk-bands to report to clinicians and for basing guidelines. These risk-bands may be adopted by future screening guidelines, which would inform the design of future commercial HPV typing tests.

**2. Proposed methods.** We first propose the prevalence-incidence family of models for full cohorts and then extend it to two-phase samples. Throughout, we assume that all outcomes and covariates have negligible measurement error.

*2.1. Complete data of a full cohort.* For full cohort data (no subsampling), denote subjects  $i = 1, \dots, N$ , the failure time,  $T_i$  has cumulative density function

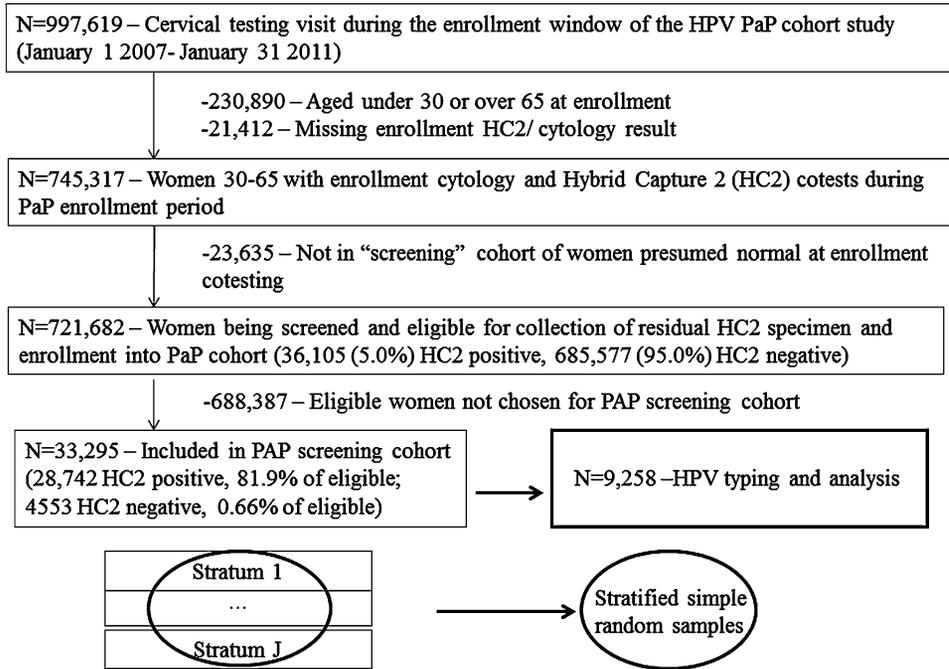


FIG. 1. Human papillomavirus (HPV) Persistence and Progression (PaP) cohort.

$F$ , and its survival function is  $S(t) = 1 - F(t)$ . The time scale is time-on-study, and we suppose the baseline time is 0 at enrollment into the study. If subject  $i$  has disease at baseline (i.e.,  $T_i \leq 0$ ), then  $Y_i = 1$ ; otherwise  $Y_i = 0$ . The prevalence indicator variable,  $Y_i$  is not necessarily observed for all subjects. As the missing indicator,  $M_i$  has value 1 if  $y_i$  is observed and 0 if  $y_i$  is missing. Failure times are interval-censored between  $L_i$  and  $R_i$ , the latest and earliest visit times at which the subject  $i$  is observed as disease-free and diseased, respectively. Intervals are defined as follows: for  $0 < L_i < R_i$ , right-censoring is  $(L_i, R_i = \infty)$ , interval-censoring where disease is definitively known to be not present at baseline is  $(0, R_i)$  or  $(L_i, R_i)$ , and disease that is diagnosed in the follow-up but might be unobserved at baseline (i.e.,  $M_i = 0$ ) is  $[0, R_i)$  for  $R_i < \infty$ . We assume that given covariates, the censoring time and observation time are independent of the failure time because visit time is predetermined by guidelines and precancers and early-phase cancers are most likely to be asymptomatic. If case status, that is, diseased versus disease-free at baseline or during the enrollment period (prevalence at baseline or incidence observed during the enrollment period), are used to determine strata, the auxiliary variable of  $V_i$  (not the risk factors of interest) includes  $(Y_i, L_i, R_i)$  in addition to other characteristics, for example, strata factors and demographics.

We assume that the prevalent disease probability,  $P_d(x_i, \beta)$  at baseline depends on  $\beta$  for a given covariate  $x_i$ , which does not overlap  $\gamma$  for incident

probability given an incidence-related covariate,  $z_i$ . The covariate vectors,  $\mathbf{X}_i$  and  $\mathbf{Z}_i$  are partially overlapped or the same, and for example, can be potential risk factors for cancers at baseline. The likelihood for complete-data,  $\mathbf{D}_c = \{(y_i, L_i, R_i, \mathbf{v}_i^T \mathbf{x}_i^T, \mathbf{z}_i^T); i = 1, \dots, N\}$  is

$$(2.1) \quad L_N(\boldsymbol{\beta}, \boldsymbol{\gamma}, S; \mathbf{D}_c) = \prod_{i=1}^N P_d(\mathbf{x}_i, \boldsymbol{\beta})^{y_i} [1 - P_d(\mathbf{x}_i, \boldsymbol{\beta})] \times \{S(L_i; \mathbf{z}_i, \boldsymbol{\gamma}) - S(R_i; \mathbf{z}_i, \boldsymbol{\gamma})\}^{1-y_i}.$$

The above likelihood defines a general class of “prevalence-incidence” mixture models. In particular, we focus on the logistic-Cox prevalence-incidence model, which models prevalent disease with a logistic regression and incident disease with a Cox model [Cox (1972)], that is,  $P_d(\mathbf{x}, \boldsymbol{\beta}) = \exp(\mathbf{x}\boldsymbol{\beta}) / \{1 + \exp(\mathbf{x}\boldsymbol{\beta})\}$  and  $S(t; \mathbf{z}, \boldsymbol{\gamma}) = \exp\{-\Lambda(t) \exp(\mathbf{z}\boldsymbol{\gamma})\}$ , where  $\Lambda(t)$  is an unknown baseline cumulative hazard function, which is nondecreasing over time and  $\Lambda(0) = 0$ . Cumulative risk from the logistic-Cox model given  $\mathbf{x}$  and  $\mathbf{z}$  is

$$(2.2) \quad CR(t | \mathbf{x}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \Lambda) = \frac{\exp(\mathbf{x}\boldsymbol{\beta})}{1 + \exp(\mathbf{x}\boldsymbol{\beta})} + \frac{1}{1 + \exp(\mathbf{x}\boldsymbol{\beta})} [1 - \exp\{-\Lambda(t) \exp(\mathbf{z}\boldsymbol{\gamma})\}].$$

