**Appendix**

*Part A: Bias in Kaplan-Meier methods when applied to screening data*

The Kaplan-Meier estimator was developed for estimation of risk of mortality or symptomatic disease.1 The Kaplan-Meier estimator was not intended to estimate unbiased absolute risks for asymptomatic disease2, except in controlled settings where all subjects return at fixed intervals3. When applied to cervical cancer screening data, Kaplan-Meier methods will essentially assume that patients are disease-free until the time of diagnosis. Therefore Kaplan-Meier methods will tend to (1) underestimate risks at early time points, because it assumes that the time of disease detection is the time of disease onset, and (2) overestimate risks at later time points, because relatively more disease is detected later when risk sets are thinner due to censoring2.

Consider the following illustrative example. Suppose 1000 patients who test HPV-positive with NILM Pap at the current screen (t0) have the following numbers with CIN3+:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Time | At timet0 | At time t0 + 1y | At timet0 + 2y | At timet0 + 3y | At timet0 + 4y | At timet0 + 5y |
| Cumulative N with CIN3+ | 20 | 25 | 30 | 35 | 40 | 45 |
| N without CIN3+ | 980 | 975 | 970 | 965 | 960 | 955 |
| Cumulative Risk of CIN3+, % | 2 | 2.5 | 3 | 3.5 | 4 | 4.5 |

If we could correctly ascertain whether all 1000 patients are <CIN3 or CIN3+ at each time point, then we could estimate the cumulative CIN3+ risk as a simple proportion. However, due to study protocols, ascertainment is only possible at visits where patients with a repeat positive test results undergo colposcopy. Thus, none of the 1000 patients undergo colposcopy at the baseline time t0 , so none of the 20 CIN3+ prevalent at baseline are diagnosed at that time. Now suppose that all 1000 patients were asked to return for testing in 1 year but only 50% returned 1 year later; 25% returned in 3 years later, and 25% returned in 5 years later. Each patient has colposcopy only for positive results and has just the 1 return visit before the end of data collection. Kaplan-Meier methods suppose also that the reason for delayed return is independent of their underlying CIN3+ status, that is, patients do not develop symptoms that accelerate their return. Then Kaplan-Meier methods will estimate the following risks based on observed data:

|  |  |  |  |
| --- | --- | --- | --- |
|  | At timet0 + 1y | At timet0 + 3y | At timet0 + 5y |
| Patients tested at that time | 500 | 250 | 250 |
| Detected CIN3+ at that time | 13 (~25x0.5) | 9 (~35x0.25) | 11 (~45x0.25) |
| Cumulative CIN3+ detected | 13 | 22 | 33 |
| Censored without CIN3+ at that time | 487 | 241 | 239 |
| N at risk at that time | 1000 | 500 | 250 |
| KM Estimate of CIN3+ Cum. Risk, % | 1.3 | 3.1 | 7.3 |
| True Cum. Risk, % | 2.5 | 3.5 | 4.5 |
| Bias in Kaplan-Meier methods, % | -48% | -11% | +62% |

The Kaplan-Meier methods estimate cumulative risks of CIN3+ that are substantially below the true cumulative risks both at 1 year and 3 years following the current screen. It also estimates cumulative risk of CIN3+ that is substantially above the true cumulative risk at 5-years. This general property of underestimation and overestimation has been shown mathematically4. Similar biases occur in other survival models (e.g. standard Cox models) that ignore interval censoring and conflate time of disease detection with time of disease onset.

Also, observe that the problem of underestimation at early times is made even more severe by the presence of prevalent (left-censored) disease. Kaplan-Meier methods require no prevalent disease, but when used to estimate disease risk from screening data, prevalent disease is often diagnosed after baseline in the future. Thus, baseline prevalence is estimated at zero (as in the above example), a severe underestimation, while risks of future (incident) disease are overestimated.

The solution to these problems is to use models that account for prevalent (left-censored) disease, incident disease that is interval-censored between two times that delimit the possible times of disease onset, and for uncertainty if disease is prevalent or incident when definitive biopsies are not conducted at baseline. These are the Prevalence-Incidence Mixture models described below.