2.2. *Two-phase stratified sample.* For two-phase stratified sample design, we follow the general inverse-probability weighting (IPW) approach [Breslow and Wellner (2007)]. The first phase is the full cohort of  $N$  subjects, which is a simple random sample from an infinite population (called superpopulation). For subjects  $i = 1, \dots, N$ , at phase 1, we observe only a vector of auxiliary variables  $\mathbf{V}_i$ , which correlates with the time-to-precancer/cancer,  $T_i$  and determines stratification. In the HPV-PaP cohort, the auxiliary information includes the currently used cotesting for cervical cancer screening (cytology and HC2) and demographics. We suppose the cohort is divided into  $J$  mutually exclusive and exhaustive strata. Let  $N_j$  denote the number of subjects in the  $j$ th stratum for  $j = 1, \dots, J$ , so  $N = \sum_{j=1}^J N_j$ . At phase 2, simple random samples without replacement of size  $n_j$  are drawn from each of the  $J$  finite phase 1 strata and  $n = \sum_{j=1}^J n_j$ . We denote  $\xi_{j,i}$  as the indicator variable equal to one if the  $i$ th subject in stratum  $j$  is sampled at phase 2 and zero otherwise. Under this two-phase stratified sample design,  $(\xi_{j,1}, \dots, \xi_{j,N_j})$  are exchangeable with  $\Pr(\xi_{j,i} = 1) = n_j / N_j$ , and the  $J$  random vectors  $(\xi_{j,1}, \dots, \xi_{j,N_j})$  are independent. With two-phase sampling,  $\mathbf{X}$  and  $\mathbf{Z}$  are not observed for all  $N$  subjects but fully observed for subjects sampled at phase 2, for example, expensive bioassay tests are only conducted on the subjects sampled in phase 1. For the general setting, let  $\pi_{j,i} = \Pr(\xi_{j,i} = 1)$  be the probability that the  $i$ th subject from stratum  $j$  is sampled at phase 2. Then served

data at phase one is  $\mathbf{D}^{(1)} = \{D_{j,i}^1, i = 1, \dots, N_j, j = 1, \dots, J\}$  where  $D_{j,i}^{(1)} = \{y_{j,i} = 1, \mathbf{v}_{j,i}, \xi_{j,i}\}$  or  $\{y_{j,i} = 0, L_{j,i}, R_{j,i}, \mathbf{v}_{j,i}, \xi_{j,i}\}$  when  $M_{j,i} = 1$ , whereas  $D_{j,i}^{(1)} = \{L_{j,i} = 0, R_{j,i}, \mathbf{v}_{j,i}, \xi_{j,i}\}$  when  $M_{j,i} = 0$ . At phase two, the observed data is  $\mathbf{D}^{(2)} = \{D_{j,i}^{(2)}, i = 1, \dots, n_j, j = 1, \dots, J\}$ , where  $D_{j,i}^{(2)} = \{\xi_{j,i}\mathbf{x}_{j,i}, \xi_{j,i}\mathbf{z}_{j,i}\}$ . Hence, the observed data from phase-two stratified sampling are  $\mathbf{D} = \mathbf{D}^{(1)} \cup \mathbf{D}^{(2)}$ . We assume the missing mechanism for  $Y$  at phase one and sample selection at phase two are missing at random (MAR). We also assume that all  $J$  strata are sampled with positive probability.

Then the weighted likelihood for the observed data  $\mathbf{D}$  and the missing indicator  $\mathbf{M}$  is

$$\begin{aligned}
 L_n^\pi(\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T, S, \Psi; \mathbf{D}, \mathbf{M}) &= \prod_{j=1}^J \prod_{\{i; m_{j,i}=1\}}^{N_j} \frac{N_j}{n_j} \xi_{j,i} P_d(\boldsymbol{\beta})^{y_{j,i}} [\{1 - P_d(\boldsymbol{\beta})\} \{S(L_{j,i}; \boldsymbol{\gamma}) \\
 (2.3) \quad &- S(R_{j,i}; \boldsymbol{\gamma})\}]^{1-y_{j,i}} P(M_{j,i} = m_{j,i} \mid D_{j,i}, \Psi) \\
 &\times \prod_{j=1}^J \prod_{\{i; m_{j,i}=0\}}^{N_j} \frac{N_j}{n_j} \xi_{j,i} [P_d(\boldsymbol{\beta}) + \{1 - P_d(\boldsymbol{\beta})\} \{1 - S(R_{j,i}; \boldsymbol{\gamma})\}] \\
 &\times P(M_{j,i} = m_{j,i} \mid D_{j,i}, \Psi),
 \end{aligned}$$

where  $\Psi$  denotes parameters in the missing data mechanism. The likelihood in (2.3) is orthogonal in  $(\boldsymbol{\beta}, \boldsymbol{\gamma})$  and  $\Psi$ . Thus, the missing data mechanism  $P(M_{j,i} \mid D_{j,i}, \Psi)$  is ignorable in maximum likelihood estimation. The MAR assumption for  $\pi_{j,i}$  is crucial to construct an unbiased estimating equation.

Then the weighted log-likelihood for the observed data,  $\mathbf{D}$  is

$$(2.4) \quad l_n^\pi(\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T, \Lambda; \mathbf{D}) = \sum_{j=1}^J N_j/n_j \sum_{i=1}^{N_j} \xi_{j,i} l(\boldsymbol{\beta}, \boldsymbol{\gamma}, \Lambda; D_{j,i}),$$

where

$$\begin{aligned}
 l(\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T, \Lambda; D_{j,i}) &= I(m_{j,i} = 1)(y_{j,i}\mathbf{x}_{j,i}\boldsymbol{\beta} - \log\{1 + \exp(\mathbf{x}_{j,i}\boldsymbol{\beta})\} \\
 (2.5) \quad &+ (1 - y_{j,i}) \log[\exp\{-\Lambda(L_{j,i}) \exp(\mathbf{z}_{j,i}\boldsymbol{\gamma})\} \\
 &- \exp\{-\Lambda(R_{j,i}) \exp(\mathbf{z}_{j,i}\boldsymbol{\gamma})\}]) \\
 &+ I(m_{j,i} = 0) \log[1 - \{1 + \exp(\mathbf{x}_{j,i}\boldsymbol{\beta})\}^{-1} \exp\{-\Lambda(R_{j,i}) \exp(\mathbf{z}_{j,i}\boldsymbol{\gamma})\}].
 \end{aligned}$$

Estimates,  $(\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\gamma}}^T, \hat{\Lambda})$  denote the corresponding arguments maximizing the objective function in (2.4). Because the indicator variables,  $\{\xi_{j,i}\}_i^{N_j}$  within  $j$  stratum are interchangeable but not independent when replacement is not allowed, we

need techniques dealing with nonindependent data to prove consistency and weak convergence of the estimates.

2.3. *Weakly parametric mixture models.* It is well known that the asymptotic distribution of cumulative hazard functions from interval-censored data is non-Gaussian converging at rates slower than root-N [Groeneboom and Wellner (1992), Sen and Banerjee (2007)]. Generally, semiparametric estimation procedures for interval censored data are computationally-intensive especially when the number of unique visit times increases, so a bootstrap method for inference is often impractical. To sidestep such challenges, we propose a weakly-parametric model by approximating the baseline cumulative hazard with an integrated B-spline. For smoothing, cubic splines are commonly used in practice [Wang et al. (2016)]. Knots can be placed at the quantiles of the finite visit time points. We also present a semiparametric estimator in Section 2.5 as a benchmark to assess how well the approximation of the baseline hazard function fits. To ensure a convergence rate of square root of sample size, we assume the number of knots for integrated B-spline are fixed. In our experience, the assumption is plausible for data analyses with rare events because the number of finite intervals in which events occur is controlled by screening guidelines, and thus is not increasing proportional to the sample size [Zhang, Hua and Huang (2010)].

We approximate the baseline cumulative hazard as  $\Lambda(t) = \sum_{k=1}^K \exp(b_k) B_k(\cdot)$ , where  $B_k(\cdot)$ 's are integrated B-spline basis functions, which are nondecreasing from 0 to 1 and the  $b_k$ 's are unknown parameters for the basis functions [using  $\exp(b_k)$  ensures nonnegative  $\Lambda(t)$ ]. We omit the subscripts  $j, i$  for simplicity. The weighted log-likelihood in the model is

$$(2.6) \quad l_n^\pi \{ \theta = (\beta^T, \gamma^T, b_1, \dots, b_K); \mathbf{D} \} = \sum_{j=1}^J N_j/n_j \sum_{i=1}^{N_j} \xi_{j,i} l(\theta; D_{j,i}),$$

where

$$\begin{aligned} l(\theta; D) = & I(m = 1) \left( (1 - y) \log \left[ \exp \left\{ - \exp(z\gamma) \sum_{k=1}^K e^{b_k} B_k(L) \right\} \right. \right. \\ & \left. \left. - \exp \left\{ - \exp(z\gamma) \sum_{k=1}^K e^{b_k} B_k(R) \right\} \right] + yx\beta - \log \{ 1 + \exp(x\beta) \} \right) \\ & + I(m = 0) \log \left[ 1 - \{ 1 - \exp(x\beta) \}^{-1} \right. \\ & \left. \times \exp \left\{ - \exp(z\gamma) \sum_{k=1}^K e^{b_k} B_k(R) \right\} \right], \end{aligned}$$

where  $K$  is the number of knots.