*Part B: Prevalence-Incidence Mixture Models*

In this section, we briefly describe prevalence-incidence mixture models mathematically. For further details, refer to articles by Cheung et al.4 and Hyun et al.5 The terminology disease and disease-free in the following description is used to refer to the endpoint of interest (for risk of CIN3+, the disease states are: diseased (CIN3+) “or case” and non-diseased (<CIN3) “or control”).

Given covariate values *x*, the cumulative risk at any time *t* ≥ *t0* (where *t0* denotes the current screen) can be modelled as the sum of the probability of having prevalent disease at the current screen and the probability of developing incident disease after the current screen:

 (1)

The , the probability of having prevalent disease at the current screen, is estimated with a logistic regression model with regression coefficients ,

, (2)

and is a proportional hazards model with regression coefficients and baseline survival function such that,

. (3)

Let Li be the last observed disease-free time point Ri be the time of diagnosis. If a patient is right-censored, we set Ri to infinity. Then the observed data log-likelihood is given as followed,

where *K1, K2,* and *K3* are the populations known to have prevalent disease, known to not have prevalent disease, and with unknown prevalent disease status (disease found in follow-up that could have been missed prevalent disease or could truly be incident disease), respectively. The parameters that maximized the log-likelihood can be jointly estimated using an iterative algorithm.4,5 We require that whether unknown disease is either prevalent or incidence depends only on known covariates (eg. MAR or Missing At Random).

The baseline hazard of the proportional hazards model was estimated using integrated B-splines5. In a few specific cases, using integrated B-splines resulted in model non-convergence, and thus instead the baseline hazard for the proportional hazards model was assumed to have a Weibull distribution4. In practice, risk estimates produced using integrated B-splines versus a Weibull distribution for the baseline hazards result in only negligible differences.

These predicted risk curves were also visually compared to the non-parametric risk estimates4 for checking. The non-parametric risk for screening data uses the Turnbull method for interval-censored time to event data6, but is adapted to handle undiagnosed prevalent disease.4 However, the non-parametric risk estimate is a step function with large jumps and cannot handle covariates, so its primary usefulness was in checking the fit of the prevalence-incidence mixture models. Rather than the non-parametric risk estimate, the prevalence-incidence models with a weakly-parametric baseline hazards allowed for greater statistical power, smoother predicted risk curves, and for comparison using time-specific risk estimates.

 Cumulative risk predictions were made by plugging in estimated parameters and covariate values into equations (1)-(3). The variance for the cumulative risk is derived using the multivariate delta method as

,

where is the gradient of and is the observed Fisher information.4,5 When not near the boundaries of zero or one, the cumulative risk is proven to be asymptotically normally distributed.5

*Part C: Illustrative example of how visit histories are used in risk estimation*

Using 6 illustrative examples, we show how patients’ visit histories are used to determine time intervals in which CIN3+ might have occurred. These time intervals are then used as outcomes in the prevalence-incidence model that is jointly a logistic regression model for prevalent CIN3+ and proportional hazards model for incidence of CIN3+.

Example Patient 1: Known prevalent CIN3+ (Left-censored at Baseline)

At the initial visit at time *T0*, a woman tests HPV-positive/ASC-US and is referred to colposcopy. She attends colposcopy within 1 month at time *T1* where CIN3+ is detected. We considered this patient to have CIN3+ prevalent at *T0*. She contributes as a case in the logistic regression model, but because she did not contribute any person-time, she does not contribute to the proportional hazards model.

Example Patient 2: Known Incident CIN3+ (Interval-censored)

At the initial visit at time *T0*, a woman tests HPV-positive/ASC-US and is referred to colposcopy. She attends colposcopy within 1 month at time *T1* and has normal biopsies, thus ruling out prevalent disease. She tests cotest-negative approximately 1 year later at time *T2*. She returns to testing 5 years later at time *T3*, tests HPV-positive/ASC-US, and is referred to colposcopy. She attends colposcopy at time *T4* where CIN3+ is detected. We consider this patient to have developed incident CIN3+ after *T2* and before/at *T4.* She contributes as a control in the logistic regression model. In the proportional hazards model, she contributes as a control up to *T2* and is considered “interval-censored” from *T2* to *T4* (contributing some time as a control before becoming a case at some unknown time in that interval).