The root of the score function,  $\sum_{j=1}^J \sum_{i=1}^{N_j} \xi_{j,i} \dot{l}(\boldsymbol{\theta}; D_{j,i})$ , where  $\dot{l}(\boldsymbol{\theta}; D_{j,i}) = \partial l(\boldsymbol{\theta}; D_{j,i}) / \partial \boldsymbol{\theta}$  (presented in Section 2 of the supplementary materials [Hyun et al. (2017)]) can be found by the Newton–Raphson iterative algorithm. Model identifiability and asymptotic consistency of the estimators obtained from the weakly-parametric procedure are proved in Lemma 3.1 and Theorem 3.1 in Section 3 of the supplementary materials [Hyun et al. (2017)], respectively. The Fisher information matrix,  $I_0 = E\{\dot{l}(\boldsymbol{\theta}_0)\dot{l}(\boldsymbol{\theta}_0)^T\}$  is invertible under the condition A2 in Section 1 of the supplementary materials [Hyun et al. (2017)], and it is shown in Lemma 3.2 in Section 3 of the supplementary materials [Hyun et al. (2017)].

2.4. *Asymptotic variance for the weakly-parametric models.* Standard parametric maximum-likelihood theory is inapplicable because the sampling is without replacement, so the sampling indicator variables  $\xi_{j,i}$  are correlated within a stratum. We follow the weighted likelihood approach from Breslow and Wellner (2007) to demonstrate weak convergence of the estimates for finite population stratified sample. We assume the number and placement of knots are known a priori and independent of sample size.

By using Taylor expansion of  $l_n^\pi(\boldsymbol{\theta}; \mathbf{D})$  in (2.6), we linearize the estimated parameters:

$$\begin{aligned} \sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) &= \sum_{j=1}^J N_j/n_j \sum_{i=1}^{N_j} \xi_{j,i} I_0^{-1} \dot{l}(\boldsymbol{\theta}_0; D_{j,i}) + o_p(1) \\ &= \sum_{i=1}^N I_0^{-1} \dot{l}(\boldsymbol{\theta}_0; D_i) + I_0^{-1} \sum_{j=1}^J \sum_{i=1}^{N_j} (N_j \xi_{j,i}/n_j - 1) \dot{l}(\boldsymbol{\theta}_0; D_{j,i}) \\ &\quad + o_p(1) \end{aligned} \tag{2.7}$$

$$\xrightarrow{d} N(0, I_0^{-1} + I_0^{-1} \Sigma_0 I_0^{-1}), \tag{2.8}$$

where  $\Sigma_0 = \sum_{j=1}^J v_j [(1 - p_j)/p_j] \{E\{\dot{l}(\boldsymbol{\theta}_0)^{\otimes 2} | \mathcal{V}_j\} - E\{\dot{l}(\boldsymbol{\theta}_0) | \mathcal{V}_j\}^{\otimes 2}\}$ ,  $\mathcal{V}_j$  is  $j$  stratum, and  $\mathbf{x}^{\otimes 2} = \mathbf{x}\mathbf{x}^T$  for a vector  $\mathbf{x}$ ; for each stratum  $j = 1, \dots, J$ . As  $N \rightarrow \infty$ , sampling fraction converges with  $p_j$  ( $= \lim n_j/N_j$ ); each stratum size increases at the same rate as  $N$  increases, that is,  $v_j = \lim N_j/N$  and  $0 < v_j < \infty$ . The asymptotic normal limit distribution of the estimators is derived in Theorem 3.2 of the supplementary materials [Hyun et al. (2017)]. The asymptotic variance estimator for  $\hat{\boldsymbol{\theta}}$  consists of two components from phase 1 and 2 finite sample design. By letting  $\ddot{l}(\boldsymbol{\theta}) = \partial \dot{l}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$  for  $\dot{l}(\boldsymbol{\theta})$ , the variance estimators are

$$\text{var}_{\text{ph1}}(\hat{\boldsymbol{\theta}}) = N^{-1} \hat{I}(\boldsymbol{\theta}_0)^{-1} = - \left\{ N \sum_{j=1}^J N_j/n_j \sum_{i=1}^{N_j} \xi_{j,i} \ddot{l}(\hat{\boldsymbol{\theta}}; D_{j,i}) \right\}^{-1}, \tag{2.9}$$

$$(2.10) \quad \text{var}_{\text{ph2}}(\hat{\theta}) = \frac{1}{N^2} \hat{I}(\theta_0)^{-1} \left[ \sum_{j=1}^J N_j \left( \frac{1-p_j}{p_j} \right) \text{var}_{0|j} \{ \dot{l}(\theta_0) \} \right] \hat{I}(\theta_0)^{-1},$$

where  $\text{var}_{0|j} = (n_j)^{-1} \sum_{i=1}^{N_j} \{ \xi_{j,i} \dot{l}(\hat{\theta}; D_{j,i}) \}^{\otimes 2} - \{ (n_j)^{-1} \sum_{i=1}^{N_j} \xi_{j,i} \dot{l}(\hat{\theta}; D_{j,i}) \}^{\otimes 2}$ . As a result, the variance estimator of  $\hat{\theta}$  is the sum of the variances in (2.9) and (2.10). Given  $\mathbf{x}$  and  $\mathbf{z}$ , the asymptotic variance estimate for  $\text{CR}(t \mid \mathbf{x}, \mathbf{z}, \hat{\theta})$  is derived by  $\text{var}(\hat{\theta})$  and the delta method. The explicit variance form is presented in Section 4 in the supplementary materials [Hyun et al. (2017)].

The sampling weights can be estimated to improve efficiency by using a parametric model  $\pi(\alpha; \mathbf{v}) = \text{Pr}(\xi_{j,i} = 1 \mid \mathbf{v}_{j,i})$  when the auxiliary variables are closely correlated with the target variables [Breslow et al. (2009)]. When we use estimated weights, the asymptotic distribution of the estimates is different from distribution (2.7), particularly from the variance due to sampling,  $\Sigma$  in (2.8). The asymptotic distribution can be derived by the result of Breslow et al. (2009), and it is presented in Section 3 in the supplementary materials [Hyun et al. (2017)].

**2.5. Semiparametric estimation procedure.** A semiparametric risk estimate is useful for checking the fit of parametric models. We propose a semiparametric estimator that maximizes the objective function in (2.4) by iterating between estimating the finite dimensional regression parameters and the infinite dimensional cumulative-hazard  $\Lambda(t)$ , estimating each with standard fitting algorithms:

1. Initialize  $\hat{\beta}^{(0)} = \beta^*$  and  $\hat{\gamma}^{(0)} = \gamma^*$ .
2. With the current estimate  $(\hat{\beta}^{(l)}, \hat{\gamma}^{(l)})$ , compute  $\hat{\Lambda}^{(l)}$  by maximizing  $l_n^\pi(\hat{\beta}^{(l)}, \hat{\gamma}^{(l)}, \Lambda; \mathbf{D})$  as a function of  $\Lambda$ . This optimization can be carried out by the Iterative Convex Minorant (ICM) algorithm [Robertson, Wright and Dykstra (1988)] (the detail follows below).
3. With the updated  $\hat{\Lambda}^{(l)}$ , we maximize  $l_n^\pi(\beta, \gamma, \hat{\Lambda}^{(l)}; \mathbf{D})$  with respect to  $(\beta^T, \gamma^T)$  using the classic iteratively reweighted least squares algorithm for generalized linear models [Nelder and Wedderburn (1972)].
4. Repeat steps 2 and 3 until convergence.