Example Patient 3: Unknown Prevalent or Incident CIN3+ (Left-censored after Baseline)

At the initial visit at time *T0*, a woman tests HPV-positive/ASC-US and is referred to colposcopy. She skips colposcopy and returns to testing 6 years later at time *T1*, tests HPV-positive/ASC-US, and is again referred to colposcopy. She attends colposcopy at time *T2* where CIN3+ is detected. We consider this patient to have CIN3+ before/at *T2*, but we cannot determine if CIN3+ was prevalent at *T0.* She contributes to the logistic regression model but her status as case or control is latent (missing information). In the proportional hazards model, she is left-censored at *T2* (becomes a case at some unknown time before *T2*). A key feature of the prevalence-incidence model is that it handles this missing information via an iterative algorithm.

Example Woman 4: Right-censored in Follow-up
At the initial visit at time *T0*, a woman tests HPV-positive/ASC-US and is referred to colposcopy. She skips colposcopy and returns to testing 1 year later at time *T1* and tests cotest-negative. After *T1*, she does not return for further testing. We consider this woman to be <CIN3 at *T1* and *T0*, because CIN3+ is not believed to regress back to normal*.* She contributes as a control in the logistic regression model and contributes as a control up to *T1* to the proportional hazards model, after which she is right-censored*.*

Example Woman 5: Right-censored at Baseline (prevalent disease is ruled out)
At the initial visit at time *T0*, a woman tests HPV-positive/ASC-US and is referred to colposcopy. She attends colposcopy and has normal biopsies. Following colposcopy, she does not return for further testing. We consider this woman to be <CIN3 at *T0*, after which she is right-censored*.* She contributes as a control in the logistic regression model, but because she did not contribute any person-time, she does not contribute to the proportional hazards model.

Example Woman 6: Uninformative patient
At the initial visit at time *T0*, a woman tests HPV-positive/ASC-US and is referred to colposcopy. She does not attend colposcopy or any follow-up visit. We have no indication if she was <CIN3 or CIN3+ at any time point, and therefore she does not contribute to risk estimation in either the logistic regression model or the proportional hazards model.

*Part D: Quantifying Uncertainty in Recommended Management*

Let rx = and rx() = denote risk of CIN3+ at the current time and 5 years after the current time for patients with test results *x* in KPNC. If we knew rx and rx(), the risk-based consensus management for patients with test results *x* would be as followed, based on the consensus risk thresholds7:

(a) expedited treatment is preferred: rx(0) ≥ 0.6,

(b) either expedited treatment or colposcopy/biopsy is acceptable: rx(0) < 0.6 and rx(0) ≥ 0.25,

(c) colposcopy/biopsy is recommended: rx(0) < 0.25 and rx(0) ≥ 0.04,

(d) retesting in 1 year: rx(0) < 0.04 and rx(5) ≥ 0.0055,

(e) retesting in 3 years: rx(5) < 0.0055 and rx(5) ≥ 0.0015, and

(f) retesting in 5 years: rx(5) < 0.0015.

Since we have only a finite censored sample of patients with test results *x* from KPNC, we cannot know the true values of rx and rx(). Instead, we determine management using estimates of rx and rx() and quantify the uncertainty in the recommended management. Let and denote estimators of rx and rx() estimated from random samples of size *N* and levels of censoring *c* from KPNC*.* The probability that an estimated risk profile [consisting of and ] will fall into the management of “retesting in 1 year” (used here as an example) can be estimated as followed,

P{ < 0.04, ≥ 0.0055} = 1 – P{ ≥ 0.04} – P{ < 0.0055}.

The above result follows from an application of de Morgan’s laws and Kolmogorov axioms as followed: Let A = {: < 0.04} and B = {: ≥ 0.0055}. Then , but because risk monotonically increases with time.