For steps 2 and 3, we define the following IPW processes:

$$\begin{aligned} A_{j,i}(t) &= I(m_{j,i} = 1 \text{ and } 0 < R_{j,i} \leq t)(1 - y_{j,i}) \frac{g(R_{j,i}) \exp(\mathbf{z}_{j,i} \boldsymbol{\gamma})}{\{g(L_{j,i}) - g(R_{j,i})\}} \\ &\quad - I(m_{j,i} = 1 \text{ and } 0 < L_{j,i} \leq t)(1 - y_{j,i}) g(L_{j,i}) \\ &\quad \times \exp(\mathbf{z}_{j,i} \boldsymbol{\gamma}) / \{g(L_{j,i}) - g(R_{j,i})\} \\ &\quad + I(m_{j,i} = 0 \text{ and } 0 < R_{j,i} \leq t) g(R_{j,i}) \\ &\quad \times \exp(\mathbf{z}_{j,i} \boldsymbol{\gamma}) / \{1 + \exp(\mathbf{x}_{j,i} \boldsymbol{\beta}) - g(R)\}, \end{aligned}$$

where  $g(t) = \exp\{-\Lambda(t) \exp(z\boldsymbol{\gamma})\}$  for  $t > 0$ ,  $g(0) = 1$  and  $\lim_{t \rightarrow \infty} g(t) = 0$ . This process  $A_{j,i}(t)$  is the time derivative of the log-likelihood in (2.5) and can only have a jump at  $t_k$ , which is at either  $L_{j,i}$  or  $R_{j,i}$ :

$$\begin{aligned}
 A_{\Lambda,n}(t) &= \sum_{j=1}^J N_j/n_j \sum_{i=1}^{N_j} \xi_{j,i} A_{j,i}(t), \\
 G_{\Lambda,n}(t) &= \sum_{j=1}^J N_j/n_j \sum_{i=1}^{N_j} \xi_{j,i} A_{j,i}^2(t), \\
 Q_{\Lambda,n}(t) &= A_{\Lambda,n}(t) + \int_0^t \Lambda(s) dG_{\Lambda,n}(s),
 \end{aligned}
 \tag{2.11}$$

where  $G_{\Lambda,n}(t)$  in (2.11) is based on a second order expansion of the log-likelihood in (2.5). To ensure identifiability of  $\Lambda(t)$ , we assume that  $\hat{\Lambda}$  is right continuous and piecewise constant, and at most only discontinuous at  $\{t_{(k)}; k = 1, \dots, K\}$ , which are ordered unique values of observed times,  $\{L_i, R_i \mid L_i \neq 0 \text{ and } R_i < \infty, i = 1, \dots, n\}$ .

For fixed  $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ , let  $\hat{\Lambda}$  be the left derivative of the greatest convex minorant of the self-induced cumulative sum diagram formed by the points,  $(0, 0)$  and  $\{G_{\hat{\Lambda},n}(t_{(k)}), Q_{\hat{\Lambda},n}(t_{(k)})\}$ . Then  $\hat{\Lambda}$  maximizes  $\sum_{j=1}^J N_j/n_j \sum_{i=1}^{N_j} \xi_{j,i} l(\boldsymbol{\beta}, \boldsymbol{\gamma}, \Lambda; D_{j,i})$  [Groeneboom and Wellner (1992)]. The consistency of the estimators obtained from the semiparametric procedure is proved in Theorem 3.1 in Section 3 of the supplementary materials [Hyun et al. (2017)].

**3. Simulation studies.** We conducted a series of simulations to assess the numerical performance for the weakly-parametric IPW logistic-Cox model and to compare estimates from it to the semiparametric IPW logistic-Cox model. We simulate two scenarios SC1 and SC2, where SC1 reflects an ideal situation with a high event rate and narrow visit-intervals, whereas SC2 reflects a realistic scenario with a moderate event rate and wide visit-intervals. Two covariates in the models (3.1) and (3.2),  $X_1$  and  $X_2$  are independently generated as a binomial with probability 0.5 and as a standard normal distribution with variance 1, respectively, and the covariate vectors for incidence and prevalence are identical:

$$\begin{aligned}
 (3.1) \quad \text{Logistic model:} \quad & \text{logit}\{P_d(X_1, X_2, \boldsymbol{\beta})\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \\
 (3.2) \quad \text{Cox model:} \quad & \Lambda(t; X_1, X_2, \boldsymbol{\gamma}) = \gamma_0 t^\tau \exp(\gamma_1 X_1 + \gamma_2 X_2).
 \end{aligned}$$

The Cox submodel baseline hazard parameters are  $(\gamma_0, \tau) = (0.135, 1)$  for SC1 and  $(0.05, 0.5)$  for SC2; the covariates-related parameters are  $(\beta_1, \beta_2, \gamma_1, \gamma_2) = (1, 1, 0.3, 0.3)$  for SC1 and 2. Visit times are independent and generated as a normal distribution with mean 3 and variance 0.5. The number of visits varies across subjects as we set a fixed end time ( $t = 20$  for SC1 and  $t = 10$  for SC2) for follow-up. Follow-up occurs if there is no prevalent disease at baseline. Whether a subject

takes a diagnostic test at each screening visit follows a binomial distribution with the probability of 0.5 and 0.07 for SC1 and 2, respectively. This means the incidental interval in SC1 is more likely to be narrower than the one in SC2.

Time interval  $(L_i, R_i)$  in which disease occurs is determined by the closest disease ascertainment date prior to and post to the true event time. We set the cohort size to be 10,000, and consider two-phase stratified sample. For stratification, we use two factors, cases-controls in certain enrollment period and a binary variable,  $V$  depending on  $X_1 + X_2$ . Among the high risk group, that is,  $X_1 + X_2 \geq Q$  (92.8%), where  $Q$  (92.8%) corresponds to the 92.8% quantile of the distribution of  $X_1 + X_2$ , namely, 2.135, we set  $P(V = 1 | X_1 + X_2 \geq 2.135) = 0.9$ , and of the low risk group, we set  $P(V = 0 | X_1 + X_2 < 2.135) = 0.9$ . This implies that the stratum variable  $V$  is strongly associated with survival time  $T$ . In SC1, cases are defined by diagnosis time up to  $t = 2$ , that is, prevalent case or  $T_i < 2$ ; whereas in SC2, cases are defined by prevalent cases only. We take all cases and randomly select samples from  $(V = 1, \text{controls})$  and  $(V = 0, \text{controls})$ , 80% and 11% for SC1 and 80% and 20% for SC2, respectively. The sampling weights for cases and controls are one and the inverse of the sampling fraction, (1.25 and 9.09) for SC1 and (1.25 and 5.0) for SC2, respectively. SC2 is meant to simulate the data of our application, while SC1 increases the number of incidental intervals.

In SC1, the average sample size is 2611. The baseline diagnosis test rate and left-/right-/interval-censoring rates are 95.5%, 30.6%, 3.9%, and 41.1%, respectively. In SC2, the average sample size is 3354. The baseline diagnosis test rate and left-/right-/interval-censoring rates are 95.5%, 12.0%, 59.0%, and 0.9%, respectively. We carried out 1000 replications for each scenario.

We first applied a naive approach, a survey-weighted Cox model for right-censored data to the simulation data by using function “svycoxph” in the R-package “Survey” [Lumley (2016)]. We imputed the minimum of  $\{L_{j,i}; i = 1, \dots, n_j, j = 1, \dots, J\} - \varepsilon$  to the event time for prevalent cases, where  $\varepsilon$  is an arbitrary positive constant so that the event time is positive; we impute  $(R_{j,i} - L_{j,i})/2$  to the event times for  $[L_{j,i} = 0, R_{j,i})$  or  $(L_{j,i}, R_{j,i})$  for  $R_{j,i} < \infty$ ; the censoring times for  $(0 < L_{j,i}, R_{j,i} = \infty)$  are imputed to  $L_{j,i}$ . In SC1, the cumulative risk estimates are substantially biased at the early times, and the bias is decreasing to 0 over time, whereas the cumulative risk estimates in SC2 are substantially biased across times because of the wide finite visit-intervals and the low event rate (Table 1).

Table 2 presents simulation results. In both scenarios, regression parameter and cumulative risk estimates have negligible bias. For the regression parameter estimates, the efficiency of both models are comparable, whereas for cumulative risk estimates, the empirical standard errors of the weakly-parametric model are smaller (relative efficiency is 1.297–1.664) than those of the semiparametric model. The resulting asymptotic variance estimates are close to the empirical standard errors except for the intercept coefficient parameter in SC2. Most coverage probabilities from the weakly-parametric model are near the nominal level 95%.

TABLE 1  
Simulation results of a naive approach: the cumulative risk estimates are for the subgroup with  $(x_1 = 1, x_2 = 0.5)$

Cumulative risk (CR)	Scenario 1 ( $n = 2611$ )		Scenario 2 ( $n = 3354$ )	
	True value	Bias	True value	Bias
CR( $t = 0.1$ )	0.138	0.135	0.141	0.185
CR( $t = 1$ )	0.288	0.058	0.185	0.146
CR( $t = 3$ )	0.534	-0.016	0.231	0.115
CR( $t = 5$ )	0.695	-0.013	0.260	0.223
CR( $t = 7$ )	0.801	-0.004	0.284	0.208

TABLE 2  
Simulation results: <sup>a</sup> parameter; <sup>b</sup> true value; <sup>c</sup> empirical standard error; <sup>d</sup> relative efficiency =  $SE_1/SE_2$ ; <sup>e</sup> asymptotic standard error; <sup>f</sup> 95% coverage probability; <sup>g</sup> cumulative risk; the cumulative risk estimates are for the subgroup with  $(x_1 = 1, x_2 = 0.5)$

Para. <sup>a</sup>	True. <sup>b</sup>	Semiparametric model			Weakly-parametric model			
		Bias	SE <sub>1</sub> <sup>c</sup>	RE <sup>d</sup>	Bias	SE <sub>2</sub> <sup>c</sup>	ASE <sup>e</sup>	CP <sup>f</sup>
Scenario 1, $n = 2611$								
$\beta_0$	-3.5	-0.007	0.100	1.004	-0.004	0.099	0.105	0.952
$\beta_1$	1.0	0.003	0.119	1.002	0.001	0.119	0.120	0.950
$\beta_2$	1.0	0.004	0.058	1.005	0.003	0.058	0.057	0.947
$\gamma_1$	0.3	-0.011	0.074	1.085	-0.004	0.068	0.065	0.938
$\gamma_2$	0.3	0.002	0.037	1.121	0.000	0.033	0.033	0.952
Prevalence	0.119	0.000	0.007	1.002	0.000	0.007	0.007	0.960
CR( $t = 1$ ) <sup>g</sup>	0.288	-0.002	0.034	1.652	0.000	0.021	0.023	0.954
CR( $t = 3$ )	0.534	-0.011	0.044	1.664	-0.001	0.026	0.026	0.939
CR( $t = 5$ )	0.695	0.004	0.032	1.390	-0.001	0.023	0.022	0.931
CR( $t = 7$ )	0.801	0.000	0.023	1.287	-0.001	0.018	0.018	0.947
CR( $t = 10$ )	0.895	-0.001	0.017	1.284	-0.001	0.014	0.013	0.946
Scenario 2, $n = 3354$								
$\beta_0$	-3.5	-0.002	0.088	1.001	-0.002	0.088	0.092	0.957
$\beta_1$	1.0	0.001	0.102	1.000	0.000	0.102	0.103	0.949
$\beta_2$	1.0	0.002	0.051	0.999	0.001	0.051	0.051	0.943
$\gamma_1$	0.3	-0.020	0.286	2.063	-0.001	0.139	0.138	0.950
$\gamma_2$	0.3	0.000	0.067	1.014	0.002	0.066	0.067	0.950
Prevalence	0.119	0.000	0.006	1.000	0.000	0.006	0.006	0.964
CR( $t = 1$ )	0.185	-0.009	0.029	1.297	0.000	0.022	0.022	0.919
CR( $t = 3$ )	0.231	-0.007	0.029	1.284	-0.002	0.022	0.023	0.925
CR( $t = 5$ )	0.260	0.003	0.027	1.266	-0.001	0.022	0.021	0.931
CR( $t = 7$ )	0.284	0.004	0.026	1.301	0.001	0.020	0.019	0.931

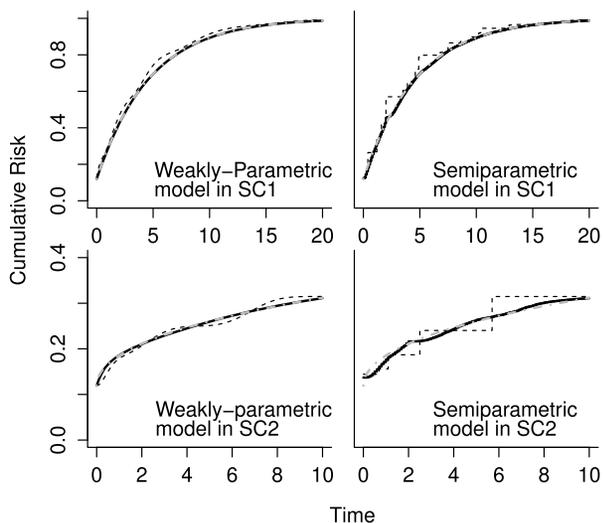


FIG. 2. Results of a simulation study with two scenarios, SC1 and SC2: the black solid lines are the average of the 1000 estimates; the black dashed lines are a single representative estimate from the 1000 estimates; the grey dashed-dot lines are the true cumulative risk.

In SC2, the relatively low coverage probability for the cumulative risk is owing to the lack of events, and consequently the few simulation estimates with relatively large bias. Results for the cumulative risk curve estimated in both scenarios are shown in Figure 2, and the bias of cumulative risk curve estimated in SC1 is much smaller than the curves estimated in SC2. The black-solid lines are the average of the 1000 estimates, and the average estimates agree well with the true curve of grey dashed-dot line. The dashed step-curve and dashed smooth-curve in Figure 2 are a single representative estimate from the 1000 estimates, and those are also close to the true curve.

We numerically evaluated the robustness of the cumulative risk estimates from the weakly-parametric and semiparametric logistic-Cox model when the true prevalence and incidence models are a probit and an additive hazard model. The cumulative risk estimates from the semiparametric and weakly-parametric logistic-Cox regression models are robust to model misspecification (Table 1 and Figure 1 in the supplementary materials [Hyun et al. (2017)]). We also evaluated the robustness of the cumulative risk estimates from the weakly-parametric logistic-Cox model when the assumptions about the cubic B-spline approximation are violated. As a violation, we considered a cumulative hazard function including abrupt change points. In the scenario with a high event rate, the cumulative risk estimate from the semiparametric model is less biased than the weakly-parametric model. However, in the scenario with a moderate event rate, the cumulative risk estimate from the weakly-parametric model is less biased than the semiparametric model (Table 2 and Figure 2 in the supplementary materials [Hyun et al. (2017)]).