Similarly, the probability that an estimated risk profile will fall into the other management options can be estimated as followed,

(a) expedited treatment is preferred: P{ ≥ 0.6} = 1 – P{ < 0.6}

(b) either expedited treatment or colposcopy/biopsy is acceptable: P{< 0.6,≥ 0.25) = 1 – P{ ≥ 0.6} – P{< 0.25},

(c) colposcopy/biopsy is recommended: P{< 0.25, ≥ 0.04) = 1 – P{≥ 0.25} – P{< 0.04},

(d) retest in 1 year: P{< 0.04, ≥ 0.0055} = 1 – P{≥ 0.04} – P{ < 0.0055},

(e) retest in 3 years: P{ < 0.0055, ≥ 0.0015} = 1 – P{ ≥ 0.0055} – P{ < 0.0015}, and

(f) retest in 5 years: P{ < 0.0015} = 1 – P{ < 0.0015}.

 When and are estimated from prevalence-incidence models, the distributions of and are normally distributed with means rx and rx() and standard errors SE{} and SE{}, respectively. We can approximate the distribution parameters by plugging in our estimated values, , , SE{}, and SE{ }, to solve for the above equations.

In some scenarios with smaller sample sizes, N, and only prevalent CIN3+, prevalence-incidence models cannot be used but immediate risks can still be estimated as the number of prevalent CIN3+ divided by (N minus the number of non-informative observation). In such scenarios, the probability that an estimated risk profile will fall into a management that involve only the immediate risk can be estimated using exact methods for the binomial distribution.

*Part E: Validation: Assessing calibration of KPNC risks to other settings*

The ratio of the numbers of observed to the number of expected CIN3+ (O/E) was estimated as the ratio of estimated risks in the validation and prediction cohorts, respectively. Using the estimated risks instead of the raw numbers of observed and expected events allow us to account for the various features of screening data (eg. prevalent asymptomatic endpoints, left-, interval-, or right-censoring). An O/E=1 means that KPNC risks exactly equal those in the other setting; O/E<1 means that KPNC risks underestimate those in the other setting.

We assume that log(O/E) is asymptotically normal and derive the variance using the Delta method. Let and be the estimated risks in the validation and prediction cohorts, respectively. The variance of log(O/E) is )+]; however, we set )=0 so as to not give credit to imprecise risk predictions. The 95% confidence intervals for the O/E can then be constructed as (exp{-1.96}, exp{+1.96}).

 Using the notation from the preceding section for immediate and 5-year risks r(0) and r(5), respectively, and clinical action thresholds t1-t5, we note that a null hypothesis that risks fall into a recommended management is equivalent to a null hypothesis that risks falls within the clinical action thresholds for that management. For example, suppose the recommended management is treatment/colposcopy, which is equivalent to having immediate CIN3+ risks bounded by t4 and t5. Suppose that the observed risk in the validation cohort exceeds t5. Then the p-value is 2P(≥t5) evaluated under the null distribution. Since prevalence-incidence models were used to estimate the observed risk in the validation cohort, the p-value can be readily estimated by assuming a normal distribution with standard errors equal to SE(

**References**

1. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Amer Statist Assoc.* 1958;53(292):457-481.

2. Landy R, Cheung LC, Schiffman M, et al. Challenges in risk estimation using routinely collected clinical data: The example of estimating cervical cancer risks from electronic health-records. *Prev Med.* 2018;111:429-435.

3. Panageas KS, Ben-Porat L, Dickler MN, Chapman PB, Schrag D. When you look matters: the effect of assessment schedule on progression-free survival. *J Natl Cancer Inst.* 2007;99(6):428-432.

4. Cheung LC, Pan Q, Hyun N, et al. Mixture models for undiagnosed prevalent disease and interval-censored incident disease: applications to a cohort assembled from electronic health records. *Stat Med.* 2017;36(22):3583-3595.

5. Hyun N, Cheung LC, Pan Q, Schiffman M, Katki HA. Flexible Risk Prediction Models for Left or Interval-Censored Data from Electronic Health Records. *Ann Appl Stat.* 2017;11(2):1063-1084.

6. Turnbull BW. The empirical distribution function with arbitrarily grouped censored and truncated data. JRSS-B. 1976;38:290-295.

7. Perkins RB, Guido RS, Castle PE, et al. 2019 ASCCP Risk-based management consensus guidelines for abnormal cervical cancer screening tests and cancer precursors. J Low Genit Tract Dis 2020;24:102–31.