**4. Application: Developing risk-bands based on HPV typing tests.** It is expected that the next cervical cancer screening guidelines will include recommendations for the use of HPV typing tests. There are thirteen oncogenic HPV types and one possibly oncogenic type commonly included in tests (HPV66), and each type has a different risk of precancer/cancer [Schiffman et al. (2011)]. However, little is known about the performance of HPV typing in clinical practice, and the best grouping of the 14 types for different triage would be useful to increase the screening benefit. Our typing assay currently groups the 14 types into 9 categories: HPV16, HPV18, HPV31, HPV45, HPV51, HPV52, HPV33/58, HPV39/68/35, and HPV59/56/66.

For the subgroup with positive on HC2 (5%) within the cohort of women undergoing screening at KPNC, we have assembled a two-phase stratified sample of 9258 (in Figure 1) with HC2-positive. From the sample, we have residual discarded HPV test specimens usable for HPV-type testing since 2007. The stratified sample was based on baseline cytology severity (normal/low/high grade), FocalPoint computer-assisted quantitative cytology (0, 1–9, 10–100%), and baseline histology result (grade 1/2/3 or cancer). Table 3 shows the sample design. The analysis dataset includes 8333 subjects with complete HPV types. Median and maximum follow-up time are 1.69 and 7.18 years, respectively. The outcome of interest is precancer (histology grade 3) or cancer. There are 744 (8.9%) prevalent cases at baseline, and baseline biopsy rate, left-/right-/interval-censored cases are 7331 (88.0%), 361 (4.7%), 7132 (94.0%), and 96 (1.1%), respectively. The 1888 (24.9%) who never got a biopsy are mostly women who have less than 1 year of follow-up or their HPV cleared at their second visit, obviating a biopsy.

We used the IPW logistic-Cox model to calculate 3-year risk of precancer or cancer for each HPV type, with the very lowest risk types grouped (Table 4). Because multiple HPV types can co-infect the cervix, the analysis is hierarchically conducted in the following manner. We calculate the marginal risk for each type, then at the next level, we excluded everyone who had all higher-risk HPV type, and recalculate marginal risks for the remaining types [Schiffman et al. (2015)], and so on. This determines the best order of introducing additional type categories for risk stratification. This strategy is sensible, in that precancer/cancer risk is dominated by the riskiest type, that is, multiple types do not “interact” [Chaturvedi et al. (2011)]. For example, a woman with both HPV16 and HPV56 will have her outcomes attributed to the higher risk type (i.e., HPV16). When estimating risk for subsets of data, a standard weighted analysis using only the subset of interest can underestimate standard errors if there is no sampled observations from the domain in some strata [Graubard and Korn (1996)], but in the hierarchical subgroup by HPV types, each domain is sampled from nearly all strata. We did not employ a multiple comparisons correction because the hierarchical analyses were done for exploratory purposes.

The estimates are obtained by applying the weakly-parametric logistic-Cox model with a covariate for HPV type in each submodel for prevalence and incidence. We chose the cubic B-spline with 7 knots placed at quantiles of visit times

TABLE 3

Sample design in women with HPV positive: for FocalPoint, “0” means result absent, “1–9” means not within most abnormal decile, “≥10” means within most abnormal decile

Severity of cytology	Histology	FocalPoint category (%)	Stratum number	Sample number	Sampling fraction	Sampling weight
Normal or low grade	<Grade 2	0	13,615	1651	0.1213	8.2
		1–9	13,826	2441	0.1766	5.7
		≥10	2845	1412	0.4963	2.0
	Grade 2	0	918	321	0.3497	2.9
		1–9	808	286	0.354	2.8
		≥10	360	184	0.5111	2.0
	Grade 3	0	541	249	0.4603	2.2
		1–9	427	185	0.4333	2.3
		≥10	184	100	0.5435	1.8
Cancer/AIS	0	82	54	0.6585	1.5	
	1–9	71	57	0.8028	1.2	
	≥10	17	12	0.7059	1.4	
High grade	<Grade 2	0	497	332	0.668	1.5
		1–9	268	215	0.8022	1.2
		≥10	251	170	0.6773	1.5
	Grade 2	0	214	169	0.7897	1.3
		1–9	107	69	0.6449	1.6
		≥10	175	116	0.6629	1.5
	Grade 3	0	251	222	0.8845	1.1
		1–9	131	80	0.6107	1.6
		≥10	299	189	0.6321	1.6
	Cancer/AIS	0	88	62	0.7045	1.4
		1–9	61	42	0.6885	1.5
		≥10	69	46	0.6667	1.5

by examining the semiparametric risk estimate. For each of the nine categories, the cumulative risk curves from the weakly-parametric approach is a good fit to the semiparametric estimates (Figure 3).

The types can be grouped into 4 bands. As expected, HPV16 had by far the greatest risk (21.9%), nearly 15 times the 1.5% risk associated with the lowest-risk types (HPV59/56/66). HPV18 has the second highest risk at 11.5%. Although HPV45 has half the risk of HPV18, they both cause a particularly worrisome subtype of cervical cancer (adenocarcinoma) so we group 18/45 together. Because types 33/58/31/52 have moderate risks between 5.6% and 8.6%, we group them together. The types 51/39/68/35/59/56/66 are grouped together because all have risk below 2.9%.

TABLE 4

*Hierarchical analysis for CIN3 plus risk by the nine HPV categories: <sup>a</sup>number of observations; <sup>b</sup>3 years-cumulative risk; <sup>c</sup>lower limit; <sup>d</sup>upper limit*

HPV	No. obs <sup>a</sup>	3yr-CR (%) <sup>b</sup>	95%LL <sup>c</sup>	95%UL <sup>d</sup>
HPV16 positive	1564	21.9	20.1	23.7
Else HPV18 positive	494	11.5	9.2	13.8
Else HPV33 or 58 positive	631	8.6	6.9	10.3
Else HPV31 positive	766	8.1	6.6	9.5
Else HPV45 positive	324	5.4	3.8	7.0
Else HPV52 positive	823	5.6	4.4	6.7
Else HPV51 positive	536	2.9	1.9	3.9
Else HPV39, 68 or 35 positive	1201	2.0	1.5	2.5
Else HPV59, 56 or 66 positive	1047	1.5	1.0	1.9

To form cervical cancer risk strata combining HPV with cytology, we calculated 3-year risk for grade 3 or cancer/AIS by histology (called CIN3+) across cytology subgroups within each band in Table 5. By comparison with established risk benchmarks and management recommendations from current U.S. guidelines [Katki et al. (2011)], we are able to propose the risk management of each stratum.

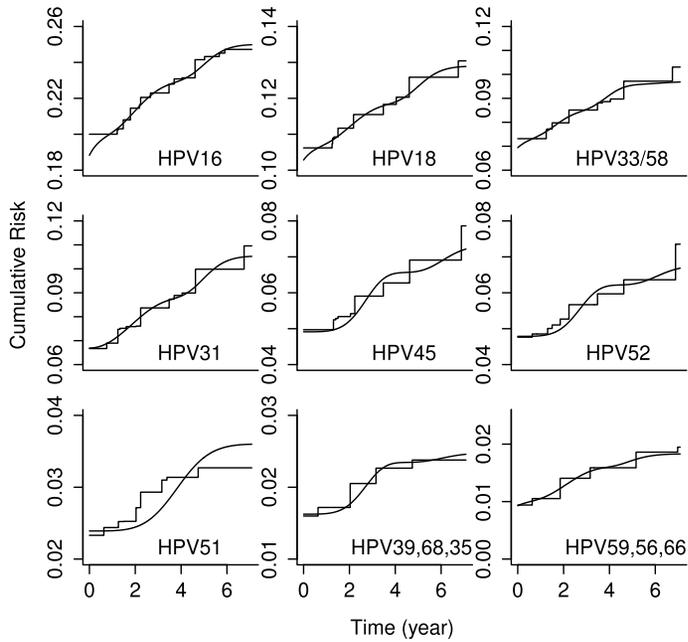


FIG. 3. *CIN3 plus cumulative risk estimates by the HPV types. the step curves are semiparametric estimates; the smooth curves are weakly-parametric estimates.*

TABLE 5

*CIN3/cancer risk strata combining the four HPV bands with cytology: <sup>a</sup> number of observations; <sup>b</sup> 3 years-cumulative risk; <sup>c</sup> 95% lower limit; <sup>d</sup> 95% upper limit*

HPV positive	Cytology	Obs. <sup>a</sup>	CR (%) <sup>b</sup>	LL <sup>c</sup>	UL <sup>d</sup>
HPV16	Overall	1564	21.9	20.1	23.7
	High	506	60.6	59.5	61.6
	Medium/low	574	17.9	16.2	19.5
	Normal	484	13.8	11.9	15.6
Else HPV 18/45	Overall	843	9.0	7.5	10.5
	High	235	40.9	39.7	42.1
	Medium/low	250	7.1	5.6	8.6
	Normal	358	4.4	3.2	5.6
Else HPV31/52/33/58	Overall	2195	7.3	6.5	8.1
	High	467	35.0	34.2	35.7
	Medium/low	850	5.7	5.0	6.4
	Normal	878	4.0	3.2	4.7
Else HPV 51/39/68/35/59/56/66	Overall	2784	2.0	1.7	2.3
	High	321	13.6	11.8	15.3
	Medium/low	1123	2.0	1.7	2.4
	Normal	1340	1.2	0.0	8.7

Risk varies from 60.6% for HPV16 and high risk cytology down to 1.2% for the 4<sup>th</sup> HPV band and normal cytology, which represents considerable risk stratification. These risk bands could be used to base future guidelines, for example, the highest risks might indicate immediate treatment, medium-high risk might indicate a biopsy, medium-low risk might indicate a 1-year return, and low-risk might indicate a 2-year return.

Cumulative risk was used to inform the screening guidelines process because it was simpler to use than separate risks of prevalent and incident disease [Katki et al. (2013)]. However, risks of prevalent versus incidence disease are separated by the model and could be used separately if so desired.

**5. Discussion.** Although potentially cost-effective and efficient, cohorts assembled from electronic health records at health providers pose analytic challenges. We addressed three challenges: prevalent left-censored outcomes and incident irregularly interval-censored outcomes, where incident disease is a mixture of truly incident disease and missed-prevalent disease when disease ascertainment is not always conducted at the baseline visit. The third challenge is complex sampling within the cohort, such as two-phase stratified case-control sampling, to ensure efficient use of biospecimen resources.

The estimates from an weighted Cox hazard model, but with ad hoc schemes to impute event onsets within intervals, are biased (Section 3). We proposed a

general family of mixture models called prevalence-incidence models and focused on the logistic-Cox model in order to estimate cumulative risk. We proposed a weighted likelihood approach, using IPW to account for different complex two-phase sampling rates. We presented a weakly-parametric model using monotone splines, whose goodness-of-fit can be checked against a semiparametric risk curve estimated by an iterative algorithm that includes a weighted-iterative convex minorant algorithm. Our approach is the obverse of the cure model for two heterogeneous subpopulations; cure models have a point mass at infinity, but prevalence-incidence models have a point mass at the origin. Cure models have identifiability problems because cure can never be observed. In contrast, prevalent disease is observable for some patients, which should mitigate identifiability issues with prevalence-incidence models. We applied the IPW logistic-Cox model to estimate risk to group the 14 HPV types into 4 risk-bands. These risk-bands may be adopted by commercial entities proposing new HPV typing tests for regulatory approval and for adoption into future cervical cancer screening guidelines.

In our example, we focused on total cervical precancer/cancer risk for HPV-positive women, which combines risks of both prevalent and incident disease. However, for other aims, one may focus on only prevalent disease risk or incident disease risk. For example, only incident disease risk is relevant for women who undergo definitive disease ascertainment and are known disease-free. In contrast, ideally only prevalent disease risk is relevant for making decisions about whether to undergo definitive disease ascertainment, such as biopsies. Our models yield proper estimates of incidence disease risk using all the data, which improves power and reduces selection bias.

Although the weakly-parametric model is flexible, it still requires assumptions. From simulation studies, we found the assumptions for the weakly-parametric model are plausible in practice, and the weakly-parametric model can sometimes have less finite-sample bias than the semiparametric model for low/moderate event rates. However, bias can dominate in a large data with many events, and weakly parametric models are more likely to have larger bias and smaller variance than semiparametric models when the assumptions are violated. To identify such situations, it is important to check whether the confidence interval from the weakly-parametric model includes the point estimate of the semiparametric model.

We linked electronic records to assemble a high-risk 5% sub-cohort of women undergoing cervical cancer screening at KPNC, and conducted HPV typing tests on a stratified two-phase sample of 8644 women. The risk curves from the weakly-parametric IPW logistic-Cox model fit well to the semiparametric curves. Because having separate guidelines for each of 14 types is too complex for clinicians, we grouped the types into 4 bands by risk: HPV16 had a uniquely high risk of precancer/cancer; HPV18/45 and HPV31/52/33/58 have intermediate risk, and HPV51/39/68/35/59/56/66 has low risk. The most common abnormality in

screening is HC2-positive and a normal cytology, for which guidelines currently recommend that patients return after 1-year. For HPV-positive women with normal cytology, if she has HPV16, her risk might be high enough to justify immediate biopsies, but if her HPV type is in the 4th (low risk) band, her risk might be low enough to justify a 2- or 3-year return. Our findings suggest that HPV typing, in conjunction with cytology test, might more precisely define management based on risk. These risk bands could be used to base future guidelines: high, medium, and low risk might indicate a biopsy, 1-year return, and 2-year return, respectively.

Our prevalence-incidence models are an incremental step on the way to developing more sophisticated models. Our models presume only progressive disease, but it is believed that some cervical precancers can spontaneously regress to normalcy without intervention. Regressive outcomes present serious identifiability problems for interval-censoring methods. Also, our model presumes a perfect outcome ascertainment, but biopsies are considered insensitive for finding cervical precancers [Schiffman et al. (2011)]. The combination of outcome measurement error and regressive outcomes present serious identifiability problems for any stochastic model, but must be addressed to develop more realistic and useful models. Cervical precancer is not deadly, so survival bias in sampling is negligible; however, if the interest was to study the natural history of cervical precancer and cancer (rather than to simply develop risk estimates valid for clinical use), we would need to account for left-truncation. Finally, we calculated risks valid only for baseline time-independent covariates, such as a baseline HPV test result. Extending the models to account for internal time-dependent covariates, such as HPV status changing over time, is an area of future work.

The semiparametric IPW logistic-Cox model is computationally intensive. Reducing the computational burden will be critical for epidemiologists who generally use only their desktop computers and are used to seeing results in a short period of time. An R package, (PIMixture) is under development to fit the IPW logistic-Cox model.

**Acknowledgments.** The first author would like to thank Dr. Barry I. Graubard for his constructive suggestions during the writing of this paper. The authors are grateful to two referees, an associate editor, and the editor for helpful comments.

#### SUPPLEMENTARY MATERIAL

**Supplement to “Flexible risk prediction models for left or interval-censored data from electronic health records”** (DOI: [10.1214/17-AOAS1036SUPP](https://doi.org/10.1214/17-AOAS1036SUPP); .pdf). Supplementary materials available in the attached file includes the proofs for

model identifiability and to establish useful asymptotic results of the estimates such as consistency and weak convergence to normal distribution under certain regularity conditions. The simulation studies and results are summarized in the supplementary materials.

## REFERENCES

- BRESLOW, N. E. and WELLNER, J. A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scand. J. Stat.* **34** 86–102. [MR2325244](#)
- BRESLOW, N. E., LUMLEY, T., BALLANTYNE, C. M., CHAMBLESS, L. E. and KULICH, M. (2009). Improved Horvitz-Thompson estimation of model parameter from two-phase stratified samples: Applications in epidemiology. *Stat. Biosci.* **1** 32–49.
- CAI, T. and ZHENG, Y. (2013). Resampling procedures for making inference under nested case-control studies. *J. Amer. Statist. Assoc.* **108** 1532–1544. [MR3174727](#)
- CASTLE, P. E., FETTERMAN, B., SCT (ASCP), POITRAS, N., LOREY, T., SHABER, R. and KINNEY, W. (2009). Five-year experience of human papillomavirus DNA and Papanicolaou test cotesting. *Obstetrics & Gynecology* **113** 595–600.
- CASTLE, P. E., STOLER, M. H., WRIGHT, JR., T. C., SHARMA, A., WRIGHT, T. L. and BEHRENS, C. M. (2011). Performance of carcinogenic human papillomavirus (HPV) testing and HPV16 or HPV18 genotyping for cervical cancer screening of women aged 25 years and older: A subanalysis of the ATHENA study. *Lancet Oncol.* **12** 880–890.
- CHATURVEDI, A. K., KATKI, H. A., HILDESHEIM, A., RODRÍGUEZ, A. C., QUINT, W., SCHIFFMAN, M., VAN DOORN, L. J., PORRAS, C., WACHOLDER, S., GONZALEZ, P. and SHERMAN, M. E. (2011). Human papillomavirus infection with multiple types: Pattern of coinfection and risk of cervical disease. *J. Infect. Dis.* **203** 910–920.
- COX, D. R. (1972). Regression models and life-tables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **34** 187–220. [MR0341758](#)
- DOREY, F. J., LITTLE, R. J. A. and SCHENKER, N. (1993). Multiple imputation for threshold-crossing data with interval censoring. *Stat. Med.* **12** 1589–1603.
- GRAUBARD, B. I. and KORN, E. L. (1996). Survey inference for subpopulations. *Am. J. Epidemiol.* **144** 102–106.
- GROENEBOOM, P. and WELLNER, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. *DMV Seminar* **19**. Birkhäuser, Basel. [MR1180321](#)
- HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685. [MR0053460](#)
- HUANG, J. and ROSSINI, A. J. (1997). Sieve estimation for the proportional-odds failure-time regression model with interval censoring. *J. Amer. Statist. Assoc.* **92** 960–967. [MR1482126](#)
- HUANG, J. and WELLNER, J. A. (1997). Interval censored survival data: A review of recent progress. In *Proceedings of the First Seattle Symposium in Biostatistics* (D. Y. Lin and T. R. Fleming, eds.) 123–169. Springer, New York.
- HYUN, N., CHEUNG, L. C., PAN, Q., SCHIFFMAN, M. and KATKI, H. A. (2017). Supplement to “Flexible risk prediction models for left or interval-censored data from electronic health records.” DOI:10.1214/17-AOAS1036SUPP.
- KAPLAN, E. L. and MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53** 457–481. [MR0093867](#)
- KATKI, H. A., KINNEY, W. K., FETTERMAN, B., LOREY, T., POITRAS, N. E., CHEUNG, L., DEMUTH, F., SCHIFFMAN, M., WACHOLDER, S. and CASTLE, P. E. (2011). Cervical cancer risk for 330,000 women undergoing concurrent HPV testing and cervical cytology in routine clinical practice at a large managed care organization. *Lancet Oncol.* **12** 663–672.

- KATKI, H. A., SCHIFFMAN, M., CASTLE, P. E., FETTERMAN, B., POITRAS, N. E., LOREY, T., CHEUNG, L. C., RAINE-BENNETT, T. R., GAGE, J. C. and KINNEY, W. K. (2013). Benchmarking CIN3+ risk as the basis for incorporating HPV and Pap cotesting into cervical screening and management guidelines. *J. Low. Genit. Tract Dis.* **17** S28–S35.
- KOVALCHIK, S. A. and PFEIFFER, R. M. (2014). Population-based absolute risk estimation with survey data. *Lifetime Data Anal.* **20** 252–275. [MR3181014](#)
- LI, C.-S., TAYLOR, J. M. G. and SY, J. P. (2001). Identifiability of cure models. *Statist. Probab. Lett.* **54** 389–395. [MR1861384](#)
- LUMLEY, T. (2016). Analyses of complex survey samples. Available at <https://cran.r-project.org/web/packages/survey/survey.pdf>.
- MA, S. (2010). Mixed case interval censored data with a cured subgroup. *Statist. Sinica* **20** 1165–1181. [MR2730178](#)
- MASSAD, L. S., EINSTEIN, M. H., HUH, W. K., KATKI, H. A., KINNEY, W. K., SCHIFFMAN, M., SOLOMON, D., WENTZENSEN, N. and LAWSON, H. W. (2013). 2012 updated consensus guidelines for the management of abnormal cervical cancer screening tests and cancer precursors. *J. Low. Genit. Tract Dis.* **17** S1–S27.
- NELDER, J. A. and WEDDERBURN, R. W. M. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A* **135** 370–384.
- ODELL, P. M., ANDERSON, K. M. and D'AGOSTINO, R. B. (1992). Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics* 951–959.
- ROBERTSON, T., WRIGHT, F. T. and DYKSTRA, R. L. (1988). *Order Restricted Statistical Inference*. Wiley, Chichester. [MR0961262](#)
- RÜCKER, R. and MESSERER, D. (1988). Remission duration: An example of interval-censored observations. *Stat. Med.* **7** 1139–1145.
- SAEGUSA, T. (2015). Variance estimation under two-phase sampling. *Scand. J. Stat.* **42** 1078–1091. [MR3426311](#)
- SCHIFFMAN, M., WENTZENSEN, N., WACHOLDER, S., WALTER, K., GAGE, J. C. and CASTLE, P. E. (2011). Human papillomavirus testing in the prevention of cervical cancer. *J. Natl. Cancer Inst.* **103** 368–383.
- SCHIFFMAN, M., VAUGHAN, L. M., RAINE-BENNETT, T. R., CASTLE, P. E., KATKI, H. A., GAGE, J. C., FETTERMAN, B., BEFANO, B. and WENTZENSEN, N. (2015). A study of HPV typing for the management of HPV-positive ASC-US cervical cytologic results. *Gynecol. Oncol.* **138** 573–578.
- SEN, B. and BANERJEE, M. (2007). A pseudolikelihood method for analyzing interval censored data. *Biometrika* **94** 71–86. [MR2307901](#)
- SHAO, F., LI, J., MA, S. and LEE, M.-L. T. (2014). Semiparametric varying-coefficient model for interval censored data with a cured proportion. *Stat. Med.* **33** 1700–1712. [MR3246689](#)
- TIAN, L. and CAI, T. (2006). On the accelerated failure time model for current status and interval censored data. *Biometrika* **93** 329–342. [MR2278087](#)
- WANG, L., MCMAHAN, C. S., HUDGENS, M. G. and QURESHI, Z. P. (2016). A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data. *Biometrics* **72** 222–231. [MR3500591](#)
- WOODWARD, M. (1999). *Epidemiology: Study Design and Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL. [MR1696292](#)
- ZHANG, Y., HUA, L. and HUANG, J. (2010). A spline-based semiparametric maximum likelihood estimation method for the Cox model with interval-censored data. *Scand. J. Stat.* **37** 338–354. [MR2682304](#)

N. HYUN  
L. C. CHEUNG  
M. SCHIFFMAN  
H. A. KATKI  
DIVISION OF CANCER EPIDEMIOLOGY  
AND GENETICS  
NATIONAL CANCER INSTITUTE  
ROCKVILLE, MARYLAND 20850  
USA  
E-MAIL: [noorie.hyun@nih.gov](mailto:noorie.hyun@nih.gov)

Q. PAN  
DEPARTMENT OF STATISTICS  
GEORGE WASHINGTON UNIVERSITY  
WASHINGTON, DC 20052  
USA